

Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Anderson dos Santos

**Transformação automatizada de regulamentos financeiros para
SBVR: uma abordagem baseada em ontologias e processamento
de linguagem natural**

São Paulo

2025

Anderson dos Santos

Transformação automatizada de regulamentos financeiros para SBVR: uma abordagem baseada em ontologias e processamento de linguagem natural

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, como parte dos requisitos para a obtenção do título de mestre em Computação Aplicada.

Data da aprovação ____ / ____ / ____

Prof. Dr. Paulo Sérgio Muniz Silva
(Orientador)

Docente do Mestrado Profissional em
Computação Aplicada

Membro da Banca Examinadora:

Prof. Dr. Paulo Sérgio Muniz Silva (Orientador)
Docente do Mestrado Profissional em Computação Aplicada

Dr. Flávio Soares Correia da Silva (Membro)
IME-USP – Universidade de São Paulo

Dr. Ricardo Luis de Azevedo Rocha (Membro)
PCS-USP – Universidade de São Paulo

Anderson dos Santos

**Transformação automatizada de regulamentos financeiros para
SBVR: uma abordagem baseada em ontologias e processamento de
linguagem natural**

Dissertação de Mestrado apresentada ao
Instituto de Pesquisas Tecnológicas do Estado
de São Paulo – IPT, como parte dos requisitos
para a obtenção do título de mestre em
Computação Aplicada.

Área de Concentração: Computação Aplicada.

Linha de Pesquisa: Sistemas de Informações
em Ambientes Organizacionais Inovadores.

Orientador: Prof. Dr. Paulo Sérgio Muniz Silva.

São Paulo
Maio/2025

Ficha Catalográfica

Elaborada pela GITEB – Gerência de Gestão da Informação Tecnológica e Bibliográfica
do IPT - Instituto de Pesquisas Tecnológicas do Estado de São Paulo

S237t Santos, Anderson dos

Transformação automatizada de regulamentos financeiros para SBVR: uma abordagem baseada em ontologias e processamento de linguagem natural. / Anderson dos Santos. São Paulo, 2025.
321p.

Dissertação (Mestrado em Computação Aplicada) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Área de concentração: Computação Aplicada.

Orientador: Prof. Dr. Paulo Sérgio Muniz Silva

1. SBVR - Semantics of Business Vocabulary and Business Rules 2. Ontologia 3. Processamento de linguagem natural 4. Regulamento financeiro 5. Tese I. Silva, Paulo Sérgio Muniz, orient. II. IPT. Unidade de Negócios em Ensino Tecnológico III. Título

2025-21

CDU 004.423:65(043)

DEDICATÓRIA

Dedico este trabalho aos professores que, com paciência, sabedoria e compromisso, marcaram minha trajetória acadêmica e me inspiraram a seguir adiante mesmo diante das incertezas. Agradeço, em especial, ao meu orientador, cuja orientação precisa e apoio constante foram determinantes para a realização deste trabalho.

AGRADECIMENTOS

Agradeço a minha esposa e ao meu filho, pela paciência e compreensão ao longo desta jornada. A presença e o apoio de vocês me deram a força necessária para superar os desafios e concluir este projeto. Este trabalho também é, de certa forma, um reflexo do amor e do incentivo que sempre recebi de vocês.

Agradeço ainda os meus gatos pela companhia, quase sempre silenciosa, pelo carinho incondicional e por tornarem os momentos de estudo mais leves e acolhedores.

RESUMO

Este trabalho propôs a transformação automatizada de regulamentos financeiros redigidos em linguagem natural para vocabulário e regras estruturadas no formato *Semantics of Business Vocabulary and Business Rules* (SBVR). O objetivo principal foi desenvolver, implementar e validar um método que empregou técnicas avançadas de processamento de linguagem natural (NLP) e ontologias, como a *Financial Industry Business Ontology* (FIBO), para a transformação dos regulamentos. A pesquisa seguiu o ciclo de *Design Science Research* (DSR), organizado em quatro etapas principais: análise do problema, elicitação de requisitos, desenvolvimento da solução e validação experimental. O método proposto, denominado CFR2SBVR, utilizou algoritmos de NLP, grafos de conhecimento e modelos probabilísticos para extrair e transformar regulamentos do *Code of Federal Regulations* (CFR) em elementos no formato SBVR. O enfoque esteve na preservação da semântica original dos regulamentos e na garantia de rastreabilidade entre os textos legais e os elementos transformados. Os experimentos mostraram que o método alcançou 92% de precisão na identificação e classificação de termos regulatórios e 89% na transformação de regras operativas para SBVR. Além disso, obteve-se uma rastreabilidade de 95%, assegurando uma correspondência verificável entre as regras transformadas e os textos originais. Esses resultados superaram abordagens baseadas em regras, evidenciando a eficiência do método na resolução de ambiguidades e complexidades textuais. A pesquisa contribuiu com a conformidade regulatória ao apresentar um arcabouço que pode ser utilizado para extrair, classificar e catalogar as regras contidas nos regulamentos, aprimorando os processos de Governança, Risco e Conformidade (GRC). Os dados, algoritmos e resultados obtidos foram disponibilizados para garantir transparência e reproduzibilidade, ampliando o impacto acadêmico e prático do trabalho.

Palavras-chave: SBVR, GRC, FRO, FIBO, Governança, Risco, Conformidade, Ontologia, NLP, LLM.

ABSTRACT

Automated Transformation of Financial Regulations into SBVR: An Ontology-Based and Natural Language Processing Approach

This study proposed the automated transformation of financial regulations written in natural language into structured vocabulary and rules in the Semantics of Business Vocabulary and Business Rules (SBVR) format. The primary objective was to develop, implement, and validate a method that employed advanced natural language processing (NLP) techniques and ontologies, such as Financial Industry Business Ontology (FIBO), for the transformation of regulations. The research followed the Design Science Research (DSR) cycle, organized into four main stages: problem analysis, requirements elicitation, solution development, and experimental validation. The proposed method, named CFR2SBVR, used NLP algorithms, knowledge graphs, and probabilistic models to extract and transform regulations from the Code of Federal Regulations (CFR) into SBVR-compliant elements. The focus was on preserving the original semantics of the regulations and ensuring traceability between the legal texts and the transformed elements. Experiments showed that the method achieved 92% accuracy in identifying and classifying regulatory terms and 89% accuracy in transforming operational rules into SBVR. Additionally, it achieved 95% traceability, ensuring a verifiable correspondence between the transformed rules and the original texts. These results outperformed rule-based approaches, demonstrating the method's effectiveness in resolving textual ambiguities and complexities. The research contributed to regulatory compliance by presenting a framework that can be used to extract, classify, and catalog the rules contained in regulations, improving Governance, Risk, and Compliance (GRC) processes. The data, algorithms, and results were made available to ensure transparency and reproducibility, increasing the academic and practical impact of the work.

Keywords: SBVR, GRC, FRO, FIBO, Governance, Risk, Compliance, Ontology, NLP, LLM.

Lista de Figuras

Figura 1 - Método de trabalho	28
Figura 2 - Organização do trabalho	29
Figura 3 - Fases do processo da Revisão Sistemática	31
Figura 4 - Execução da RSL	37
Figura 5 - Fluxo conceitual para conformidade, baseado no livro vermelho GRC (OCEG, 2023) combinado com processo de três etapas para gerenciar a conformidade (RACZ <i>et al.</i> , 2010).....	39
Figura 6 - Taxonomia de Witt (2012)	51
Figura 7 - Arcabouço para o processo de transformação de NL para SBVR.	114
Figura 8 - Anotação semântica.....	122
Figura 9 – KGs conceitual	125
Figura 10 - Associação dos elementos no modelo de conteúdo SBVR.....	132
Figura 11 - Arquitetura do CFR2SBVR.....	143
Figura 12 - Visão geral da transformação de CRF para SBVR.	145
Figura 13 - Componente “create kg”	146
Figura 14 - Consulta SPARQL para termos FIBO.	147
Figura 15 - Consulta SPARQL para seção da CFR.....	147
Figura 16 - Algoritmo orquestrador – “transform2sbvr”.....	148
Figura 17 - Algoritmo “semantic annotation”	151
Figura 18 - Pseudocódigo do <i>prompt</i> para extração de termos – primeira abordagem	175
Figura 19 - Algoritmo “Transformation”.....	179
Figura 20 - Algoritmo “elements association and creation”	185
Figura 21 - Algoritmo “define vocabulary namespace”	194
Figura 22 - Fragmento da organização dos <i>namespaces</i> para os vocabulários da CFR	195
Figura 23 - Processo CFR2SBVR	202
Figura 24 - Extração e classificação das regras operativas e tipos de fatos	217

Figura 25 - Similaridade entre as declarações extraídas e do conjunto de dados ouro.	218
Figura 26 - Matriz de confusão da identificação da origem dos elementos	219
Figura 27 - Classificação das regras operativas (P1).....	221
Figura 28 – Classificação (P2) das regras operativas nos subníveis da taxonomia.	222
Figura 29 - <i>Templates</i> para as classificações de termos.....	223
Figura 30 - Classificação (P2) para termos	224
Figura 31 - Classificação (P2) para nomes.	225
Figura 32 - Classificação (P2) para tipo de fatos.....	225
Figura 33 - Distribuição para todos os elementos das pontuações de similaridade semântica	230
Figura 34 - Histograma de distribuição do <i>Similarity Score</i> e <i>SemScore</i>	231
Figura 35 - Diferença entre <i>Similarity Score</i> e <i>SemScore</i> por tipo de elemento.....	231
Figura 36 – Uso dos <i>tokens</i> por processo.....	234
Figura 37 - Tempo total por processo	234
Figura 38 - CFR2SBVR semi-automático.....	239

Lista de Quadros

Quadro 1 - Fontes de busca para a revisão sistemática	32
Quadro 2 - <i>Strings</i> de busca para pergunta PICO nas bases de dados.....	34
Quadro 3 - <i>Strings</i> de busca para pergunta de conhecimento nas bases de dados .	36
Quadro 4 - Classificação das regras segundo Von Halle (2001).....	49
Quadro 5 - Fragmento da ontologia SBVR transformada do XSD	54
Quadro 6 - Relações jurídicas	58
Quadro 7 - Requisitos	91
Quadro 8 - Comparação das abordagens que transformam regras de negócio para SBVR	93
Quadro 9 - Ontologia SBVR adaptada	130
Quadro 10 - Inserção de um termo no KG	133
Quadro 11 - Problema com o KG CFR.....	149
Quadro 12 - <i>Prompt</i> para extração de termos – primeira abordagem	153
Quadro 13 – Fragmento do resultado do algoritmo “extract / classify elements”....	156
Quadro 14 - <i>Prompt</i> do algoritmo “extract / classify elements” parte 1. Segunda abordagem.....	159
Quadro 15 - <i>Prompt</i> do algoritmo “extract / classify elements” parte 2. Segunda abordagem.....	162
Quadro 16 - <i>Prompt</i> para classificação dos elementos no topo da taxonomia de Witt (2012).	166
Quadro 17 - <i>Prompt</i> para classificação na taxonomia e obtenção do <i>template</i> para transformação (parte 1 – regras operativas).	170
Quadro 18 - Instruções para a transformação para SBVR.....	180
Quadro 19 - Exemplo de instruções para utilização do <i>template</i> na transformação.	182
Quadro 20 - Instruções para uso dos templates.....	184
Quadro 21 - Triplas para regras de definição.	186
Quadro 22 - Tripla para regras operativas.....	188
Quadro 23 - Consulta SPARQL para recuperar referências.	189
Quadro 24 - Pesquisa por elementos sem definição.....	190
Quadro 25 - Busca por similaridade no KG FIBO.....	192

Quadro 26 - Consulta SPARQL para recuperar vocabulários e <i>namespaces</i> dos vocabulários	196
Quadro 27 - Consulta SPARQL para recuperar o <i>namespace</i> do vocabulário associado a designação.....	196
Quadro 28 - Prompt do <i>LLM as a judge</i>	204
Quadro 29 - <i>Prompt</i> para classificação do tipo da regra.	207
Quadro 30 - Fragmento do <i>prompt</i> para a subtipo de regras	208
Quadro 31 - Exemplo da saída do algoritmo “ <i>classify</i> ”.....	209
Quadro 32 - Comparativo dos modelos T8 e T10 de escrita de regras	211
Quadro 33 - Comparativo entre modelos de escrita T22 e T23.....	212
Quadro 34 - Resultado dos autores.	245
Quadro 35 - Artigos base para a revisão sistemática	267
Quadro 36 - Artigos retornados por base de dados.....	270
Quadro 37 - Principais revistas e congressos consultados	287
Quadro 38 - Uso de IA pelos produtos	288
Quadro 39 - Fornecedores de GRC	289
Quadro 40 - Taxonomia de regras definidas por Witt (2012)	303
Quadro 41 - Elementos transformados	306
Quadro 42 - Classificação e <i>templates</i> usados na transformação dos elementos ..	318

Lista de Tabelas

Tabela 1 - Conjunto de dados CFR	213
Tabela 2 - Conjunto de dados ouro para validação (P1)	214
Tabela 3 - Termos e nomes com definição.	214
Tabela 4 - Resumo métricas do processo de extração.....	216
Tabela 5 - Conjunto de dados ouro para a classificação (P1)	221
Tabela 6 - Resumo das métricas de classificação.....	226
Tabela 7 - Estatística das métricas combinadas.....	227
Tabela 8 - Pontuações de qualidade da transformação	229
Tabela 9 - Pontuações de confiança no julgamento do LLM	229
Tabela 10 - Pontuações de validação dos elementos	315

Lista de abreviaturas e siglas

AMEX	American Stock Exchange
ANBIMA	Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais
BiLM	Bi-directional Language Model
BPMN	Business Process Model and Notation
CCG	Combinatory Categorial Grammar
CFR	Code of Federal Regulations
CFTC	Commodity Futures Trading Commission
CNL	Controlled Natural Language
COLIEE	Competition on Legal Information Extraction and Entailment
CNL	Controlled Natural Language
CoT	Chain of Thought
CVM	Comissão de Valores Mobiliários
DPR	Dense Passage Retrieval
DSR	Design Science Research
ECA	Event condition action
EDMC	Enterprise Data Management Council
FIBO	Financial Industry Business Ontology
FINRA	Financial Industry Regulatory Authority
FN	False Negative
FP	False Positive
FRO	Financial Regulation Ontology
GNN	Graph Neural Network
GPO	Government Publishing Office
GPT	Generative Pre-trained Transformer
GRC	Governance, Risk, and Compliance
GRCTC	Governance, Risk, and Compliance - Technical Center
GRCA	Governance, Risk, and Compliance
IA	Inteligência Artificial
KG	Knowledge Graph

KGE	Knowledge Graph Embedding
LKIF	Legal Knowledge Interchange Format
LLM	Large Language Model
M2M	Model-to-model
M2T	Model-to-text
MBSE	Model-Based Systems Engineering
MDA	Model Driven Architecture
ML	Machine Learning
MOF	Meta Object Facility
NASDAQ	National Association of Securities Dealers Automated Quotations
NER	Named Entity Recognition
NL	Natural Language
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OCEG	Open Compliance and Ethics Group
OCL	Object Constraint Language
OGD	Open Government Data
OLRC	Office of the Law Revision Counsel
OMG	Object Management Group
OSS	Open-Source Software
OWL	Web Ontology Language
OWL-DL	Web Ontology Language Description Logics
OWL2	Web Ontology Language 2
POS	Parts of Speech
PSI	Public Sector Information
RAG	Retrieval Augmented Generation
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RLAIF	Reinforcement Learning from AI Feedback
RLHF	Reinforcement Learning with Human Feedback
SaaS	Software as a Service
SBVR	Semantic for Business Vocabularies and Business Rules

SBVR-SE	Semantic for Business Vocabularies and Business Rules - Structured English
SE	Structured English
SEC	Securities and Exchange Commission
SKOS	Simple Knowledge Organization System
SME	Subject Matter Expert
SRS	Software Requirements Specification
T2M	Text-to-model
TP	True Positive
UML	Unified Modeling Language
URI	Uniform Resource Identifier
USC	United States Code
USE	UML-based Specification Environment
W3C	World Wide Web Consortium
XAI	Explainable AI
XMI	XML Metadata Interchange

SUMÁRIO

1 INTRODUÇÃO	16
1.1 Motivação	19
1.2 Declaração do problema	23
1.3 Objetivo	25
1.4 Contribuições	27
1.5 Método de trabalho	27
1.6 Organização do trabalho	29
2 REVISÃO SISTEMÁTICA DA LITERATURA.....	31
2.1 Método	31
2.2 Protocolo	32
2.3 Execução.....	34
3 FUNDAMENTAÇÃO CONCEITUAL.....	38
3.1 Abordagens formais de verificação de conformidade	38
3.2 Semantic for Business Vocabularies and Business Rules (SBVR).....	42
3.3 Fonte de regras de negócio e vocabulário	55
3.4 Posições jurídicas e modalidades deôntica e aléticas.....	58
3.5 Aplicações de NLP no contexto legal	61
3.6 Transformação em vocabulário e regras	63
3.7 Ontologias	64
3.8 Grandes modelos de linguagem (LLMs).....	67
3.9 Grafos de conhecimento	74
3.10 Geração aumentada por recuperação (RAG).....	79
3.11 Validação e testes.....	81
3.12 Resumo	86
4 DISCUSSÃO DAS SOLUÇÕES	87
4.1 Padrões para comparação das abordagens de transformação.....	87

4.2 Abordagens de transformação de NLP para SBVR.....	93
4.3 Preocupações éticas e legais.....	106
4.4 Resumo	109
5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR	111
5.1 Introdução	111
5.2 Principais artefatos	116
5.3 Cenários de validação	133
5.4 Resumo	138
6 FERRAMENTAS DE SUPORTE	139
6.1 Ferramentas e infraestrutura	140
6.2 Implementação dos principais componentes.....	144
7 VALIDAÇÃO	198
7.1 Validação por similaridade.....	199
7.2 Métodos aplicados ao experimento	202
7.3 Resultados	215
8 CONSIDERAÇÕES FINAIS	237
8.1 Discussão.....	237
8.2 Conclusão	247
8.3 Trabalhos futuros.....	249
REFERÊNCIAS.....	253
APÊNDICE A – RESULTADOS DA RSL	267
APÊNDICE B – FORNECEDORES DE SOLUÇÕES PARA A GRC	288
APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO	300
APÊNDICE D – RESULTADO DA TRANSFORMAÇÃO	306

1 INTRODUÇÃO

Segundo o grupo aberto de ética e conformidade (*Open Compliance and Ethics Group - OCEG*), a governança, o risco e a conformidade (*Governance, Risk, and Compliance - GRC*) são elementos essenciais para o setor financeiro e, anualmente, mais de USD 1 trilhão de dólares são perdidos devido a condutas inescrupulosas, erros e cálculos imprecisos. Para endereçar esse problema, as organizações, os indivíduos e a sociedade contam com os processos e profissionais de GRC. Em 2020, o setor bancário acumulou USD 15 bilhões em multas por não conformidades, sendo 73% de bancos americanos¹. Em um ambiente marcado por condições voláteis, incertas, complexas e ambíguas, organizações de todos os tipos e tamanhos buscam soluções de GRC. O mercado global de GRC foi avaliado em USD 4.4 milhões em 2021 e tem previsão de crescimento para USD 1.04 bilhão até 2027 (360 MARKET UPDATES, 2023). Porém, uma breve avaliação da aplicação de Inteligência Artificial (IA) nas empresas listadas no relatório da Forrester Wave² GRC Q4 2023 mostrou que poucas empresas abordam a questão do uso de IA para os tratamentos de regras de negócio.

A OCEG define GRC como a capacidade, ou conjunto integrado de capacidades, que permite a uma organização alcançar objetivos de forma confiável, lidar com a incerteza e agir com integridade; incluindo a governança, garantia e gestão do desempenho, risco e conformidade, o que está em linha com a definição obtida por Racz *et al.* (2010), na sua revisão da literatura para GRC. Segundo os autores, a conformidade regulatória é o principal impulsionador do GRC, seguida pela gestão de riscos.

As soluções de GRC auxiliam as empresas a lidarem com a conformidade regulatória, porém, na sua maioria, elas se baseiam em gerenciamento de conteúdo, documentos e dependem de especialistas. Essas abordagens, geralmente, são parcialmente formais. A verificação formal de conformidade

¹ Ver mais em: <https://www.corporatecomplianceinsights.com/banks-15b-in-fines-in-2020>. Acesso em: 15/mar./2025.

² Ver mais em: <https://www.forrester.com/policies/forrester-wave-methodology/>. Acesso em: 15/mar./2025.

oferece várias vantagens em termos de análise (SUNKLE *et al.*, 2015a). Dado que a não conformidade é sujeita a penalizações severas na maioria dos países e em várias áreas de negócios, resolver a conformidade regulatória de forma eficaz e eficiente se tornou uma prioridade. O aumento no volume de mudanças regulatórias representa desafios de conformidade, incluindo custos não monitorados, expectativas de crescimento das equipes de conformidade e a necessidade de reter recursos qualificados, com baixa moral da equipe emergindo como um risco de conduta (REUTERS, 2023).

No mercado financeiro, a baixa maturidade do GRC pode levar a consequências desastrosas, não só para as empresas, como para a sociedade. Exemplos foram as crises financeiras mundiais desencadeadas pela má conduta de empresas do setor financeiro e que culminaram, nos EUA, com uma profunda reformulação do setor, pelas leis Dodd-Frank *Wall Street Reform and Consumer Protection Act* de 2010 (CONGRESS, 2010), promulgada em resposta à crise financeira de 2008 (Crise do Subprime³) e a SOx (OXLEY, 2002), promulgada em resposta a uma série de escândalos contábeis corporativos que abalaram a confiança dos investidores nos mercados financeiros, entre eles o colapso da empresa de energia Enron, os problemas contábeis na empresa de comunicações WorldCom e a empresa de contabilidade Arthur Andersen, que auditava várias das empresas envolvidas nos escândalos⁴.

Essas crises tiveram profundo impacto nas economias mundiais, porque afetaram a confiança dos investidores. Para efeito de comparação, o mercado mobiliário no Brasil, pela Brasil, Bolsa e Balcão (B3), é responsável pela negociação, registro, oferta, liquidação e operacionalização de ativos e demais instrumentos financeiros do mercado de mais de 400 empresas. Esse mercado negocia, em média, cerca de 30 bilhões de reais diariamente, tendo movimentado mais de 7 trilhões de reais em 2022 (B3, 2023a). Em 2021, foram 28 IPOs⁵ e 26 *Follow-ons*⁶ com volume total captado cerca de 130 bilhões de

³ Ver mais em: <https://warren.com.br/magazine/crise-do-subprime/>. Acesso em: 15/mar./2025.

⁴ Ver mais em: <https://www.jusbrasil.com.br/artigos/enron-corporation-uma-analise-sobre-a-tragedia-da-6-maior-empresa-de-energia-do-mundo/1283408527>. Acesso em: 15/mar./2025.

⁵ Oferta Pública Inicial (Initial Public Offering - IPO).

⁶ O *Follow-on* significa oferta subsequente. É realizado por empresas que já têm capital aberto e ações negociadas na bolsa que decidem fazer uma nova oferta pública de ações.

reais (B3, 2023b). Em 2022, não houve IPOs, devido ao cenário interno e internacional (FERREIRA; LI; NOSSIG, 2023), mas foram captados mais de 57 bilhões de reais em *Follow-ons*. Mesmo com essa evolução, o Brasil possui pouquíssimas empresas listadas quando comparado com outros países. Na Inglaterra, o número se aproxima de duas mil empresas, enquanto na Ásia, destaca-se a China e o Japão, que possuem quase quatro mil companhias listadas. Liderando estão os EUA, com um número que passa de sete mil empresas, nas bolsas da *National Association of Securities Dealers Automated Quotations* (NASDAQ), *New York Stock Exchange* (NYSE) e *American Stock Exchange* (AMEX) (NASDAQ, 2023).

Somente na NYSE são mais de três mil empresas listadas, incluindo empresas domésticas e internacionais e na NASDAQ há mais de 3,3 mil (STATISTA, 2023). A NYSE é a bolsa com a maior capitalização total (*market cap*) de USD 27 trilhões e tem uma média diária de movimentação de USD 1.2 trilhões, sendo seguida pela NASDAQ, conhecida por abrigar empresas do setor tecnológico, com um volume médio diário aproximado de USD 1.8 trilhões e capitalização total de USD 24.5 trilhões. Juntas, essas duas bolsas representam mais de 50% do volume médio diário negociado em bolsas de valores ao redor do mundo (INFOMONEY, 2022). Para efeito de comparação, a capitalização total da B3 é de USD 837 bilhões (B3, 2023d).

No mercado financeiro, as bolsas de valores negociam *securities*, um termo amplo que engloba vários instrumentos financeiros que representam uma participação acionária, uma dívida ou outro valor financeiro e *equities* um subconjunto de *security*, referindo-se, especificamente, à propriedade em uma empresa. Nos EUA, *equities* representam mais de 64% do mercado de capitais⁷, seguido pelo Reino Unido com 59%, Zona do Euro com 55% e o Japão com 50% (SIFMA, 2023). No Brasil, não ultrapassam 3% (ANBIMA, 2023).

Como mencionado, crises nestes mercados têm grande impacto econômico, tanto para empresas já estabelecidas, quanto entrantes. Esse segundo grupo depende fortemente da confiança do investidor para obter capital. Os órgãos

⁷ O mercado de capitais é um conjunto de instituições, agentes econômicos e instrumentos legais dedicados à distribuição de valores mobiliários.

reguladores, juntamente com o legislativo, são responsáveis por estabelecer normas e leis que regulam o mercado e aumentam a sua confiabilidade. No mercado de capitais brasileiro, os órgãos reguladores são a Comissão de Valores Mobiliários (CVM) e entidades privadas, como B3 e Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA), por delegação da CVM. Nos EUA, o órgão regulador é a comissão de valores mobiliários e câmbio (*Securities and Exchange Commission - SEC*). Essa agência federal foi criada pela lei de bolsas de valores de 1934 (*Securities Exchange Act of 1934*). Os regulamentos gerais e permanentes promulgados pelos departamentos executivos e agências do governo federal dos EUA são mantidos pelo código de regulamentos federais (*Code of Federal Regulations - CFR*). Os regulamentos mantidos pela CFR são os de interesse dessa pesquisa.

1.1 Motivação

A pesquisa é impulsionada pela necessidade crescente de gerenciar a conformidade regulatória de maneira eficiente no setor financeiro. As instituições financeiras estão submetidas a um crescente volume de regulamentos complexos e em constante mudança. Esse cenário impõe desafios significativos na interpretação e implementação das regulamentações, resultando em riscos elevados de não conformidade e penalidades severas, que podem impactar a saúde de todo o ecossistema financeiro, podendo levar a crises mundiais. As práticas de GRC, chamadas de tradicionais, são as que as auditorias conduzem com auditores para a detecção de não conformidades “após o fato” (SADIQ *et al.*, 2007; KULKARNI *et al.*, 2021). Essas práticas tradicionais de gestão regulatória são, muitas vezes, manuais, lentas e propensas a erros, destacando a necessidade de uma abordagem mais automatizada e confiável (KULKARNI *et al.*, 2021; BOELLA *et al.*, 2013; ROYCHOUDHURY *et al.*, 2017; SUNKLE *et al.*, 2015a; BHAROSA *et al.*, 2011; SUNKLE *et al.*, 2015b). Esse tema é abordado com mais detalhes nesta pesquisa.

Na gestão comercial, o reconhecimento e a interpretação das legislações aplicáveis ao setor são essenciais, envolvendo a análise detalhada de normas e regulamentos pertinentes para a operação dos negócios. Esse processo, frequentemente, depende da competência de especialistas jurídicos e de

negócios, muitas vezes provenientes de consultorias externas, o que demanda um investimento significativo de tempo e recursos. Tais processos tendem a depender de soluções de governança, risco e conformidade (*Governance, Risk and Compliance - GRC*) que, por vezes, não incorporam o rigor dos formalismos teóricos discutidos no âmbito acadêmico (SUNKLE *et al.*, 2015b). A complexidade dos textos legais, aliada às constantes atualizações regulatórias, representam obstáculos significativos para verificar e assegurar a conformidade (SUNKLE *et al.*, 2015a). Nesse cenário, a capacidade de rastrear e vincular as ações da empresa às exigências regulamentares é vital, habilitando os auditores a determinarem, com precisão, quais procedimentos precisam ser revisados em conformidade com a legislação pertinente (BOUZIDI *et al.*, 2011).

Nos últimos anos, as iniciativas de governo aberto avançaram além da simples publicação de dados, buscando transparência, governança participativa e valorização social e comercial. Essas iniciativas, como a diretiva de informações do setor público (*Public Sector Information - PSI*), a iniciativa de dados abertos do presidente Obama, a parceria para o Governo Aberto⁸ e a carta de dados abertos assinada pelo G8⁹, visam permitir que cidadãos e partes interessadas monitorem as ações governamentais e promovam o controle social. A transparência não se limita ao acesso aos dados, mas inclui a capacidade de usá-los, reutilizá-los e distribuí-los, fomentando uma participação ativa nos processos de governança. As iniciativas de dados abertos do governo possibilitam que os interessados sejam mais informados e tomem decisões melhores, impactando diversas áreas, como gestão urbana, marketing, melhoria de serviços e qualidade de vida (ATTARD; ORLANDI; AUER, 2016).

Os governos são grandes produtores de dados em vários domínios, destacando o valor social e comercial dessas informações. Os processos existentes de criação de valor em dados abertos do governo (*Open Government Data - OGD*) envolvem várias etapas projetadas para transformar dados governamentais brutos em recursos valiosos para diversos interessados. Esses processos incluem a coleta e a disponibilização de conjuntos de dados governamentais,

⁸ Ver mais em: <https://obamawhitehouse.archives.gov/open>. Acesso em: 15/mar./2025.

⁹ Ver mais em: <https://opendatacharter.org/resources/g8-open-data-charter/>. Acesso em: 15/mar./2025.

garantindo a sua acessibilidade e usabilidade, promovendo o seu reuso em diferentes setores. Agências governamentais, empresas privadas e organizações da sociedade civil se engajam com OGD para desenvolver novos serviços, aprimorar a tomada de decisões e promover transparência e responsabilidade. O processo de criação de valor também envolve abordar barreiras técnicas, legais e culturais para maximizar o impacto e a utilidade dos dados abertos do governo (JANSSEN *et al.*, 2011).

Essas iniciativas, aliadas a recentes avanços nas tecnologias de processamento de linguagem natural (*Natural Language Processing* - NLP), entendimento de linguagem natural (*Natural Language Understanding* - NLU) e geração de linguagem natural (*Natural Language Generation* - NLG) têm habilitado a interpretação de textos em linguagem natural (*Natural Language* - NL) sem precedentes (KHURANA *et al.*, 2022), permitindo um incremento significativo na qualidade dos resultados de pesquisas com o mesmo objetivo. Guiado também pela necessidade de melhorar processos de negócio, ontologias específicas de domínios têm surgido, como a ontologia de negócios para a indústria financeira (*Financial Industry Business Ontology* - FIBO) (BENNETT, 2022; FORD *et al.*, 2016), que reúne o conhecimento de diversos profissionais do setor, pela coordenação do conselho de gestão de dados empresariais (*Enterprise Data Management Council* - EDMC).

A legislação, redigida em linguagem natural, impõe desafios consideráveis à interpretação precisa. A complexidade reside não só na identificação léxica, mas na organização e aplicação consciente de palavras para cumprir objetivos bem definidos, com uma compreensão clara das intenções subjacentes (HOUGH *et al.*, 2019). A tarefa é complicada pela polissemia das palavras e pela vasta gama de possíveis interpretações de uma única frase. A ambiguidade, uma propriedade intrínseca da linguagem natural, ocorre, frequentemente, não por uma falha na comunicação, mas devido a um contexto que tende a privilegiar determinadas leituras em detrimento de outras (JACKSON, 2020).

Em contrapartida, as linguagens naturais controladas (*Controlled Natural Languages* - CNLs) propõem um mapeamento direto e bilateral com linguagens formais, como a de lógica de primeira ordem (*First-Order Logic* - FOL) ou a lógica

de descrição (*Description Logic - DL*), visando clareza (GRUZITIS; BARZDINS, 2009). A semântica de vocabulário de negócios e regras de negócios (*Semantics of Business Vocabulary and Business Rules - SBVR*) da OMG exemplifica essa abordagem, estruturando o conhecimento empresarial em uma CNL e oferecendo uma estrutura para definir e gerenciar vocabulários e regras de negócios. Ela proporciona expressividade, formalidade e suporte para a linguagem natural, possibilitando uma comunicação clara dos conceitos de negócio. O SBVR suporta múltiplos idiomas, promovendo a internacionalização e garantindo a interoperabilidade entre sistemas. A sua abordagem estruturada permite a reutilização e a extensibilidade de termos e regras, facilitando a consistência e a adaptabilidade em ambientes de negócios em evolução (OMG, 2019).

No entanto, mesmo com o uso de uma linguagem controlada, ao tratar regulamentos e normas, não se enfrenta apenas a ambiguidade linguística, mas também o desafio de conciliar regras que, por vezes, encontram-se em conflito. Por exemplo: suponha que existe uma norma n_1 que prescreve C sob a condição A , e a sanção por não fazer C é S_1 ; ao mesmo tempo, existe uma segunda norma n_2 , independente da primeira, que proíbe C quando B é verdade; a violação da segunda norma é sancionada por S_2 . Além disso, as duas normas estão no mesmo nível na hierarquia do sistema normativo subjacente; assim, não há meios claros para avaliar se uma das duas normas prevalece sobre a outra (GOVERNATORI; ROTOLI, 2023).

Nesse cenário, a afirmação de que “ C é obrigatório”, do ponto de vista deontológico, é ambígua, uma vez que existem duas linhas de raciocínio, uma sustentando essa conclusão e outra sustentando o contrário. Mitra *et al.* (2018) buscam resolver o problema de como fusões e aquisições, atualizações e incorporação de uma nova aplicação tornam necessário comparar um conjunto de regras de negócios de uma organização com as regras de um modelo de referência. Para obter uma medida de similaridade entre a funcionalidade de negócios, eles propõem uma ferramenta que, dado um conjunto de regras, detecta lacunas entre elas. Esses são exemplos de problemas que o avanço das tecnologias de processamento de linguagem natural pode auxiliar na resolução.

A constante necessidade de alinhar processos de TI com os de negócio é um diferencial competitivo e enfatizado nos últimos anos pelos movimentos de transformação digital. A representação das regras e vocabulário em SBVR servem a esse propósito, passando a ser um passo importante para esse alinhamento. Isso deverá permitir que processos como os de GRC sejam implementados mais rapidamente e de forma confiável, levando as empresas, os governos e os cidadãos a obterem os benefícios desses avanços tecnológicos.

Essa é uma área de pesquisa bastante ativa. Abi-Lahoud *et al.* (2013) descrevem uma metodologia semiautomática para interpretar e formalizar regulamentações com SBVR envolvendo especialistas em todas as etapas, o que torna os resultados dependentes do conhecimento e velocidade dos especialistas. Joshi e Saha (2020) convertem regulamentações do CFR em um formato estruturado e legível por máquinas, usando IA para extrair termos e regras e organizá-los em um grafo de conhecimento (*Knowledge Graph - KG*). O arcabouço não visa transformar regulamentos em regras SBVR, mas fornece uma base para um sistema de perguntas e respostas. Haj *et al.* (2021a) apresentam um método automático para transformar regras de negócios em linguagem natural para SBVR, utilizando NLP para análise linguística e mapeamento de elementos SBVR. O método se mostra eficaz para regras explícitas e bem estruturadas, mas enfrenta limitações com regras complexas, ambíguas ou implícitas. Essas pesquisas contribuem para resolver ou minimizar alguns desses desafios, mas têm as suas próprias limitações. Essas e outras pesquisas são discutidas com mais profundidade na Seção “4.2 Abordagens de transformação de NLP para SBVR”.

1.2 Declaração do problema

No setor financeiro, as perdas anuais, devido à má conduta, erros e cálculos equivocados não apenas desperdiçam recursos, mas também podem levar a danos significativos à sociedade, incluindo perda de ativos e crises financeiras globais, como já mencionado. A governança, risco e conformidade (*Governance, Risk and Compliance – GRC*) desempenha papel importante na proteção da sociedade. As agências reguladoras impõem regulamentos que devem ser

seguidos, com sanções severas para a não conformidade. A complexidade e a dinâmica dos regulamentos, que estão em constante evolução, apresentam desafios significativos, que incluem a interpretação precisa dos regulamentos, a extração de regras de negócio relevantes e a transformação dessas regras em uma forma estruturada e formalizada, para facilitar a implementação e a verificação automatizada de conformidade. Assim, a desambiguação dos conceitos, a conversão das regras de negócio em linguagem natural para uma linguagem natural controlada, a criação de um catálogo e a rastreabilidade entre as regras e os trechos originais dos regulamentos são ferramentas necessárias para aprimorar os processos de GRC.

Dada a complexidade e a constante evolução dos regulamentos financeiros, bem como os desafios associados à sua interpretação e implementação, a presente pesquisa busca responder à seguinte pergunta, estruturada de acordo com a metodologia PICO (PETTICREW; ROBERTS, 2008):

- a) População (P): Regulamentos financeiros do CFR escritos em linguagem natural;
- b) Intervenção (I): Aplicação de técnicas avançadas de NLP e uso de ontologias para transformar regulamentos financeiros em regras e vocabulário estruturados, conforme a especificação SBVR;
- c) Comparação (C): Métodos existentes para interpretação manual e automatizada de textos regulatórios;
- d) Resultado (O): Melhoria na precisão da transformação de regulamentos financeiros em regras e vocabulário estruturados.

Dessa forma, a pergunta de pesquisa central que este trabalho pretende responder é: como a aplicação de técnicas avançadas de processamento em linguagem natural e ontologias, como a FIBO, podem aprimorar a transformação automatizada de regulamentos financeiros em regras e vocabulário estruturados, conforme a especificação SBVR?

Para apoiar o ciclo de projeto e contextualizar o método com a prática de GRC, além da pergunta de pesquisa, foram respondidos os questionamentos de conhecimento a seguir.

- a) Quais são as abordagens existentes para transformação de regulamentos em linguagem natural para linguagem controlada? Identifica-se, aqui, as abordagens, as barreiras técnicas, semânticas e de implementação.
- b) Quais e como as técnicas de processamento de linguagem natural podem ser aplicadas para melhorar o processo de transformação?
- c) De que maneira ontologias podem ser utilizadas para resolver ambiguidades? Investiga-se, neste caso, como ontologias, como a FIBO, podem ser integradas ao processo de conversão para esclarecer termos ambíguos e garantir que as regras convertidas refletem com precisão o significado pretendido.
- d) Como são realizados os processos de GRC? Como os processos são conduzidos?
- e) Como a transformação de regulamentos financeiros para SBVR podem ser validadas? São observadas as estratégias para verificar a precisão das regras de negócios transformadas para SBVR.
- f) Quais são as questões éticas envolvidas no processo?

1.3 Objetivo

O objetivo deste trabalho é propor, implementar e validar um método automatizado que empregue técnicas avançadas de processamento de linguagem natural (*Natural Language Processing – NLP*) para transformar regulamentos do código federal de regulamentos dos EUA (*Code of Federal Regulations – CFR*), escritos em linguagem natural, para regras e vocabulário em inglês estruturado (*Structured English - SE*), conforme a especificação da semântica de vocabulário de negócios e regras de negócios (*Semantics of Business Vocabulary and Business Rules - SBVR*) da OMG, abreviada para CFR2SBVR (CFR to SBVR).

A hipótese central desta pesquisa é que a aplicação de um processo automatizado baseado em técnicas avançadas de processamento de linguagem

natural com um modelo de Inteligência Artificial (IA) probabilístico pré-treinado, o uso de grafo de conhecimento (*Knowledge Graph* - KG) para representar textos legais e o suporte da ontologia de negócios para a indústria financeira (*Financial Industry Business Ontology* - FIBO) para o vocabulário deverão resultar em uma precisão superior aos trabalhos estudados. A seguir, são apresentados os resultados esperados:

- a) na desambiguação de conceitos nos textos regulatórios, a FIBO deverá aumentar a precisão na transformação dos regulamentos, reduzindo ambiguidades e erros de interpretação (BENNETT, 2021; FORD *et al.*, 2016). Essa ontologia foi desenvolvida pelo EDMC e padronizada pelo OMG, com o propósito de fornecer um recurso padrão para uso operacional e definições de negócios, facilitando a interoperabilidade semântica e o gerenciamento eficiente de informações dentro do setor financeiro (AMARAL *et al.*, 2021);
- b) na extração de regras de negócio em NL, a metodologia e as técnicas utilizadas para identificar, extrair e transformar regulamentos escritos em NL em regras candidatas serão eficazes em capturar as regras e enriquecê-las para a transformação para SBVR. O emprego de modelo probabilístico na classificação e extração das regras deve apresentar uma precisão e generalização melhor em relação aos métodos estudados do que as obtidas com soluções de NLP especialistas baseadas apenas em marcação de classes gramaticais (*Part-of-speech* - POS), reconhecimento de entidades nomeadas (*Named Entity Recognition* - NER) e árvore gramatical (*parser tree*) (MIN *et al.*, 2023; KULKARNI, 2023);
- c) o uso do grafo de conhecimento (*Knowledge Graph* – KG) deverá preservar a característica conectada inherente dos textos, permitindo que, na fase de transformação, o conhecimento das relações possibilite uma melhor interpretação do texto em questão (FILTZ, 2017);
- d) a transformação para a SBVR das regras candidatas em regras e vocabulário SBVR deverá ser precisa e manter a semântica original dos regulamentos, facilitando a sua implementação e verificação de conformidade em sistemas de TI. O emprego novamente de modelos

probabilísticos deve lidar com mais eficiência e generalização do que os modelos baseados em regras (PAN *et al.*, 2024).

1.4 Contribuições

Esta pesquisa contribui para o campo da conformidade regulatória financeira, conforme destacado a seguir, pela:

- a) identificação do estado da arte para abordagens de transformação de regras em linguagem natural para linguagem controlada;
- b) criação de um método automatizado para transformar textos de regras de negócio em linguagem natural para regras em SBVR;
- c) validação do método por um experimento, aplicando métodos estatísticos quantitativos para compará-la com um conjunto de dados de referência (padrão-ouro) gerados por esta pesquisa e métricas de similaridade semântica;
- d) discussão de como este trabalho pode ser aplicado na área de conformidade regulatória.

A pesquisa também contribui para a comunidade acadêmica e prática, disponibilizando instruções, dados, resultados e códigos para garantir a transparência e a reproduzibilidade da pesquisa.

1.5 Método de trabalho

O método de trabalho desta pesquisa é baseado no ciclo de projeto (*design cycle*) de Wieringa (2014), que discute a aplicação do *Design Science Research* (DSR) para a área de engenharia de software e sistemas de informação. Segundo o autor, o ciclo de projeto consiste nas seguintes tarefas:

- a) investigação do problema: quais fenômenos precisam ser melhorados? Por quê?
- b) projeto do tratamento: projetar um ou mais artefatos que possam tratar o problema;
- c) validação do tratamento: esses projetos tratarão o problema?

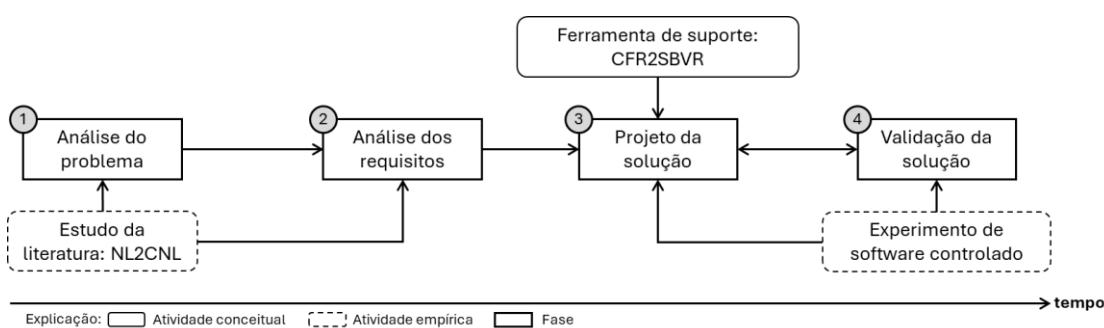
- d) implementação do tratamento: tratar o problema com um dos artefatos projetados;
- e) avaliação da implementação: quão bem-sucedido foi o tratamento?

No ciclo de engenharia (*engineering cycle*), a implementação, segundo a definição do autor, corresponde à aplicação do tratamento ao contexto original do problema. Assim, em vez de “implementação”, pode-se entender como “transferência para o contexto do problema”, interpretação adotada nesta pesquisa. Este estudo segue a abordagem de *Design Science Research* (DSR) e está restrito às três primeiras etapas do ciclo de engenharia: investigação do problema, concepção e validação do tratamento (WIERINGA, 2014).

Por ser mais usual para engenheiros de software, optou-se pelo uso do termo solução em vez de tratamento, no escopo desta pesquisa, entendendo os termos como sinônimos. Além disso, a pesquisa adapta o ciclo de projeto para quatro fases: (1) análise do problema (o mesmo que investigação do problema); (2) análise dos requisitos; (3) projeto da solução (o mesmo que projeto do tratamento) e (4) validação da solução (validação do tratamento), como ilustrado na Figura 1.

Na Fase 2, a elicitação dos requisitos foi realizada com base nos casos de uso das soluções analisadas na literatura da Fase 1, dado que o projeto não possui partes interessadas (*stakeholders*) externas. Assim, tornou-se necessário enfatizar os requisitos considerados na solução proposta.

Figura 1 - Método de trabalho



Fonte: Elaborado pelo autor.

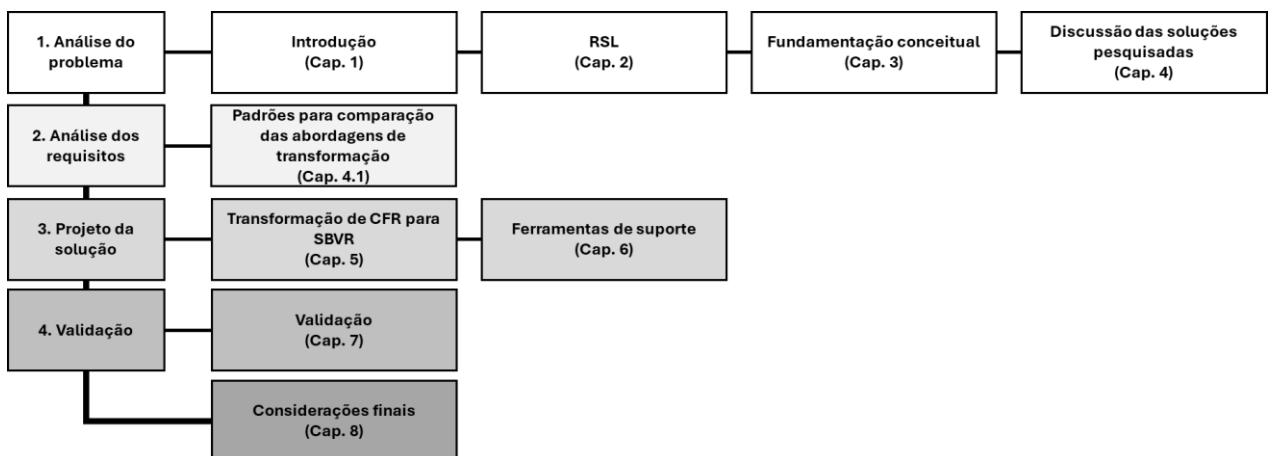
Figura 1 ilustra as fases da pesquisa que se inicia com a análise do problema a ser investigado (Fase 1), conduzida por uma revisão exploratória e sistemática da literatura. Baseando-se nessas atividades, bem como na experiência prática do autor no domínio financeiro, são derivados requisitos para a avaliação do projeto de engenharia (Fase 2). A partir dos requisitos identificados, desenvolve-se a abordagem de solução proposta (Fase 3). Finalmente, a solução proposta é validada por um experimento controlado de software (Fase 4). O ciclo de projeto desta pesquisa é caracterizado por uma única iteração.

Entende-se aqui por artefato um objeto ou construto feito pelo homem criado para abordar um problema específico ou alcançar um objetivo particular. Os artefatos podem ser tangíveis, como software, protótipos e ferramentas, ou intangíveis, como modelos, métodos, estruturas e algoritmos. O desenvolvimento, a avaliação e o refinamento de artefatos são atividades-chave na DSR, pois representam a aplicação do conhecimento teórico a desafios práticos (WIERINGA, 2014). Os artefatos produzidos por esta pesquisa são apresentados na seção “5.2 Principais artefatos”.

1.6 Organização do trabalho

Este trabalho está organizado em uma série de capítulos, cada um dedicado a explorar diferentes aspectos da pesquisa. A Figura 2 ilustra a organização do trabalho orientada pelas fases descritas anteriormente e serve de guia para a sua leitura.

Figura 2 - Organização do trabalho



Fonte: Elaborado pelo autor.

A Fase 1 do método de trabalho (análise do problema) é desenvolvida no Capítulo “1 INTRODUÇÃO” e no Capítulo “2 REVISÃO SISTEMÁTICA DA LITERATURA”. No Capítulo “3 FUNDAMENTAÇÃO CONCEITUAL”, são discutidos os trabalhos importantes para o entendimento desta pesquisa e no Capítulo “4 DISCUSSÃO DAS SOLUÇÕES” são analisados os trabalhos relacionados à transformação. A Fase 2 (análise dos requisitos) elenca os requisitos com base na discussão do Capítulo “4 DISCUSSÃO DAS SOLUÇÕES” e resume os requisitos na seção “4.1 Padrões para comparação das abordagens de transformação”. A Fase 3 desenvolve o projeto da solução no Capítulo “5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR”. O Capítulo “6 FERRAMENTAS DE SUPORTE” descreve as ferramentas construídas, o software de coleta de dados e a infraestrutura necessária para a validação. A Fase 4 compreende a estratégia de validação da solução de software elaborada no Capítulo “7 VALIDAÇÃO”, que descreve como a pesquisa experimental foi conduzida e analisa os resultados. O Capítulo “8 CONSIDERAÇÕES FINAIS” apresenta os dados e as informações obtidas durante a pesquisa, discute os impactos e a eficácia do processo de transformação em relação à literatura e as implicações práticas e teóricas dos achados e apresenta a conclusão para as perguntas de pesquisa e uma perspectiva sobre trabalhos futuros.

2 REVISÃO SISTEMÁTICA DA LITERATURA

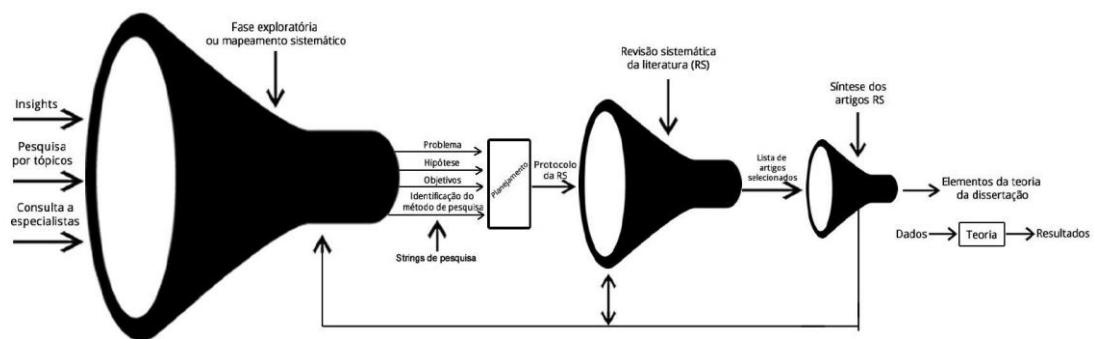
Esta seção tem como objetivo estudar a literatura existente para identificar o estado da arte na transformação de regras de negócio para uma linguagem controlada como SBVR. A revisão sistemática procura identificar estudos que possam auxiliar na resposta à questão principal da pesquisa, definida na seção “1.2 Declaração do problema” e repetida a seguir: Como a aplicação de técnicas de NLP e ontologias, como a FIBO, podem aprimorar a transformação automatizada de regulamentos financeiros em regras e vocabulário estruturados conforme a especificação SBVR?

2.1 Método

Para Biolchini *et al.* (2005) e Kitchenham (2004), a revisão sistemática é um protocolo de pesquisa para analisar e interpretar a literatura disponível sobre um determinado tópico, questão de pesquisa ou fenômeno de interesse. Segundo Nakagawa *et al.* (2017), a possibilidade de conferir, reproduzir e auditar a revisão sistemática é o seu principal diferencial.

Segundo Gava (2022), a revisão sistemática envolve as fases com refinamentos sucessivos desde o planejamento, condução e análise dos dados para publicação dos resultados. As três fases estão detalhadas na Figura 3.

Figura 3 - Fases do processo da Revisão Sistemática



Fonte GAVA (2022)

A revisão adotou essas fases. Na fase exploratória, os estudos de apoio à revisão sistemática foram identificados pela consulta bibliográfica de artigos publicados em congressos, periódicos e dissertações de mestrado (Biblioteca

IPT¹⁰), obtidas por consultas na base de dados eletrônica Google Scholar¹¹ e que tinham acesso livre. Foram identificados 12 artigos que foram julgados relevantes e representativos. A lista completa pode ser encontrada no “APÊNDICE A – RESULTADOS DA RSL”. O resultado dessa fase culminou na elaboração do Capítulo “1 INTRODUÇÃO” e na elaboração da RSL. As demais seções detalham as fases da RSL e sintetizam os artigos que são a base para os Capítulos “3 FUNDAMENTAÇÃO CONCEITUAL” e “4 DISCUSSÃO DAS SOLUÇÕES”.

2.2 Protocolo

As bases de dados escolhidas são reconhecidas academicamente pela produção literária de alta qualidade e pela inclusão dos principais periódicos e eventos científicos nos campos da engenharia de software e da ciência da computação, listadas no Quadro 1.

Quadro 1 - Fontes de busca para a revisão sistemática

Fonte de Busca	Endereço Online
ACM	https://dl.acm.org/
IEEE Explorer (IEEE)	https://ieeexplore.ieee.org/
ScienceDirect	https://www.sciencedirect.com/
Springer Link (Springer)	https://link.springer.com/
Web of Science (WoS)	https://www.webofscience.com/

Nota: A Scopus Digital Library (Scopus) estava inacessível durante o período de execução dessa RSL pelo acesso CAFé¹²

Fonte: Elaborado pelo autor.

O idioma selecionado foi o inglês, por ser o mais usado na produção de artigos e trabalhos acadêmicos para ciência da computação e engenharia de software. Foram identificadas as seguintes palavras-chave com base na pesquisa exploratória e nas perguntas da pesquisa, em língua inglesa: *finance, financial, natural language processing, nlp, ontology, semantics of business vocabulary*

¹⁰ Ver mais em: <http://aleph.upt.br/F>. Acesso em: 15/mar./2025.

¹¹ Ver mais em: <http://scholar.google.com/>. Acesso em: 15/mar./2025.

¹² Ver mais em: <https://www.periodicos.capes.gov.br/index.php/acesso-cafe.html>. Acesso em: 15/mar./2025.

and business rules, sbvr, controlled natural language, financial industry business ontology.

Em termos gerais, as *strings* específicas serão apresentadas na etapa de condução. A *string* genérica foi dividida em duas:

- a) *String* de busca para a pergunta PICO: (“financial” OR “finance”) “semantics of business vocabulary and business rules” OR “SBVR”);
- b) *String* de busca para perguntas de conhecimento: (“governance, risk and compliance”) AND (“financial” OR “finance”)

As *strings* de busca foram usadas, separadamente, para explorar diferentes aspectos da pesquisa e identificar artigos relevantes nas bases de dados. Os resultados são combinados para obter uma lista sem duplicações.

A busca automática foi realizada pela aplicação das *strings* de busca e suas variações nas bases de busca selecionadas, no período entre 2018 e 2024, incluindo apenas artigos de periódicos e anais. A busca manual foi realizada pela aplicação da técnica de *Snowballing* (GREENHALGH; PEACOCK, 2005; WOHLIN, 2014), que utiliza a lista de citações e referências dos estudos aceitos com o objetivo de identificar novos trabalhos.

Todos os trabalhos retornados pelas buscas estão documentados no “APÊNDICE A – RESULTADOS DA RSL”, bem como os critérios de inclusão e exclusão.

O protocolo da RSL foi dividido em duas avaliações. A primeira avaliação (A1) foi realizada a partir da leitura dos seus títulos, resumos, introdução, conclusão e palavras-chave, considerando os critérios de inclusão e exclusão definidos; em seguida, levando em conta os mesmos critérios, foi realizada uma segunda avaliação (A2), pela leitura completa dos estudos que passaram pela primeira avaliação. O Quadro 36 do “APÊNDICE A – RESULTADOS DA RSL” documenta as informações e registra quais critérios de inclusão e/ou exclusão foram atribuídos para cada um dos trabalhos avaliados.

2.3 Execução

A execução da RSL contou com o suporte da ferramenta stArt (UFSCAR, 2024) e todos os trabalhos retornados estão registrados no “APÊNDICE A – RESULTADOS DA RSL”.

Quadro 2 - *Strings* de busca para pergunta PICO nas bases de dados

Base de dados	Data da execução	String de busca ajustada	Tipo de documento	Total
ACM	2024-04-16	[All: “controlled natural language”] OR [All: “cnl”] OR [All: “sbvr”] OR [All: “semantics of business vocabulary and business rules”] AND [E- Publication Date: (01/01/2018 TO 12/31/2024)]	Research articles	17
IEEE	2024-04-16	(“controlled natural language” OR “cnl” OR “sbvr” OR “semantics of business vocabulary and business rules”) You Refined By: Year: 2018-2024	Livros, jornais e conferências	11
ScienceDirect	2024-04-16	(“controlled natural language” OR “sbvr” OR “semantics of business vocabulary and business rules”) Year: 2018-2024	Artigos de revisão, artigos de pesquisa e capítulo de livros	14

Base de dados	Data da execução	String de busca ajustada	Tipo de documento	Total
Springer	2024-04-16	'financ*' AND ("sbvr" OR "semantics of business vocabulary and business rules")' within 2018 - 2024	Capítulos, conferências e artigos	30
WoS	2024-04-05	("financial" OR "finance") AND ("nlp" OR "natural language processing") AND ("ontology" OR "financial industry business ontology" OR "controlled natural language" OR "semantics of business vocabulary and business rules" OR "SBVR") (All Fields) and 2018 or 2019 or 2020 or 2021 or 2022 or 2023 (Publication Years)	Anais de eventos e artigos	37

Fonte: Elaborado pelo autor.

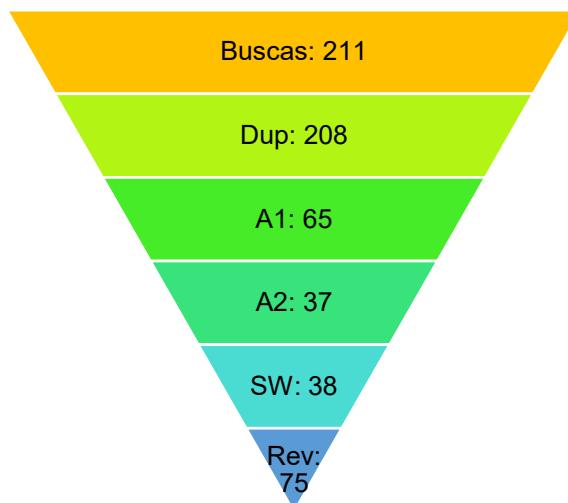
Quadro 3 - *Strings* de busca para pergunta de conhecimento nas bases de dados

Base de dados	Data da execução	Amplitude da busca	String de busca	Tipo de documento	Total
ACM	2024-04-16	Todos os metadados	[[All: "governance, risk and compliance"] OR [All: "grc"]]] AND [All: financ*] AND [E-Publication Date: (01/01/2018 TO 12/31/2024)]	Research articles	13
IEEE	2024-04-16	Todos os metadados	("governance, risk and compliance" OR "grc") AND financ* You Refined By: Year: 2018-2024	Livros, jornais e conferências	14
Science Direct	2024-04-16	Título, resumo e palavras-chave	("governance, risk and compliance") AND ("financial" OR "finance") Year: 2018-2024	Artigos de revisão, artigos de pesquisa e capítulo de livros	48
Springer	2024-04-16	Todos os metadados	'financ*' AND ("governance, risk and compliance")' within Article, 2018 - 2024	Capítulos e artigos	18
WoS	2024-04-16	Todos os metadados	("governance, risk and compliance") AND ("financial" OR "finance") (All Fields) and 2018 or 2019 or 2020 or 2021 or 2022 or 2023 (Publication Years)	Anais de eventos e artigos	9

Fonte: Elaborado pelo autor.

O Quadro 2 sumariza o resultado da pesquisa nas bases de dados para a pergunta PICO (*string* 1) e o Quadro 3 para as perguntas de conhecimento (*string* 2), apresentando-os por base de dados, data da execução, período de busca, *string* de busca específica e tipo de documento.

Figura 4 - Execução da RSL



Fonte: Elaborado pelo autor.

A Figura 4 ilustra o resultado quantitativo da execução da RSL. O total de artigos recuperados na pesquisa das bases de dados foram 211; após a remoção de duplicações, o total foi de 208. Na aplicação do protocolo, a primeira análise (A1) reduziu o total de artigos para 65. O destaque para o alto número de rejeições foi a *string* de busca 1 (SBVR) nas bases de dados WoS (18%) e Springer (11%) e a *string* de busca 2 (GRC) na Science Direct (30%), respondendo por 60% das rejeições. Na análise final (A2), o total de artigos foi 37. Foram adicionados 38 artigos das referências encontradas nos artigos lidos em A2. Esses artigos são listados no Quadro 36 do Apêndice A. Por fim, o total de 75 artigos foram considerados e compõem o corpus de conhecimento discutido nas demais seções deste trabalho.

Com base nos dados bibliométricos dos artigos selecionados, os autores mais influentes para esta pesquisa são: Bajwa, Imran Sarwar, Danenas, Paulis; Skersky, Tomas; Haj, Abdellatif; Sunkle, Sagar; Abi-lahoud, Elie, obtida com base no cálculo de coautoria do stArt.

3 FUNDAMENTAÇÃO CONCEITUAL

Nesta seção, serão tratados os principais conceitos que formam a base do conhecimento para o entendimento desta pesquisa.

3.1 Abordagens formais de verificação de conformidade

Racz *et al.* (2010) apresentam um quadro de referência para a pesquisa de GRC, com base na revisão sistemática da literatura. Os autores analisaram definições existentes e resultados de pesquisas anteriores para derivar uma definição única de GRC. Essa definição foi então avaliada e aprimorada por uma pesquisa entre profissionais da área. A definição final fornecida pelos autores e adotada nesta pesquisa é:

GRC é uma abordagem integrada e holística para a governança, risco e conformidade em toda a organização, garantindo que a organização aja de forma éticamente correta e de acordo com seu apetite ao risco, políticas internas e regulamentos externos, por meio do alinhamento de estratégia, processos, tecnologia e pessoas, melhorando assim a eficiência e eficácia (RACZ *et al.*, 2010, p. 112, tradução do autor deste trabalho).

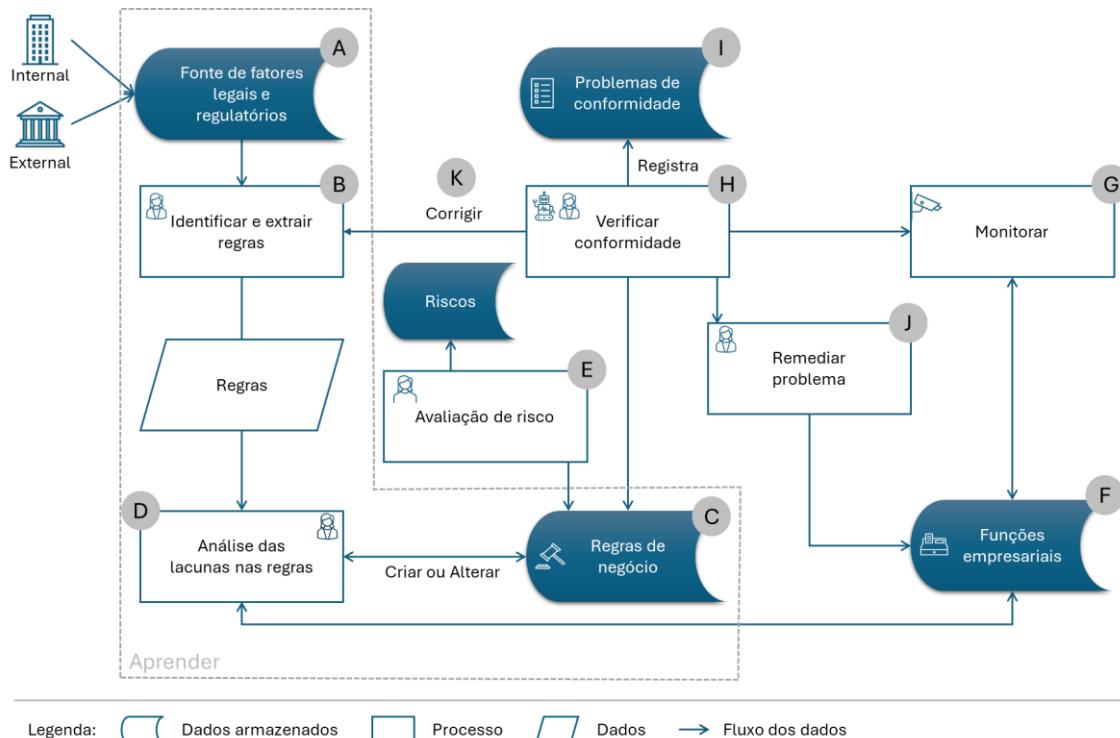
Esse quadro de referência destaca quatro componentes principais: estratégia, processos, tecnologia e pessoas, focados na governança, gerenciamento de riscos e conformidade. A tecnologia é essencial, aprimorando o gerenciamento de dados, automação de processos e avaliação de riscos, facilitando o monitoramento de riscos em tempo real e a integração de sistemas para uma gestão de GRC mais ágil e eficaz.

Abdullah (2019) discute a GRC como uma estratégia integrada essencial para supervisionar pessoas, processos e tecnologia em organizações. Esta abordagem holística visa agregar valor ao negócio, estabelecendo uma estrutura que permite gerenciar riscos de forma proativa. A eficácia de GRC é evidenciada na melhoria da eficiência organizacional, decisões baseadas em risco, agilidade nos negócios, conformidade regulatória, bem como na proteção da reputação e redução de custos, pela prevenção de penalidades e mitigação de riscos.

Contudo, persistem os desafios como resistência à mudança, a necessidade de processos consistentes de avaliação de risco, governança e custos tecnológicos altos. Superar esses obstáculos requer uma estratégia cuidadosa, que inclui comunicação clara, treinamento abrangente e gestão eficaz de mudanças.

Para Sunkle *et al.* (2015a), os desafios para GRC incluem a natureza semiformal das soluções, a forte dependência de especialistas, a falta de formalismo, a abordagem baseada em documentos, a complexidade dos textos legais, a falta de mapeamento explícito e a disparidade terminológica entre regulamentações e operações. A ausência de uma conexão clara entre os conceitos dos textos regulatórios e os termos usados nas operações das empresas tornam difícil: identificar onde as regulamentações se aplicam, a garantia de conformidade, a automação de verificações de conformidade, a adaptação a mudanças regulatórias e a comunicação eficaz com partes interessadas.

Figura 5 - Fluxo conceitual para conformidade, baseado no livro vermelho GRC (OCEG, 2023) combinado com processo de três etapas para gerenciar a conformidade (RACZ *et al.*, 2010)



Fonte: Elaborado pelo autor.

Os autores propõem o uso do SBVR combinado com medidas de similaridade semântica para criar mapeamentos explícitos que melhorem o alinhamento entre

conceitos regulatórios e operacionais, oferecendo uma abordagem mais formal e estruturada. Eles destacam que, enquanto as abordagens acadêmicas assumem a existência desse mapeamento sem fornecer ferramentas para defini-lo, as soluções de GRC na indústria frequentemente recorrem a taxonomias semiautomatizadas que ainda dependem de intervenção especializada e carecem de formalismo.

A Figura 5 ilustra um processo de GRC dividido em três etapas principais:

- a) identificação e extração: Na “Fonte de Fatores Legais e Regulatórios” (A), as regras regulatórias relevantes são identificadas e extraídas (B) para criar o conjunto de “Regras de Negócios” (C) específicas para o domínio. Isso é feito a partir da “Análise das lacunas nas regras” (D), em que as diferenças entre as políticas atuais da empresa e as novas regulamentações são avaliadas. Esse passo é importante para entender em que parte não há conformidade com os regulamentos mais recentes;
- b) identificação de riscos: “Avaliação de Risco” (E) ocorre após a compilação das regras, nas quais as lacunas foram identificadas. Novas regras ou mudanças podem acarretar riscos para o negócio e precisam ser identificadas e medidas para serem priorizadas e mitigadas na operacionalização das regras;
- c) operacionalização das regras: As “Regras de Negócio” (C) extraídas são então operacionalizadas em “Funções empresariais” (F) para garantir que as práticas empresariais estejam alinhadas com as políticas de conformidade atualizadas. Ela inclui o monitoramento (G) contínuo das “Funções empresariais” (F) e a verificação de conformidade (H) para garantir que os problemas de conformidade (I) sejam identificados e remediados (J) ou que as regras de negócios sejam revisitadas (K), fechando o ciclo de GRC e permitindo que a organização aprenda e se adapte continuamente às mudanças no ambiente regulatório.

O diagrama da Figura 5 representa um ciclo de *feedback* contínuo, no qual o aprendizado e a melhoria são integrados em todas as etapas, refletindo a

natureza dinâmica e iterativa do processo de GRC. Essa pesquisa desenvolve uma abordagem para ser integrada as atividades (A) e (B) da primeira etapa, pré-requisito para as demais atividades, ficando as demais como trabalhos futuros.

3.1.1 Visão geral das soluções comerciais para a GRC

Esta pesquisa não tem como objetivo comparar as soluções comerciais de GRC nem obter o estado-da-arte dessas soluções, mas oferece uma visão geral sobre as soluções disponíveis e as suas principais características, examinando como esses produtos incorporam a tecnologia de IA. O Forrester Wave™ é um guia para compradores com análises de produtos comerciais, usadas nas empresas para o processo de seleção de software, semelhante ao Quadrante Mágico do Gartner¹³. As análises e as opiniões são temáticas e frequentemente usadas pelos fornecedores dos produtos para se promoverem. Esta pesquisa não teve acesso às publicações da Forrester e usou fragmentos publicados por fornecedores para identificar a lista de produtos disponíveis no mercado. Essa lista foi obtida do relatório Forrester Wave™ GRC Q4 2023 disponibilizado pela MetricStream¹⁴.

A metodologia aplicada e os dados obtidos estão disponíveis no “APÊNDICE B – FORNECEDORES DE SOLUÇÕES PARA A GRC”. O resultado não foi surpreendente. Todos os fornecedores anunciam algum tipo de IA em seus produtos, porém, apenas dois fornecedores (1 e 7) descreveram o uso de IA para inferência sobre as regras de negócio, semelhante aos trabalhos avaliados nesta pesquisa, embora as informações publicadas *on-line* sejam limitadas.

As empresas avaliadas formam o principal grupo de fornecedores, segundo o Forrester Wave™, e empregam IA nas suas soluções para resolver problemas mais simples do que os propostos pela academia. Na opinião dos autores, isso reafirma a relevância da presente pesquisa para comunicar os seus achados

¹³ Ver mais em: <https://www.gartner.com.br/pt-br/metodologias/magic-quadrant>. Acesso em: 15/mar./2025.

¹⁴ Ver mais em: <https://www.metricstream.com/pressNews/forrester-wave-report-grc-platform.html>. Acesso em: 15/mar./2025.

para a indústria. Como trabalhos futuros, uma pesquisa estruturada com fornecedores pode ser conduzida.

3.2 Semantic for Business Vocabularies and Business Rules (SBVR)

O SBVR é um padrão desenvolvido pela OMG para fornecer uma maneira formal e inequívoca de definir o vocabulário e as regras específicas de um determinado domínio de negócios. O SBVR é projetado para ser usado por analistas de negócios, especialistas no assunto (*Subject Matter Experts - SMEs*) e outros interessados para capturar e documentar a semântica dos negócios, de uma maneira que seja legível por humanos e processável por máquinas. O objetivo do SBVR é permitir uma comunicação clara entre negócios e TI, melhorando, assim, o alinhamento das operações dos negócios com os seus objetivos e facilitando uma tomada de decisão mais eficaz (OMG, 2019).

3.2.1 Visão geral

Segundo a especificação do SBVR da OMG (2019), o SBVR fornece uma estrutura para definir e intercambiar vocabulários e regras de negócios entre organizações e ferramentas de software. Ele se baseia na lógica formal e linguística, permitindo definições unívocas, centradas no significado, multilíngues e semanticamente ricas de conceitos e regras do negócio. Os seus pontos-chave incluem:

- a) escopo e aplicabilidade: SBVR é aplicável a todos os tipos de atividades e organizações de negócios, fornecendo uma maneira de definir os significados dos conceitos e regras de negócios, de forma unívoca e consistente;
- b) conformidade: a especificação define a conformidade para o software que implementa o SBVR e para documentos de intercâmbio do modelo de conteúdo SBVR. A conformidade é categorizada em conformidade com a sintaxe abstrata, conformidade de intercâmbio de dicionário terminológico e/ou livro de regras, bem como a conformidade com a semântica SBVR;
- c) fundamentos linguísticos: o triângulo semiótico/semântico é introduzido como a base teórica para a estrutura do SBVR, que separa a

representação do significado. Conceitos-chave incluem significado, coisa, enunciado e as suas relações;

d) tipos de significado: os significados são classificados em conceitos (significados de palavras ou sinais) e proposições (significados de declarações);

e) tipos de declarações (*statements*): as declarações são categorizadas em designações (declarações que significam conceitos) e formulações (declarações que significam proposições);

f) conexões entre conceitos e coisas: extensões (conjuntos de coisas que são instâncias de um conceito) e instâncias (coisas específicas que são exemplos de um conceito);

g) conexões entre tipos de significado e estados de coisas: como as proposições (significados de declarações) se relacionam com situações do mundo real ou estados de coisas. Uma proposição é considerada verdadeira se descrever com precisão um estado de coisas que realmente existe. Um exemplo é a proposição "*Inventory levels are sufficient*" que é verdadeira se os níveis reais de estoque atenderem aos padrões necessários. Da mesma forma, a proposição "*The store is open*" deve corresponder ao status operacional real da loja para ser verdadeira. Nas potencialidades, as proposições como "*We could increase sales by 20% if we extended our hours*" descrevem cenários possíveis sob condições específicas. Além disso, são abordados papéis e orientações dentro dos contextos de negócios, como "*Employees must secure approval for expenses over \$500*", garantindo que as regras e as políticas de negócios sejam diretamente aplicáveis às operações e governança dentro de uma organização;

h) conexões entre declarações e coisas no negócio: aborda como as declarações estão relacionados a coisas no negócio, incluindo a identificação e desambiguação. As declarações se relacionam, então, com os objetos, ações ou conceitos que representam em um contexto de negócios, garantindo uma reflexão precisa das operações de negócios;

i) necessidades relacionadas à extensão: descreve regras sobre a extensão de conceitos, como a necessidade de cada instância de um conceito ser uma instância do seu conceito geral.

Esses pontos-chave permitem que a linguagem seja semanticamente rica e logicamente precisa para definir vocabulário e regras de negócios. Como será discutido na próxima seção, o SBVR não é em si uma CNL, mas fornece uma base para criá-la para domínios de negócios.

A especificação SBVR (OMG, 2019) faz distinção entre dicionário e vocabulário. O dicionário terminológico é definido como a parte de um modelo de conteúdo SBVR que documenta e organiza sistematicamente o vocabulário usado no negócio. Isso inclui fornecer definições precisas, diretrizes de uso e as relações entre os termos. O dicionário terminológico é especificamente projetado para garantir consistência e precisão no uso dos termos de negócios em várias partes de uma organização. Ele contém definições detalhadas de termos para garantir que não haja ambiguidade sobre os seus significados, organiza os termos de maneira estruturada, mostrando relações como sinônimia, antônimia e classificações hierárquicas, servindo como uma ferramenta de referência para qualquer pessoa na organização que precise entender os significados específicos dos termos, conforme são, oficialmente, reconhecidos e usados nas operações de negócios.

Por outro lado, o vocabulário de negócios é o conjunto abrangente de termos (incluindo substantivos, verbos e adjetivos) usados para expressar os conceitos pertinentes a um negócio ou indústria específicos. Esse conjunto de termos encapsula a linguagem usada nas interações diárias de negócios e processos operacionais. Ele cobre todos os elementos linguísticos necessários para descrever e conduzir negócios dentro de um domínio específico ou em toda a empresa, sendo ativamente usado nas comunicações diárias de negócios, documentação e dentro dos processos de negócios, tornando-o dinâmico e contextual e sendo projetado para facilitar uma comunicação clara e eficaz em diferentes áreas dele.

A especificação é bastante extensa e não é intenção desta seção substituí-la.

3.2.2 Linguagem Controlada

Ashfaq e Bajwa (2021) discutem a ambiguidade na linguagem natural em documentos de especificação de requisitos de software (*Software Requirements Specification* - SRS) e como isso pode causar mal-entendidos e erros no desenvolvimento de software. Eles exploram várias CNLs que visam reduzir essa ambiguidade. Entre elas, o *Attempto Controlled English* (ACE) é usado para requisitos de software e consultas de banco de dados, devido à sua legibilidade e fácil interpretabilidade por máquinas. O *Processable English* (PENG) permite a redação clara de especificações para processamento automático. O *Controlled Language for Ontology Editing* (CLOnE) é adequado para a criação e edição de ontologias e o Formalized-English é, frequentemente, aplicado em contextos legais e técnicos para expressar informações de forma estruturada. Os autores favorecem o uso do inglês estruturado baseado no SBVR (*Semantic for Business Vocabularies and Business Rules - Structured English* - SBVR-SE), uma CNL projetada para articular regras e vocabulários de negócios, por ser uma estrutura padronizada que facilita tanto a compreensão humana quanto o processamento por máquinas. Eles destacam que o conhecimento armazenado no vocabulário padronizado do SBVR é reutilizável em diferentes projetos dentro do mesmo domínio, levando a potenciais ganhos de eficiência.

Kowalski e Datoo (2021) desenvolvem uma abordagem para padronizar e automatizar documentos legais, especificamente contratos financeiros, pela introdução do *Logical English* (LE). LE é uma CNL, legível por humanos e executável por máquina, agindo como intermediária entre o inglês jurídico e as linguagens de programação. Fundamentada nos princípios da programação lógica, LE incorpora variáveis tipificadas e quantificadas, mantendo a estrutura padrão de programas lógicos com enunciados atômicos e condicionais, usando sentenças em inglês controlado em vez de símbolos. Enquanto LE se concentra na automação de documentos legais no setor financeiro, SBVR abrange uma variedade maior de domínios, oferecendo um padrão formal para expressar regras de negócios e vocabulários, útil em modelagem de processos de negócios e gestão de conformidade. Ambas as abordagens visam facilitar a transição entre

NL e lógica formal, mas LE é, especificamente, voltada para a automação legal no setor financeiro e não é padronizada.

Para Roychoudhury *et al.* (2017), especialistas do domínio, frequentemente, encontram desafios ao escreverem regulamentações em uma especificação formal, devido a uma desconexão entre o seu conhecimento e as ferramentas disponíveis. Os autores visaram identificar uma linguagem de especificação que os especialistas do domínio adotariam facilmente, especialmente com o suporte de ferramenta adequado. Essa linguagem também deveria ser facilmente transformável em uma especificação formal, funcionando como um intermediário. A linguagem escolhida pelos autores foi o inglês estruturado especificado no ANEXO C do SBVR (OMG, 2019).

O SBVR é um padrão que especifica um metamodelo para definir vocabulários e regras de negócios de uma forma que seja compreensível para humanos e interpretável por máquinas. Uma CNL baseada em SBVR usa um vocabulário e uma gramática restritos e derivados da NL para expressar regras e conceitos de negócios de maneira clara e inequívoca. A especificação da OMG sugere o uso de Inglês Estruturado (*Structured English* - SE) para expressar regras de negócios.

Embora os exemplos na especificação não sejam prescritivos, eles demonstram como uma linguagem controlada pode ser construída usando as estruturas semânticas do SBVR. A especificação inclui um anexo não normativo que fornece diretrizes para o SBVR-SE, oferecendo uma sintaxe e vocabulário sugeridos para expressar vocabulário de negócios e regras de maneira estruturada e consistente. Embora as diretrizes no anexo C não sejam obrigatórias, elas servem como uma referência útil para criar regras de negócios e definições de vocabulário claras e inequívocas.

Esta pesquisa utiliza o inglês estruturado descrito por Witt (2012), que segue as formulações do RuleSpeak¹⁵, uma variante semelhante que influenciou a versão

¹⁵ RuleSpeak™ é um conjunto de diretrizes para expressar regras de negócios de forma concisa e amigável aos negócios, desenvolvida por Ronald G. Ross, em 1996 (Ver mais em: <https://www.rulespeak.com/en/>. Acesso em: 15/mar./2025).

apresentada no anexo C da especificação da OMG (2019). O autor critica certas formulações complexas, como as do anexo C, e propõe alternativas mais acessíveis, visando maior clareza e consistência na definição de regras. A escolha do SBVR e do inglês estruturado baseia-se em sua precisão e clareza na expressão de regras de negócio, bem como na proximidade com a linguagem natural. Além disso, sua ampla adoção na academia contrasta com o que se observa nas ferramentas comerciais de GRC, sugerindo que ainda existem barreiras para sua disseminação.

3.2.3 Termos, fatos e regras

Witt (2012) discute como as regras de negócios, fatos e termos se interrelacionam. Uma regra de negócios é uma declaração que define ou restringe um aspecto de um negócio, com o objetivo de influenciar ou controlar o comportamento da organização. Uma controvérsia na definição de regras de negócios surge da questão de incluir ou não regulamentos externos, como leis e normas da indústria, como parte dessas regras. A visão mais ampla inclui essas restrições externas, enquanto o padrão SBVR, mais restrito, as exclui, focando apenas nas regras que as empresas têm autoridade para modificar ou descartar.

Em termos de relacionamento, os fatos, as regras e os termos interagem de forma hierárquica. Os fatos descrevem as relações entre termos, que representam conceitos de negócios. As regras sobre esses fatos impõem restrições ou diretrizes sobre o comportamento da organização. Tanto os fatos quanto as regras são compostos por termos, o que faz dos termos os blocos fundamentais da linguagem de negócios. A frase “termos expressam conceitos de negócios; fatos fazem afirmações sobre esses conceitos; regras restringem e apoiam esses fatos” (artigo 3, parágrafo 2 do manifesto sobre regras de negócios¹⁶) resume essa hierarquia, em que os termos definem os conceitos, os fatos expressam as relações entre esses termos e as regras governam as ações (Witt, 2012).

Um fato é uma afirmação específica sobre o mundo que pode ser verdadeira ou falsa, como "João é um funcionário da ABC Corp". Um tipo de fato é uma

¹⁶ Ver mais em: <https://www.businessrulesgroup.org/brmanifesto.htm>. Acesso em: 15/mar./2025.

estrutura geral, um modelo, que define uma categoria de fatos, especificando relacionamentos ou propriedades, como "Funcionário trabalha para Empresa". Esse modelo apoia a criação e o gerenciamento de regras, fornecendo um mapa claro dessas relações. Embora Witt (2012) não defina explicitamente um modelo, ele implica tal modelo pela sua abordagem estruturada de categorização e gerenciamento de regras de negócios. Na especificação SBVR (OMG, 2019), um tipo de fato é uma representação conceitual de um domínio de negócios expressa em termos de fatos e regras. Ele consiste em um conjunto de enunciados que capturam tanto regras de negócios quanto “fatos fundamentais” - fatos relacionados a uma instância específica ou a um conjunto de instâncias dentro do domínio de negócios. Esses fatos são articulados usando lógica de predicados, mapeando tipicamente indivíduos ou entidades para relacionamentos ou propriedades.

As regras são classificadas em duas categorias, segundo Witt (2012), (1) as regras de definição que especificam como os termos ou conceitos de negócios devem ser definidos, incluindo os fatos e tipos de fatos; (2) e as regras operativas que são subdivididas em: regras de dados que restringem a integridade dos dados e os processos de entrada; as regras de atividades que governam os processos, comportamentos e a tomada de decisões no negócio; e as regras de papéis que especificam quais funções ou indivíduos podem realizar determinadas ações.

Nesta pesquisa, são extraídos dos textos regulatórios as regras de definição e as operativas.

3.2.4 Taxonomia para as regras de negócio

Não há uma taxonomia padrão para as regras de negócios, autores como Ross (1997) propõe classificar as regras em obrigatoriedade (utilizando a palavra “must”), proibição (palavra “must not”), permissão condicional (“may not”) e aconselhamento (“may” e “need not”). A especificação SBVR separa as regras em comportamentais (operativas), que compreende as regras de obrigação, permissão e proibição, bem como de definição (ou estruturais), que define como as entidades do negócio são caracterizadas e relacionadas entre si.

Quadro 4 - Classificação das regras segundo Von Halle (2001)

Nome do modelo	Definição	Modelo de escrita
Termo	Um termo é definido como um substantivo ou frase nominal, representando um conceito, propriedade, valor ou conjunto.	<Termo1> É DEFINIDO COMO <texto>
Fato	Conecta termos para formar observações significativas sobre relacionamentos ou atributos de negócios.	<Termo1> É UM <Termo2>, <Termo1> É COMPOSTO DE <Termo2>, <Termo1> TEM UMA PROPRIEDADE DE <Termo2>
Cálculo	Representa um algoritmo ou operação numérica para determinar um valor.	<Termo1> É CALCULADO COMO <fórmula>
Restrição Obrigatória	Condições obrigatórias para as informações de entrada.	<Termo1> DEVE TER <pelo menos, no máximo, exatamente n de> <Termo2> <Termo1> DEVE SER <comparação> <Termo2> <Termo1> DEVE ESTAR NA LISTA <a, b, c>
Diretriz	Expressa condições que idealmente deveriam ser verdadeiras para as informações de entrada, semelhantes às restrições, mas menos rigorosas.	<Termo1> DEVERIA TER <pelo menos, no máximo, exatamente n de> <Termo2> <Termo1> DEVERIA ESTAR NA LISTA <a, b, c>
Conhecimento Inferido	Define como novas informações são inferidas, com base em determinadas condições.	SE <Termo1> <operador> <Termo2>, valor, lista de valores> E <novamente ENTÃO <Termo3> <operador> <Termo4>
Habilitador de Ação	Condições que levam a um evento externo.	SE <Termo1> <operador> <Termo2> ENTÃO <ação>

Fonte: Elaborado pelo autor.

Von Halle (2001, p. 39) introduz modelos (*templates*) de regras como padrões estruturados usados para representar regras de negócios, combinando cláusulas de regras, geralmente na forma de “<termo1> <operador> <termo2>”, em que os termos são conceitos definidos e os operadores os conectam logicamente. Esse método proporciona uma maneira formalizada de expressar lógica de negócios complexa, facilitando a definição e a automação das regras de forma consistente. O autor define os modelos, listados no Quadro 4, que são modelos bastante genéricos, se comparados aos de Witt (2012).

A especificação SBVR (OMG, 2019) e Von Halle (2001, p. 39) tratam de regras do tipo conselho (*advice* ou *guideline*). Essas regras são elementos de orientação que reconhecem ou permitem um determinado estado de coisas, oferecendo segurança ou confirmando um grau de liberdade, em vez de impor restrições. Esses conselhos incluem o conselho de possibilidade, que reconhece um estado possível implícito por regras de definição, o conselho de contingência, que evita a suposição equivocada de regras inexistentes, e o conselho de permissão, que afirma que um comportamento específico é permitido. Essencialmente, os conselhos funcionam como “não-regras”, enfatizando ações permitidas e evitando restrições não intencionais nas práticas empresariais.

Witt (2012) classifica as regras de conselho em dois tipos principais: declarações de permissão, que indicam que uma determinada ação é permitida, mas não obrigatória, e declarações de possibilidade, que esclarecem situações em que as ações são viáveis sem restrições explícitas. A taxonomia das regras divide todas as regras de negócios em categorias, como regras de definição, de dados, de atividade e de papéis, sendo que as regras de conselho fornecem uma clarificação suplementar na ausência de proibições ou requisitos explícitos.

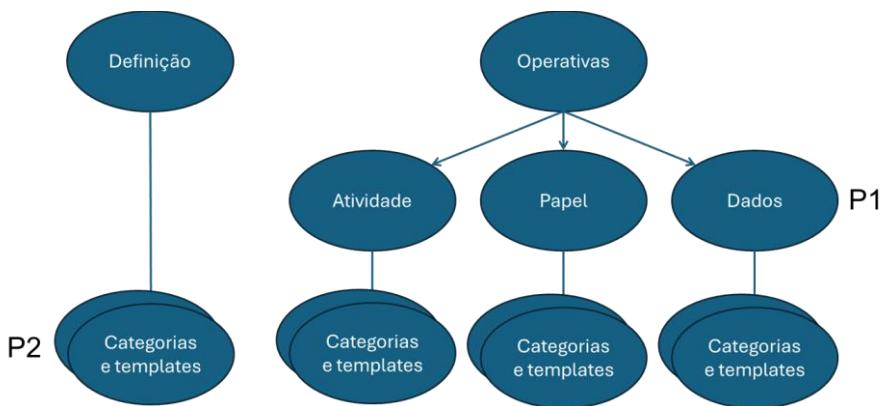
A especificação SBVR (OMG, 2019) define dois tipos principais de regras: regras de definição e regras comportamentais. As regras de definição, também conhecidas como regras estruturais, definem características essenciais de conceitos e/ou relacionamentos, estabelecendo condições que são sempre verdadeiras, como atributos inerentes a um conceito. As regras comportamentais, também chamadas de regras operativas, obrigam ou restringem ações, especificando comportamentos aceitáveis ou inaceitáveis,

governando diretamente a conduta, a partir de declarações executáveis, como obrigações, proibições ou permissões. Assim, o SBVR diferencia as regras que definem propriedades intrínsecas das que regulam comportamentos.

Seguindo a especificação SBVR, as formulações de regras podem ser classificadas em várias categorias, incluindo formulações lógicas e projeções. As formulações lógicas são usadas para estruturar proposições, sejam simples ou complexas e podem incluir operações como quantificações, conjunções, disjunções e outros construtos lógicos. Exemplos disso inclui formulações atômicas (declarações indivisíveis que envolvem conceitos verbais), formulações de instância (relacionadas à instanciação de conceitos) e formulações modais (expressando necessidade, obrigação ou possibilidade).

Projeções são outro tipo de formulação usada para estruturar definições, agregações e perguntas, essencialmente organizando um conjunto de itens que atendem a restrições específicas. Além disso, as objetificações e as nominalizações proposicionais são usadas para transformar estados ou eventos em sujeitos ou objetos dentro de outras formulações, permitindo articular regras mais complexas temporais ou relacionais. Embora bastante abrangente, o SBVR não prescreve nenhum modelo de escrita.

Figura 6 - Taxonomia de Witt (2012)



Fonte: Elaborado pelo autor.

Witt (2012, p. 84-99) define uma taxonomia alinhada com a classificação do SBVR para regras de definição e comportamento. No topo da taxonomia de Witt, estão as regras de definição, que correspondem as regras de definição do SBVR

e ao lado as regras de dados, atividades e papéis, sendo essas regras comportamentais ou operativas. Parte da taxonomia é apresentada na Figura 6 e a taxonomia completa se encontra no “APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO”.

Essa taxonomia se relaciona com os conceitos do SBVR de formulações lógicas e projeções, ao categorizar a maneira como as regras de negócios e significados são estruturados. Ela fornece uma classificação para a organização de termos, relacionamentos e lógica, assim como o uso das Formulações Lógicas no SBVR para definir proposições, incluindo operações como conjunções, disjunções e quantificações.

As projeções no SBVR servem como um método para estruturar definições, agregações e questões, organizando conjuntos que atendem a restrições específicas. De forma semelhante, a taxonomia de Witt trata da categorização com base em características comuns ou distintivas, proporcionando uma base estrutural para o vocabulário de negócios, que se alinha à categorização formal e projeção encontradas no SBVR. Essencialmente, ambas as classificações visam organizar termos, papéis e restrições para garantir representações claras e abrangentes da semântica dos negócios.

Os *templates* de Witt (2012) para fatos e regras visam melhorar a consistência na forma como eles são estruturados e enunciados. Um exemplo disso é que há *templates* para termos que ajudam a padronizar a definição de conceitos de negócios, como o T7, que é usado para definições intensionais formais. Ademais, existem *templates* de fatos, como o F10 (“pessoa prepara refeição”) ou F31 (“cliente faz pedido”), que definem as relações entre os termos. Os *templates* de regras garantem que as regras sejam expressas de forma clara e consistente, reduzindo a ambiguidade e melhorando a comunicação entre as equipes de negócios e técnicas. Eles ajudam a formalizar lógicas complexas, tornando-as comprehensíveis e reutilizáveis por diferentes partes interessadas, incluindo desenvolvedores, analistas e sistemas automatizados. O uso da taxonomia de Witt (2012) nesta pesquisa se dá pelo alinhamento com a especificação SBVR, pela sua abrangência, com mais de 60 categorias e pelos *templates* de escrita associados a cada categoria.

Como discutido nesta seção, não existe um padrão de jure para a classificação das regras de negócio, sendo a sua escolha dependente do propósito. Nesta pesquisa, o objetivo é identificar padrões de regras e associá-los a *templates* de escrita a fim de obter um resultado consistente na sua formulação, em que a taxonomia de Witt (2012), ao combinar a classificação e um *template* de escrita, atende a esse propósito.

3.2.5 Modelo de conteúdo

O modelo de conteúdo SBVR (OMG, 2019) é um modelo formal baseado em XML, que permite a representação e a troca de vocabulários e regras de negócios entre sistemas e organizações, garantindo a consistência semântica e eliminando ambiguidades. Ele facilita a interoperabilidade, ao definir de maneira clara os termos e as regras de negócios, sendo útil tanto para profissionais quanto para sistemas automatizados. Com suporte multilingüístico e a capacidade de transformar definições em regras executáveis, o modelo ajuda as empresas a documentarem e automatizarem os seus processos, mantendo conformidade e integração eficiente em diferentes ambientes empresariais.

Ele comprehende duas partes principais: o dicionário terminológico e o livro de regras. O dicionário terminológico documenta o vocabulário de negócios, definindo termos, conceitos e outras representações usadas no contexto empresarial. Já o livro de regras contém orientações comportamentais, incluindo regras de negócios de definição e comportamentais. O modelo utiliza o formato XML e é estruturado de acordo com o esquema SBVR XMI (XML Metadata Interchange), permitindo a interoperabilidade entre diferentes sistemas de software e a troca de vocabulários e regras de negócios (OMG, 2019).

O modelo não é especificado como uma ontologia, mas como um modelo conceitual que documenta vocabulários e regras de negócios, focado na definição de significados e as suas representações. Embora capture toda a semântica necessária para transformações em modelos de tecnologia da informação, ele não é usado para raciocínio sobre dados, algo típico de ontologias formais. O seu objetivo principal é garantir a interoperabilidade e a

clareza dos conceitos e regras de negócios entre sistemas, sem atuar como uma ontologia no sentido computacional.

O modelo usa o MOF (Meta-Object Facility) como sintaxe para representar a semântica do SBVR, garantindo que o modelo resultante não seja uma metamodelo da MOF ou UML, mas uma documentação dos significados de termos e regras usadas nos contextos de negócios. A especificação do SBVR afirma que a reflexão MOF (a capacidade de MOF de refletir sobre si mesma como um metamodelo) é explicitamente proibida no uso do SBVR. Isso reforça que o metamodelo SBVR MOF é apenas uma representação dos significados de termos e regras e não uma ferramenta para raciocínio semântico ou modelagem computacional de alto nível. O foco do SBVR é a interoperabilidade semântica no intercâmbio de vocabulários e regras de negócios entre sistemas (OMG, 2019).

O modelo MOF e o metamodelo XMI do SBVR fornecem apenas a documentação para o intercâmbio de vocabulários e regras de negócios, sem a pretensões de serem usados como modelos de dados ou para raciocínio sobre informações de negócios, sendo necessária uma transformação para esse fim.

Quadro 5 - Fragmento da ontologia SBVR transformada do XSD

```

sbvr:Rule a owl:Class ;
  rdfs:label "rule" ;
  rdfs:comment "This is the rule property or class within the SBVR ontology." .

sbvr:Fact a owl:Class, sbvr:SententialForm ;
  rdfs:label "Fact" ;
  rdfs:comment "Proposition that is taken as true" .

```

Fonte: Elaborado pelo autor.

Na ausência de uma prescrição na especificação SBVR para essa transformação, uma alternativa é o uso de ontologias existentes. Uma das ontologias é a *Multiple Vocabulary Facility* (MVF)¹⁷, uma especificação da OMG, que permite a representação de elementos de um modelo em várias linguagens e terminologias. O MVF facilita a associação de vocabulários alternativos aos

¹⁷ Ver mais em: <https://www.omg.org/spec/MVF>. Acesso em: 15/mar./2025.

elementos do modelo, permitindo que eles sejam criados, exibidos e referenciados, de acordo com os termos e definições de diferentes vocabulários. Isso é útil em contextos em que há necessidade de expressar definições e conceitos em múltiplas formas, para facilitar a compreensão e a interoperabilidade entre domínios e disciplinas. O MVF dispõe de uma ontologia para termos e definições (ISO 1087), porém não há uma ontologia para as regras de negócio. Para trabalhos futuros, o mapeamento do modelo SBVR (XSD) para o MVF ou grafos conceituais (SOWA, 2008) poderá representar melhor o modelo de dados para o dicionário terminológico.

Para esta pesquisa, o esquema definido pelo SBVR¹⁸ foi adaptado para uma ontologia RDF mapeando a estrutura e os tipos de dados de XSD para RDF¹⁹. Essa ontologia será utilizada para definir os termos e as regras de negócio no KG. O Quadro 5 apresenta um exemplo com dois elementos da ontologia transformada do esquema XML da SBVR para turtle. O mapeamento das triplas para a ontologia é apresentada na seção “6.2 Implementação dos principais componentes”.

3.3 Fonte de regras de negócio e vocabulário

As regras de negócio e os vocabulários são compilados de diversas fontes. Nos setores fortemente regulados, eles podem ser obtidos nas leis e regulamentos. Porém, fontes como contratos e documentos técnicos também são fontes comuns. As seções seguintes apresentam uma breve discussão sobre algumas dessas fontes.

3.3.1 Leis e regulamentos

Regras de negócio e vocabulários nos setores regulamentados são extraídos de leis e regulamentos, bem como de contratos e documentos técnicos. A complexidade linguística desses textos legais, com as suas frases longas e termos especializados, apresenta desafios para o NLP, como discutido por Anh *et al.* (2023) e Wyner e Peters (2011). Nos EUA, a legislação do mercado de

¹⁸ Ver mais em: <https://www.omg.org/spec/SBVR/20190601/SBVR-XML-Schema.xsd>. Acesso em: 15/mar./2025.

¹⁹ Ver mais em: <https://www.w3.org/TR/swbp-xsch-datatypes/>. Acesso em: 15/mar./2025.

valores mobiliários é mantida pelo congresso e órgãos reguladores, como a autoridade reguladora da indústria financeira americana (*Financial Industry Regulatory Authority - FINRA*), que garante a equidade do mercado sob supervisão da SEC, a qual protege investidores e mantém a integridade dos mercados (SEC, 2023). Leis como a Securities Act de 1933, a Securities Exchange Act de 1934 e a Investment Company Act e Investment Advisers Act de 1940, foram motivadas pela quebra da bolsa de valores de 1929 e estabelecem a base regulatória. Respostas a crises mais recentes incluem a Lei Sarbanes-Oxley de 2002 e a Lei Dodd-Frank de 2010, que implementaram reformas significativas no controle financeiro e governança corporativa. As empresas são obrigadas a incorporar esses regulamentos e se adaptarem às suas atualizações, empregando práticas de GRC para governança, avaliação de riscos e conformidade, conforme apresentado por Racz *et al.* (2010).

Organizado em 50 títulos para refletir diferentes áreas de serviços governamentais, o CFR é atualizado trimestralmente para manter as regras administrativas atualizadas e legalmente vinculadas, fundamentais para a conformidade legal nos EUA (CFR, 2024a; 2024b). Especificamente, essa pesquisa se concentra no Título 17, que regula o setor de investimentos e é dividido em várias partes, abordando as regulamentações da comissão de negociação de futuros de commodities americana (*Commodity Futures Trading Commission - CFTC*) e da SEC. O CFR fornece o quadro regulatório para as atividades de investimento nos EUA, incluindo negociação de valores mobiliários e gerenciamento de fundos mútuos (JOSHI; SAHA, 2020). As leis estatutárias e o CFR são distintas, mas interligadas; as leis estabelecem os princípios legais pelo congresso, enquanto o CFR contém regulamentações detalhadas das agências federais que aplicam essas leis, com a lei estatutária prevalecendo em caso de conflito (CFR, 2024b).

A avaliação automatizada de informações textuais em NL é fundamental, devido ao constante desenvolvimento e complexidade das leis, exigindo ferramentas eficazes para análise jurídica, já que métodos convencionais envolvendo advogados e consultores fiscais são custosos e demorados, com risco de

omissão de informações vitais, o que pode impactar a precisão de opiniões legais e decisões de negócios (STRĄK; TUSZYNSKI, 2022).

Além da precisão, Governatori e Rotolo (2023) apresentam o problema da ambiguidade deôntica presente nos textos legais e a define como: uma literal p é deonticamente ambígua se existirem duas cadeias de raciocínio, uma sustentando a conclusão Op , enquanto a outra sustenta a conclusão $O\neg p$, e a relação de superioridade não resolve esse conflito. Embora a ambiguidade seja comum no raciocínio deôntico, ainda não foi amplamente explorado como esse tipo de conflito pode influenciar os subsequentes raciocínios deônticos. No âmbito desta pesquisa, conflitos entre regras não são tratados.

A IA tem atraído atenção para esses desafios, como abordado na *Competition on Legal Information Extraction and Entailment* (COLIEE) 2019 (Rabelo, 2019) e nos 30 anos de retrospectiva sobre o progresso em NLP e *Machine Learning* (ML), descritos por Sansone e Sperlí (2022).

3.3.2 Outras fontes

Joshi *et al.* (2021) usam a técnica de adaptação de domínio para empregar tecnologias de NLP treinadas em grandes quantidades de textos para o domínio regulatório em tarefas para contratos. Bouzidi *et al.* (2011) abordam uma técnica para modelar padrões para a verificação de conformidade usando ontologias. Os autores destacam a importância de convertê-los em um formato como o SBVR. Omrane *et al.* (2011) descrevem uma plataforma que utiliza ontologias lexicalizadas na indústria automotiva. Essa plataforma visa manter a conexão entre fontes textuais e regras de negócios formalizadas. Ela auxilia especialistas do domínio a entender e gerenciar as suas regras de negócios, fornecendo acesso aos documentos originais usados para criar conceitos ontológicos.

No processo de análise de sistemas, são produzidos documentos que traduzem requisitos funcionais e não-funcionais dos sistemas a serem desenvolvidos. Esses artefatos contêm informações de regras e vocabulários do negócio. Em um processo de engenharia reversa, podem ser extraídas e transformados novamente em regras e vocabulário, com possível perda semântica e de rastreabilidade com a sua origem legal, como demonstrado por Skersys *et al.*

(2018), que utiliza a transformação *model-to-model* (M2M), para desenvolver uma solução que extrai vocabulários e regras de negócios SBVR de diagramas Unified Modeling Language (UML). Cabot *et al.* (2010) apresentam um método de tradução automática baseada em padrões de esquemas conceituais em UML/*Object Constraint Language* (OCL) para SBVR. Em Skersys *et al.* (2022), a transformação é baseada em regras, em que os processos *Business Process Model and Notation* (BPMN) são convertidos em vocabulário e regras de negócio em SBVR. Bajwa *et al.* (2011) transformam as regras de negócio textuais em regras SBVR, com o auxílio de modelos de classes UML.

3.4 Posições jurídicas e modalidades de ôntica e aléticas

Hohfeld (1917) introduziu os conceitos “posições jurídicas” ou “correlativo jurídicos” e “opostos jurídicos” na teoria jurídica. Nas posições, destacam-se as interações entre direitos e deveres, privilégios e não-direitos, poderes e responsabilidades, imunidades e incapacidades, descrevendo as relações fundamentais dentro do direito. Por exemplo, se uma parte *A* possui um direito, a parte correspondente *B* tem um dever, como o dever de *B* de pagar por um serviço que *A* tem o direito de receber. Similarmente, se *A* tem um privilégio, como o de falar livremente, *B* não pode impedir essa ação. Nos casos de poder e responsabilidade, *A* pode alterar os status legais de *B*, e *B* é responsável por aceitar essas mudanças, como a rescisão de um contrato. As imunidades implicam que *A*, estando imune, não pode ter sua posição legal alterada por *B*, como é o caso de diplomatas com imunidade diplomática. Esses conceitos são complementados pelos opostos jurídicos que ajudam a ilustrar contrastes, em que não ter um direito é o oposto de ter um direito, ter um privilégio é o oposto de um dever, ter poder é o oposto de uma incapacidade e ser imune é o oposto de ser responsável.

Quadro 6 - Relações jurídicas

	direito	privilégio	poder	imunidade
Opostos	não-direito	dever	impotência (<i>disability</i>)	sujeição (<i>liability</i>)
Correlativos	dever	não-direito	sujeição	impotência

Fonte: Adaptado pelo autor de Hohfeld (1917)

A abordagem de Hohfeld (1917) revela a natureza relacional dos conceitos legais, fundamental para compreender as complexidades das relações jurídicas. Os opostos jurídicos no Quadro 6 servem para destacar os contrastes nas relações legais. Para os propósitos desta pesquisa, consideram-se apenas duas relações: as obrigações e poderes, que, segundo a teoria jurídica de Hohfeld (1917), são conceitos fundamentais, que delineiam a natureza das relações jurídicas.

Sharif *et al.* (2020) exploram o conceito de poderes no contexto jurídico, definindo-os como a capacidade legal de um indivíduo ou entidade para modificar, criar ou extinguir direitos, deveres ou outras posições jurídicas. Os poderes são correlativos jurídicos, vinculados à responsabilidade, permitindo a uma parte alterar unilateralmente as relações jurídicas, como ao terminar um contrato. Esse conceito é crucial em contextos que envolvem autoridades como juízes ou legisladores, que possuem autoridade para estabelecer e alterar normas jurídicas. A lógica deôntica é destacada como uma ferramenta eficaz para explicar posições legais usando princípios formais para definir obrigações e proibições. Os direitos são expressos como obrigações de outros (operador modal *O*), enquanto as permissões e as proibições são indicadas pelos operadores *P* e *F*, respectivamente. Um exemplo disso é que a proibição de divulgação de informações confidenciais seria expressa como *F(A divulga informações)*. A lógica deôntica também aborda a capacidade de alterar estados ou relações jurídicas, o que pode envolver complexas interações de ações e as suas consequências legais.

A modalidade deôntica no SBVR se refere à modalidade da obrigação, permissão e às suas relações. Ela é central para as regras de negócios comportamentais. Essas regras são sobre o que é permitido, proibido, ou obrigatório em um contexto de negócios. Por exemplo, uma regra de negócio que diz “um empregado deve reportar conflitos de interesse” é deôntica, pois implica uma obrigação. Essa modalidade também abrange a permissão: certas ações ou comportamentos são permitidos dentro do contexto organizacional (OMG, 2019).

Joshi e Saha (2020) integram sentenças deônticas ao KG, empregando um arcabouço para reconhecer e representar esses elementos. Isso envolveu a identificação e a classificação de frases e termos que expressam esses conceitos deônticos, garantindo que o KG não apenas capturasse as regulamentações e regras literais, mas também as obrigações e as permissões subjacentes que impõem ou concedem. Neill *et al.* (2017) apresentam um estudo sobre a classificação da modalidade na linguagem jurídica, focando, especialmente, nas regulamentações, atos e diretrizes financeiras. Os autores visam abordar os desafios na leitura e interpretação de textos jurídicos, especialmente, na identificação de modalidades deônticas relevantes (obrigações, proibições e permissões), devido à natureza estrita da linguagem jurídica.

A verificação de conformidade legal na Web Semântica, particularmente em um ambiente de *Linking Open Data* (LOD) proposta por Francesconi e Governatori (2022), modela as normas deônticas como classes de ontologia e restrições de propriedade em *Web Ontology Language 2* (OWL2). Essa abordagem de modelagem permite um raciocínio jurídico que se mostra eficaz e uma verificação de conformidade usando tecnologia disponível. Os autores ilustram a sua abordagem como exemplos, incluindo normas sobre termos e condições contratuais e limites de velocidade. Ademais, eles abordam a anulação da norma, que pode ser substituída por outra norma sob condições específicas. Os autores demonstram como modelar conflitos de normas e exceções em um arcabouço de lógica descritiva, garantindo que o raciocínio jurídico permaneça dentro de um perfil de complexidade computacional aceitável.

A modalidade alética na especificação SBVR (OMG, 2019) se refere às modalidades de necessidade, possibilidade e impossibilidade, conforme se relacionam com a verdade das proposições dentro de um domínio de negócios. A modalidade alética é usada para definir a estrutura lógica do vocabulário de negócios, especificando o que deve ser verdade, o que pode ser verdade e o que não pode ser verdade. Um exemplo disso é que a necessidade de um produto ter um preço; a possibilidade de cancelar um pedido em até 24 horas e a impossibilidade de um cliente ter saldo negativo. Essas modalidades são

fundamentais para estabelecer as verdades essenciais e a estrutura lógica dos conceitos de negócio. Entretanto, as pesquisas estudadas no presente trabalho não exploram explicitamente essa modalidade, não tendo sido identificados trabalhos sobre a modalidade alética, com exceção de Neill *et al.* (2017), que citam a necessidade de trabalhos futuros para essa modalidade.

Parvizimosaed *et al.* (2022) sugerem que contratos e regulamentações desempenham papéis distintos em contextos jurídicos. Um contrato independente típico, geralmente, contém um número limitado de propriedades que são específicas ao acordo entre as partes envolvidas. Esse contrato é um documento autônomo, que delinea as obrigações e poderes, termos e condições acordados pelas partes, por outro lado, regulamentações são regras abrangentes com as quais os contratos devem estar em conformidade. Os autores se referem a direitos, deveres, poderes e responsabilidades designados a entidades em contextos legais, como contratos ou legislação. Enquanto as modalidades deônticas fornecem uma compreensão teórica das normas, as posições legais se concentram nas suas implicações práticas e específicas, detalhando obrigações e poderes oriundos de acordos contratuais ou estatutos legais.

Nesta pesquisa, as regras são classificadas segundo a taxonomia de Witt (2012) e as modalidades são derivadas dessa classificação.

3.5 Aplicações de NLP no contexto legal

O NLP é um campo da IA que se concentra na interação entre máquinas e humanos pela linguagem natural. É utilizada em uma ampla gama de aplicações, incluindo análise de texto, tradução automática, *chatbots*, recuperação de informações, reconhecimento de fala, resumo de texto, reconhecimento de entidades nomeadas, sistemas de perguntas e respostas, análise de sentimentos e geração automática de texto. O NLP permite que as máquinas entendam, interpretem e gerem a linguagem humana, facilitando a comunicação eficiente entre humanos e computadores e automatizando a análise de grandes volumes de dados de texto em várias indústrias e domínios.

Zhong *et al.* (2020) destacam o impacto crescente do NLP no domínio jurídico, melhorando significativamente o acesso e a análise de informações por profissionais do direito. O NLP facilita a pesquisa jurídica, auxiliando na recuperação eficiente de documentos legais relevantes e agiliza a análise de contratos, ao identificar automaticamente cláusulas chave e riscos, assegurando a conformidade com padrões legais. A tecnologia também é aplicada para resumir extensos documentos legais, permitindo uma compreensão rápida dos pontos principais. Adicionalmente, o NLP está emergindo na análise preditiva, ajudando advogados a preverem resultados de casos e na descoberta legal, acelerando a coleta de evidências. Outras aplicações incluem monitoramento de conformidade, em que o NLP garante a aderência a leis e regulamentos em evolução, bem como sistemas de perguntas e respostas que proporcionam consultas legais precisas. O NER e a análise de sentimentos são também utilizados para categorizar informações e discernir perspectivas em argumentos legais.

A aplicação de NLP no contexto legal é uma área bastante ativa, congressos como *International Conference on Artificial Intelligence and Law* (ICAL)²⁰ e competições como COLIEE, são exemplos disso. Goebel *et al.* (2023) resume a competição COLIEE que envolveu quatro tarefas relacionadas à jurisprudência e legislação. A tarefa 1 focou na recuperação de jurisprudência, com a equipe vencedora empregando modelos de linguagem pré-treinados e processamento heurístico. A tarefa 2 envolveu a implicação de jurisprudência, em que a equipe de melhor desempenho usou um modelo de sequência a sequência (um tipo de RNN) com técnicas de ajuste fino. A tarefa 3 exigiu que os participantes recuperassem artigos do código civil japonês relevantes para questões de exames da ordem dos advogados, com o melhor desempenho alcançado, usando uma abordagem com *Large Language Models* (LLMs). A tarefa 4, implicação textual e resposta a perguntas da legislação, alcançou a maior

²⁰ As submissões para a ICAL 19^a - 2023 cobriram a mineração de argumentos, classificação e sumarização de textos legais, além de abordagens computacionais para negociação, formação de contratos e resolução de disputas. Elas incluem modelos de raciocínio jurídico usando lógicas deônticas, sistemas baseados em conhecimento e implicações éticas da IA no direito. As técnicas envolvem aquisição de conhecimento, processamento de NL específico para o direito e uso de ML para análises jurídicas.

precisão com LLMs. A competição destacou tendências em recuperação e implicação de informações legais, como o uso de LLMs, métodos de conjunto e técnicas tradicionais de recuperação da informação.

Anh *et al.* (2023) examinam o impacto dos LLMs no NLP, particularmente no domínio jurídico, no qual os textos são caracteristicamente complexos e contextuais e os LLMs ajudam a desambiguar e interpretar frases longas, aprimorando a compreensão contextual. As aplicações dos LLMs incluem sumarização de documentos, análise de contratos, respostas a perguntas, classificação de texto e extração de informações, aumentando a eficiência na pesquisa jurídica e acessibilidade das informações. Entretanto, surgem preocupações éticas e legais, como viés, privacidade e propriedade intelectual, que demandam diretrizes claras para uso responsável. Desafios futuros envolvem o processamento de texto jurídico multilíngue, ajuste fino específico ao domínio e integração com outras ferramentas jurídicas.

O campo de NLP em IA é vital para a interação entre máquinas e humanos pela linguagem natural. No domínio jurídico, como destacam Zhong *et al.* (2020), o NLP está revolucionando a forma como informações são acessadas e analisadas, facilitando a pesquisa jurídica, análise de contratos e a identificação de cláusulas e riscos, além de ajudar na conformidade legal. O NLP também está sendo empregado em análises preditivas para prever resultados de casos, bem como na automação da descoberta de documentos durante litígios. A conferência ICAL e a competição COLIEE exemplificam a integração crescente do NLP no campo jurídico, mostrando a aplicação de técnicas avançadas e modelos LLM na melhoria da eficiência e acessibilidade das informações jurídicas.

3.6 Transformação em vocabulário e regras

O principal objetivo desta pesquisa é a transformação de regras de negócio escritas em NL para o formato de inglês estruturado (*Structured English* - SE), conforme a especificação do SBVR (OMG, 2019). Essa transformação de formato deve ocorrer sem alterar o significado original, o que implica a preservação de todas as características essenciais do texto de origem. Nesse

processo, identifica-se no texto original os elementos que são adequados para serem transformados, pois o destino da transformação, a regra SBVR, não aceita declarações que não são regras.

Em adição, a especificação do SBVR define o termo “comunidade” como um grupo de pessoas ou organizações que compartilham um entendimento comum de conceitos e terminologia em um domínio específico (OMG, 2019). Assim, a transformação deve ocorrer dentro de um domínio bem definido, para garantir a preservação do significado. Define-se então a transformação de regras de negócio de NL para SBVR como a transformação das regras de negócio em NL para um formato de SE, respeitando os conceitos e a terminologias acordados pela comunidade do domínio e em concordância com a especificação SBVR.

Os trabalhos relacionados a tal transformação são discutidos no capítulo “4 DISCUSSÃO DAS SOLUÇÕES”.

3.7 Ontologias

As ontologias são vocabulários formalizados de termos, muitas vezes cobrindo um domínio específico e compartilhados por uma comunidade de usuários. Elas especificam as definições dos termos descrevendo as suas relações com outros termos na ontologia. A OWL 2 é uma extensão e revisão da linguagem de ontologia web (*Web Ontology Language - OWL*) desenvolvida pelo grupo de trabalho de ontologia da web do W3C e publicada inicialmente em 2004 (referida aqui como “OWL 1”). Como o OWL 1, o OWL 2 é projetado para facilitar o desenvolvimento e compartilhamento de ontologias via web, com o objetivo final de tornar o conteúdo da web mais acessível para máquinas (W3C, 2012).

Este trabalho não tem como objetivo principal construir uma ontologia para o vocabulário de negócios e as regras, conforme discutido na seção “3.2 Semantic for Business Vocabularies and Business Rules (SBVR)”. No entanto, considerando que os vocabulários e regras são representados por meio de grafos de conhecimento, o uso de ontologias pode ampliar significativamente sua compreensão e, assim, contribuir para os objetivos desta pesquisa. Com isso em vista, são apresentados estudos que apoiam essa abordagem. O processo de

desenvolvimento é detalhado no capítulo “5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR”.

Joshi e Saha (2020) desenvolveram uma ontologia proprietária para estruturar um grafo de conhecimento para CFR, refletindo a sua organização hierárquica e relações semânticas, similar à ontologia formato de intercâmbio de conhecimento jurídico (*Legal Knowledge Interchange Format - LKIF*²¹), usada na ontologia de regulamentos financeiros (*Financial Regulation Ontology - FRO*). Desenvolvida em OWL, a ontologia suporta consultas complexas em SPARQL, facilitando a pesquisa e análise dentro do corpus regulatório do CFR.

Bouzidi *et al.* (2011) desenvolvem uma metodologia para modelar padrões técnicos usando ontologias e a linguagem SBVR, facilitando a verificação de conformidade por representar regras de negócios em CNL. Eles criaram a OntoDT, uma ontologia com 50 classes e 26 propriedades, extraída de documentos do centro científico e técnico de construção (*Centre Scientifique et Technique du Bâtiment - CSTB*²²) e a integraram com um dicionário de sinônimos para formar um vocabulário que vincula avaliações técnicas. Esse trabalho também destaca os desafios da interpretação ambígua e a necessidade de especialistas para garantir a clareza dos textos regulatórios. Os autores propõem o uso de regras baseadas no modelo evento-condição-ação (*Event Condition Action - ECA*) para clarificar decisões de não conformidade, mas reconhecem limitações como a necessidade de intervenção manual, representação insuficiente de conhecimento especializado, extração trabalhosa das regras e reformulação de textos para maior clareza.

Sharif *et al.* (2020) desenvolvem uma ontologia para mapear as relações jurídicas em contratos, detalhando a interação entre entidades e conceitos contratuais. A estrutura da ontologia gira em torno do elemento “contrato”, que se liga a “partes”, podendo integrar um “contrato pai” ou ser subdividido em

²¹ O LKIF é um formato para compartilhar conhecimento jurídico, como políticas, leis e casos, entre diferentes sistemas. Ele utiliza OWL para definir conceitos e SWRL para expressar regras. Ademais, possui uma ontologia básica de conceitos jurídicos. O LKIF foi desenvolvido pelo projeto europeu denominado ESTRELLA (ver mais em: <https://github.com/RinkeHoekstra/lkif-core>. Acesso em: 15/mar./2025).

²² https://en.m.wikipedia.org/wiki/Scientific_and_Technical_Centre_for_Building

“subcontratos”. As “partes” são entidades com “ativos” e assumem “papéis” como credor ou devedor, definindo as suas obrigações e direitos. A ontologia abrange “posições legais” como “obrigações”, que são deveres ativados por um “antecedente” e resultam em um “consequente”, e “poderes”, que permitem a uma parte alterar o estado do contrato. Ademais, são incluídos conceitos de “evento”, “situação” e “tempo” para contextualizar as obrigações e poderes, detalhando como uma situação pode desencadear uma obrigação com uma temporalidade específica. Essa estrutura facilita a análise, redação e interpretação de contratos, evidenciando a complexidade das interações jurídicas.

Omrane *et al.* (2011) descrevem uma plataforma para a indústria automotiva, que usa ontologias lexicalizadas para vincular textos a regras de negócio formalizadas, melhorando a gestão e a compreensão dessas regras por especialistas do domínio. A plataforma permite acesso direto aos documentos originais que dão suporte aos conceitos ontológicos, aumentando a flexibilidade e assegurando que as regras refletem os significados intencionados. A lexicalização associa um léxico aos conceitos ontológicos, enriquecendo a estrutura semântica e possibilitando a anotação de documentos para vincular textos a conceitos relevantes. Apesar de facilitar a exploração textual por consultas semânticas e apoiar especialistas do domínio na transformação de textos em regras, a plataforma enfrenta desafios, como a complexidade de expressar conhecimento em linguagens lógicas, custos elevados na criação de ontologias detalhadas, riscos de ambiguidades textuais e perda de nuances durante a transformação textual.

Sansone e Sperlí (2022) destacam os desafios para a interpretação de textos legais por IA, devido à escassez de ontologias específicas do setor e dificuldades na gestão e atualização dessas estruturas. Os autores apontam que a linguagem utilizada em documentos legais é frequentemente abstrata e especializada, contribuindo para ruídos na análise e classificação, devido ao extenso vocabulário e falta de contexto.

Hoekstra *et al.* (2007) apresentam uma ontologia legal que faz parte de uma arquitetura genérica para sistemas de conhecimento jurídico. Essa ontologia,

chamada LKIF, é projetada para facilitar a troca de conhecimento entre sistemas de conhecimento jurídico existentes, permitindo a tradução de bases de conhecimento jurídico escritas em diversos formatos e formalismos de representação. Os autores detalham a metodologia por trás da ontologia LKIF Core, que inclui conhecimento terminológico ou definições intensionais de conceitos representados como classes. A ontologia visa apoiar a aquisição de conhecimento e a modelagem em domínios jurídicos, definindo conceitos jurídicos comuns como normas, juízos, responsabilidade e documentos. A ontologia é estruturada em vários módulos que representam um grupo de conceitos estreitamente relacionados e utiliza uma combinação de *Web Ontology Language Description Logics* (OWL-DL) e *Semantic Web Rule Language* (SWRL) para a sua estrutura.

A presente pesquisa reutiliza duas ontologias: a FIBO desenvolvida pela EDMC (2024), criada após a crise financeira de 2008 para melhorar a clareza dos dados, e a FRO, baseada em LKIF, desenvolvidas pela Jayzed Data Models inc. (2021), ambas sob licença de código-livre (*Open Source Software – OSS*).

3.8 Grandes modelos de linguagem (LLMs)

Os LLMs são um tipo de modelo de IA projetado para entender e gerar linguagem humana. Esses modelos são “grandes”, no sentido de que são treinados em vastas quantidades de dados de texto e contêm um número significativo de parâmetros. Exemplos de LLMs incluem a série *Generative Pre-trained Transformer* (GPT) da OpenAI²³, como o GPT-4, e o Gemini²⁴ (antigo BARD) da Google. Esses modelos são capazes de realizar uma ampla gama de tarefas de NLP.

Min *et al.* (2023) exploram o impacto dos LLMs no campo do NLP, destacando três paradigmas principais de utilização desses modelos: pré-treinamento seguido de ajuste fino, aprendizado baseado em *prompts* e NLP como geração de texto. No primeiro paradigma, os LLMs são pré-treinados em um extenso

²³ Ver mais em: <https://openai.com/>. Acesso em: 15/mar./2025.

²⁴ Ver mais em: <https://deepmind.google/technologies/gemini/#introduction>. Acesso em: 15/mar./2025.

corpus com tarefas gerais de modelagem de linguagem, o que permite ao modelo desenvolver representações linguísticas. Posteriormente, esses modelos são ajustados a conjuntos de dados menores e específicos para aperfeiçoar seu desempenho em tarefas direcionadas, uma prática que se tornou padrão em NLP. O segundo paradigma, o aprendizado baseado em *prompts*, utiliza *prompts* de linguagem natural para direcionar os modelos a realizarem tarefas específicas sem a necessidade de ajuste fino adicional, sendo ideal para situações de aprendizado com poucos exemplos. O terceiro paradigma trata o NLP como uma tarefa de geração de texto, cujo objetivo é gerar um texto específico a partir de um texto de entrada, utilizando as capacidades gerativas de modelos como GPT-3 e T5 para aplicações como tradução automática, sumarização e resposta a perguntas.

Sun *et al.* (2022) discutem a evolução das metodologias em NLP, destacando a transição de paradigmas tradicionais para novas abordagens facilitadas pelo aprendizado profundo e modelos de linguagem pré-treinados. Os autores categorizam sete paradigmas principais em NLP, incluindo classificação, correspondência, rotulagem de sequência, compreensão de leitura de máquina, sequência para sequência, sequência para sequência de ações e modelagem de linguagem (com Máscara), fornecendo percepções sobre suas aplicações e tarefas representativas. Os autores exploram, ainda, como várias tarefas de NLP, como classificação de texto, NER e extração de relações passaram por mudanças de paradigma, resultando em melhoria do desempenho do modelo. Os autores também discutem o potencial de certos paradigmas para unificar diversas tarefas de NLP em uma única estrutura, destacando a tendência para abordagens de modelagem mais versáteis e eficientes.

Chen *et al.* (2023) investigam o uso de LLMs, como GPT-3.5 e GPT-4, para automação total na modelagem de domínio usando a UML. O estudo avalia a automação sem treinamento supervisionado, aplicando técnicas de engenharia de *prompts* em um conjunto limitado de exemplos de modelagem. Embora os LLMs exibam um entendimento notável do domínio, eles ainda não são completamente práticos para uma automação total. O GPT-4 se destacou, mas

com limitações: atingiu pontuações F1-Score²⁵ de 76% em geração de classes, 61% em atributos e apenas 34% em relacionamentos, indicando uma habilidade decrescente de geração de classes para relacionamentos. Os resultados mostraram que a engenharia de *prompts* pode melhorar a recuperação de classes e relações, porém, o aumento de exemplos nem sempre amplia o desempenho e etapas adicionais de raciocínio podem até reduzir a eficácia. Comparativamente, o GPT-4 foi consistentemente superior ao GPT-3.5, especialmente nos cenários em que nenhum exemplo é fornecido para o modelo (*zero-shot*).

Anh *et al.* (2023) abordam o processamento de texto jurídico, destacando os desafios, abordagens tradicionais e o papel dos LLMs na superação desses desafios. Os principais desafios incluem a ambiguidade da linguagem jurídica, a presença de frases longas e complexas e a necessidade de processamento sensível ao contexto. As abordagens tradicionais, como métodos baseados em regras, sistemas de recuperação de informações e NER têm limitações ao lidar com esses desafios. Embora eficazes para tarefas simples, esses métodos lutam com a complexidade e ambiguidade presentes nos textos jurídicos. Os LLMs oferecem uma solução promissora para superar essas limitações. Eles são capazes de entender o contexto, desambiguar termos e frases legais complexas e processar frases longas de maneira eficiente. Além disso, as incorporações contextuais (*contextual embeddings*) dos LLMs permitem um processamento de linguagem mais sensível ao contexto. O uso de LLMs pode resultar em uma análise de texto mais eficiente, precisa e contextualmente rica, beneficiando profissionais jurídicos nas suas práticas.

Câmara *et al.* (2023) investigam o uso do LLM ChatGPT, na modelagem de software UML. Apesar do ChatGPT, frequentemente, gerar diagramas de classes UML sintaticamente corretos, esses modelos carecem de precisão semântica e nem sempre correspondem à intenção do usuário. O desempenho do modelo varia conforme o contexto e o domínio, mostrando-se mais eficaz em áreas com as quais o modelo tem maior familiaridade. O ChatGPT enfrenta desafios ao lidar

²⁵ O F1-Score é uma métrica calculada a partir da média harmônica entre a precisão e a sensibilidade. Ele ornece uma média equilibrada do desempenho de um modelo de classificação.

com modelos UML que envolvem mais de 10 a 12 classes e com conceitos complexos como herança múltipla e classes de associação. A pesquisa também destaca a sensibilidade do ChatGPT às variações nos *prompts*, o que pode levar a inconsistências significativas nas respostas. Os autores sugerem que os LLMs têm potencial para apoiar a engenharia de sistemas baseada em modelos (*Model-Based Systems Engineering - MBSE*), propondo a publicação de mais modelos de software e a integração de LLMs na prática e no ensino de modelagem. No entanto, os autores ressaltam a necessidade de melhorias nas capacidades de modelagem do ChatGPT e na adaptação das práticas de modelagem para integrar eficazmente essas ferramentas de IA.

Combemale *et al.* (2023) discutem a integração potencial do ChatGPT e outros LLMs na modelagem de software, explorando vários papéis, como a transformação direta de requisitos em código, modelagem intermediária, assistência de modelagem, geração de casos de teste e substituição de documentação. Os autores destacam preocupações sobre a confiabilidade dos resultados gerados por IA e o impacto para os profissionais da área. Eles defendem uma exploração e experimentação mais aprofundadas com o ChatGPT no campo da modelagem de software, enfatizando a necessidade de um entendimento equilibrado das suas capacidades e limitações.

Tonmoy *et al.* (2024) exploram técnicas para mitigar a alucinação em LLMs, pois esses modelos podem gerar informações não fundamentadas em fatos, um risco em áreas críticas como medicina e finanças. Eles apresentam uma taxonomia das abordagens de mitigação divididas em engenharia de *prompt* e desenvolvimento de modelos. Na engenharia de *prompt*, destacam-se métodos como geração aumentada de recuperação (*Retrieval Augmented Generation - RAG*), que integra conhecimento externo para ancorar respostas em fatos, e autorrefinamento (*self refinement*), permitindo que os modelos refinem as suas respostas após a geração inicial. No desenvolvimento de modelos, abordagens incluem novas técnicas de decodificação que direcionam a geração de texto, uso de grafos de conhecimento para confirmar precisão factual, funções de perda baseadas em fidelidade para manter a consistência com dados de entrada ou verdade fundamental, e ajuste fino (*fine-tuning*), supervisionado para aumentar

a precisão e a confiabilidade. Essas técnicas são avaliadas pela sua eficácia em prevenir alucinações, crucial para a aplicação segura dos LLMs.

Essa é uma área de pesquisa ativa, como exemplificado com o trabalho de Yu *et al.* (2023), que investiga o uso de LLMs nas tarefas de raciocínio jurídico. Os autores se concentram na engenharia de *prompt*. A pesquisa é especificamente direcionada para a tarefa de implicação do COLIEE, baseada no exame da ordem dos advogados do Japão. Os autores experimentam abordagens variando o número de exemplos para o modelo de zero (*zero-shot*) a alguns (*few-shot*) e ajuste fino, combinando várias estratégias de *prompting*. Eles também discutem as limitações e os potenciais riscos associados ao uso de LLMs nas tarefas de raciocínio jurídico, destacando a importância das considerações éticas e a necessidade de mais pesquisas nessa área.

A aplicação de LLMs nas tarefas de NLP demonstra ser uma evolução da forma tradicional, como modelos dedicados a tarefas específicas ou algoritmos, porém utilizar o LLM para orquestrar todo o processo não parece ser a solução. Um ponto médio é combinar a capacidade de entendimento dos LLMs com outras tecnologias para obter melhor eficiência, reproduzibilidade e transparência.

3.8.1 Engenharia de *prompts* e alucinações

Não há uma única definição para engenharia de *prompt* e alucinação dos LLMs. As alucinações, no contexto dos LLMs, referem-se à geração de conteúdo pelo modelo que parece factual, mas na verdade é infundado ou incorreto. Isso pode manifestar-se quando os LLMs produzem respostas que contêm imprecisões, fabricações ou informações irrelevantes que não têm apoio nos dados de entrada ou fontes externas de verdade (TONMOY *et al.*, 2024).

A engenharia de *prompt* é uma técnica usada para melhorar o desempenho e a precisão de LLMs, ao criar, cuidadosamente, consultas de entrada fornecidas ao modelo. O *prompt* é o texto de entrada que orienta o modelo na geração de uma resposta. Na engenharia de *prompt*, o objetivo é projetar *prompts* que comuniquem, efetivamente, a tarefa ou questão para o modelo, provocando respostas mais precisas, relevantes e úteis. Isso envolve experimentar diferentes

formulações, estruturas e estilos de *prompts* para identificar aqueles que levam aos melhores resultados para uma determinada tarefa (TONMOY *et al.*, 2024).

Opsahl-Ong *et al.* (2024) definem a engenharia de *prompt* como o processo de criação e otimização de instruções e exemplos para guiar modelos de linguagem em tarefas complexas. No contexto de pipelines ou programas de modelos de linguagem, essa prática se concentra em estruturar cadeias de *prompts* para que diferentes módulos do modelo interajam de maneira eficaz. A engenharia de *prompt* envolve identificar e ajustar os elementos específicos de cada instrução e exemplo que contribuem para a performance desejada, evitando abordagens de tentativa e erro exaustivas.

O autor destaca que a engenharia de *prompt* em *pipelines* multietapas é especialmente desafiadora, devido ao grande espaço de possibilidades de *prompts* (problema de proposta) e à dificuldade de identificar quais partes do *prompt* são responsáveis pelo desempenho (problema de atribuição de crédito). A solução envolve técnicas como a geração de exemplos adequados, a criação de instruções ancoradas no objetivo da tarefa e a otimização probabilística para encontrar as melhores combinações de elementos, tornando o processo mais sistemático e eficiente.

Tonmoy *et al.* (2024) investigam a alucinação em LLMs, um desafio crítico em domínios sensíveis, como saúde, atendimento ao cliente, finanças e direito, nos quais a precisão é crucial. Para combater a alucinação, desenvolveram-se técnicas de mitigação agrupadas em engenharia de *prompt* e desenvolvimento de modelos. A engenharia de *prompt* aprimora as respostas do LLM usando técnicas como RAG, que integra conhecimentos externos durante a geração e o autorrefinamento, que permite ao modelo refinar suas respostas com *feedback* contínuo. O desenvolvimento de modelos inclui novas estratégias de decodificação para limitar alucinações, uso de grafos de conhecimento para manter a precisão factual e ajuste fino supervisionado que alinha os LLMs com dados rotulados específicos da tarefa. Contudo, essas abordagens enfrentam desafios como a dependência de bases de conhecimento, que podem estar desatualizadas, altos custos computacionais, a qualidade variável do feedback

e a necessidade de equilibrar a redução da alucinação com a preservação da criatividade nas respostas do modelo.

Agrawal *et al.* (2023) apresentam três abordagens principais para integrar KG em LLMs para mitigar as alucinações, a inferência, o treinamento e a validação cientes de conhecimento (*knowledge-aware*). Na inferência ciente de conhecimento, o foco é melhorar o processo de inferência dos LLMs com técnicas como RAG por KG, que fornecem documentos relevantes dos KGs durante a geração de texto. O treinamento ciente de conhecimento visa otimizar o processo de aprendizagem dos LLMs, incorporando conhecimento dos KGs durante o treinamento. A validação ciente de conhecimento usa dados estruturados do KG como referência para verificar fatos e validar as informações geradas pelos LLMs.

Opsahl-Ong *et al.* (2024) tratam o problema de otimização de prompts com o método MIPRO, um otimizador para programas de modelos de linguagem multiestágio, projetado para melhorar o desempenho desses modelos em tarefas complexas que exigem múltiplas etapas de processamento. Ele resolve problemas comuns na construção de *prompts* eficazes, como a necessidade de gerar instruções personalizadas e realizar a atribuição de crédito correta para cada módulo do programa. O MIPRO utiliza um modelo bayesiano para ajustar *prompts* de maneira eficiente, automatizando o processo de criação de instruções e alguns exemplos (*few-shots*) para cada etapa, sem depender de ajustes manuais ou rótulos intermediários.

Os principais desafios que o MIPRO aborda incluem a vasta quantidade de possibilidades de *prompts* e a dificuldade de identificar como cada módulo contribui para o desempenho final. O MIPRO gera instruções fundamentadas no contexto da tarefa, enquanto a otimização bayesiana avalia quais combinações de instruções e demonstrações trazem melhores resultados. Isso permite que o MIPRO atribua créditos com precisão, ajustando o programa de forma que cada etapa melhore o desempenho geral, especialmente em tarefas que possuem regras condicionais complexas, em que as instruções e os exemplos otimizados são críticos.

Nos testes o MIPRO obteve resultados superiores, melhorando a precisão em até 13% em tarefas como perguntas com mais de um salto (*multi-hop*), classificação e verificação de fatos. A otimização conjunta de instruções e exemplos se mostrou particularmente eficaz, especialmente nas tarefas condicionais, em que apenas demonstrações não são suficientes para o aprendizado completo da atividade. Com isso, o MIPRO se destaca por automatizar de forma eficiente a otimização de *prompts* para programas de LM multiestágio, aumentando o desempenho e reduzindo a necessidade de intervenção manual.

Para esta pesquisa, o uso de *prompts* é uma parte fundamental de como as instruções são dadas ao modelo para atingir um objetivo específico, obter respostas precisas e minimizar as variações (diferentes respostas para a mesma pergunta), são requisitos para o uso dos LLMs nos processos do CFR2SBVR.

3.9 Grafos de conhecimento

O KG é definido como um grafo de dados destinado a acumular e transmitir conhecimento do mundo real, cujos nós representam entidades relevantes para um domínio específico e as arestas representam relações potencialmente diferentes entre essas entidades (HOGAN *et al.*, 2021).

Joshi e Saha (2020) exploram o uso de KGs para transformar textos legais complexos e densos, como o CFR, em formatos ricos em semântica e legíveis por máquinas. Os KGs ajudam a organizar e capturar a complexa rede de regulamentos, termos e regras, bem como as suas interconexões, facilitando a análise e o processamento automatizados desses documentos. Para construir o grafo, os autores começam extraíndo elementos-chave do CFR, incluindo termos e suas relações, tópicos e regras. Esses elementos são interligados para refletir a estrutura do texto legal e enriquecidos com terminologias semanticamente semelhantes e sentenças deônticas, que tratam de necessidade, obrigação e permissão. O resultado é um KG que não só organiza de forma clara o texto legal, mas também capture as suas nuances e as complexas interrelações inerentes aos documentos legais.

Sequeda *et al.* (2024) destacam o papel dos KGs na melhoria dos LLMs como o GPT-4, particularmente nas aplicações complexas como consultas a bancos de dados SQL empresariais. Os KGs proporcionam uma representação estruturada dos dados que auxilia os LLMs a compreenderem as consultas complexas de maneira mais eficaz, a partir de uma melhor contextualização e enriquecimento semântico. Essa estrutura permite uma interpretação mais precisa das consultas, ajudando na desambiguação e refinamento delas pela integração de informações detalhadas sobre entidades e as suas interrelações. Essa abordagem melhorou, significativamente, a precisão das respostas do GPT-4 em consultas de bancos de dados SQL. Originalmente, sem KGs, a precisão das respostas do GPT-4 era de apenas 16,7%. Com a integração dos KGs, essa precisão saltou para 54,2%. A precisão em consultas simples, menos complexas, alcançou 71,1% com KGs, contra 25,5% sem eles. Nos cenários de maior complexidade, a precisão variou de 35,7% a 66,9% com KGs, enquanto a abordagem direta sem KGs mostrou 0% de precisão nas consultas bastante complexas. Esses resultados destacam a eficácia dos KGs em aumentar a precisão dos LLMs em tarefas de consulta de dados, especialmente em contextos de alta complexidade.

Louis *et al.* (2023) desenvolveram o modelo G-DSR, que combina técnicas avançadas de NLP e rede neural de grafos (*Graph Neural Network - GNNs*) para melhorar a recuperação de leis estatutárias. Eles utilizaram redes hierárquicas e mecanismos de atenção para recuperar informações que superam as técnicas baseadas em correspondência de palavras-chave. O modelo foca no entendimento do conteúdo semântico e do contexto dos documentos legais, essencial para a aplicação no domínio jurídico. O G-DSR inclui um codificador grafo legislativo que mapeia a complexa estrutura dos documentos legais em grafos, assim como um recuperador denso que transforma textos em representações vetoriais densas, capturando significados e contextos mais profundos. Essa abordagem permite uma recuperação mais precisa e eficiente de informações legais, lidando, eficazmente, com a complexidade e as particularidades da linguagem jurídica.

As LLMs podem auxiliar nesse processo, automatizando vários aspectos da construção e manutenção de grafos de conhecimento. Um exemplo disso é que eles podem traduzir consultas em NL para consultas SPARQL, usadas para recuperar informações de grafos. Eles também podem ajudar na exploração e sumarização do conteúdo de grafos de conhecimento existentes (MEYER *et al.*, 2024).

Junior *et al.* (2019) discutem a criação de grafos de conhecimento usando a estrutura de descrição de recursos (*Resource Description Framework* - RDF) para documentos legais, visando melhorar a precisão de busca e tornar as informações legais mais acessíveis e compreensíveis. A principal abordagem é transformar documentos legais armazenados em formato XML em grafos de conhecimento RDF usando diferentes linguagens de mapeamento, entre elas: XSPARQL, que combina XQuery e SPARQL; SPARQL-Generate, que estende SPARQL para gerar RDF de fontes variadas; *RDF Mapping Language (RML)-Mapper*, que amplia o R2RML para formatos semiestruturados, e CARML, que implementa RML para gerar conjuntos de dados RDF. Cada linguagem tem características únicas, como a capacidade de manipular e transformar dados de maneiras específicas. Os autores avaliam essas linguagens com base em critérios como compatibilidade de formato de dados, seleção de dados e transformação de valores literais em Identificadores de Recursos Internacionalizados (IRIs). O CARML (LEE *et al.*, 2011) é uma ferramenta de código livre²⁶, que dá suporte a geração de conjuntos de dados RDF a partir de uma variedade de formatos de dados, incluindo XML. Além disso, possui capacidades de funções de transformação, apoiadas por uma ontologia de funções²⁷, permitindo manipulações complexas de dados, frequentemente necessárias nos documentos jurídicos.

Casellas *et al.* (2011) descrevem a conversão do CFR de XML para um formato RDF próprio para melhorar a pesquisa semântica e a recuperação de informações. O processo se inicia com a identificação de seções relevantes do CFR, seguida pela análise e estruturação desses dados em objetos RDF,

²⁶ Ver mais em: <https://github.com/carml/carml>. Acesso em: 15/mar./2025.

²⁷ Ver mais em: <https://fno.io/>. Acesso em: 15/mar./2025.

marcando e anotando com metadados ligados a tesouros jurídicos²⁸. Componentes como títulos e parágrafos são organizados eficientemente em RDF. Eles desenvolvem um esquema RDF para estabelecer relações e propriedades, facilitando a criação de um modelo de dados estruturado. A extração e a transformação dos dados são feitas a partir de uma análise programática, mapeando elementos textuais para classes e propriedades do esquema RDF, com o apoio de ferramentas de análise de texto. Metadados e anotações enriquecem a utilidade dos dados. A qualidade e a precisão dos dados RDF são asseguradas por revisões manuais e verificações automatizadas no final do processo.

3.9.1 Grafos de conhecimento para finanças

Abu-salih (2021) discute a importância dos KGs no setor financeiro, destacando a distinção entre KGs genéricos, que servem para tarefas amplas e independentes de domínio, e KGs específicos de domínio, que são essenciais para resolver problemas intrínsecos a áreas específicas como as finanças. Os KGs específicos de domínio no setor financeiro são construídos com base em uma ontologia de domínio predefinida, representando o conhecimento de alto nível do domínio e os seus subdomínios por entidades e relações interrelacionadas. Esses KGs são especialmente valiosos, devido à complexidade, dinamismo dos mercados financeiros e grandes volumes de dados heterogêneos, permitindo a vinculação de diversos tipos de dados para uma visão mais holística das entidades financeiras e as suas complexas inter-relações. Essa capacidade é crucial nas aplicações como detecção de fraudes e tomada de decisões de investimento, nas quais os KGs facilitam a identificação de padrões sutis e fornecem percepções que suportam melhores estratégias de investimento e avaliações de risco mais precisas.

Em suma, devido à sua capacidade de representar relações complexas entre entidades, os KGs encontram aplicações diversas em vários domínios (Fensel *et al.*, 2020). Eles também são usados para a detecção de fraude financeira

²⁸ O dicionário de sinônimos de termos de indexação apresenta um vocabulário básico de indexação para regulamentações federais publicadas no registro federal e no CFR (Ver mais em: <https://www.archives.gov/federal-register/cfr/thesaurus.html>. Acesso em: 15/mar./2025).

(MAO *et al.*, 2022) e sistemas de resposta a perguntas para responder a consultas usando *chatbots* e assistentes virtuais (OMAR *et al.*, 2023; AIT-MLOUK; JIANG, 2020). Como resultado do processo de transformação, o KG obtido do CFR2SBVR pode ser utilizado para impulsionar a eficiência de outros processos, como os discutidos nesta seção.

3.9.2 Representação do grafo de conhecimento no espaço vetorial

A representação vetorial dos nós e propriedades do KG é importante para a aplicação de métodos numéricos e computacionais, permitindo a resolução de problemas como reconhecimento de entidades e previsão de links. O conceito principal *Knowledge Graph Embedding* (KGE) consiste em mapear cada entidade e relação para um espaço vetorial. Nesse espaço, a distância e a direção entre vetores capturam as semelhanças e diferenças semânticas subjacentes: entidades e relações estreitamente relacionadas são posicionadas mais próximas umas das outras, enquanto as não relacionadas ficam mais distantes. Esse arranjo reflete as propriedades e a semântica do KG. O processo, frequentemente, envolve três etapas principais: codificação de entidades e relações em vetores, definição de funções de pontuação para quantificar a correção das relações inferidas dentro do KG e aplicação de um procedimento de otimização para refinar as representações vetoriais (ABU-SALIH, 2021).

3.9.3 Grafos de conhecimento como fonte da verdade para LLMs

A pesquisa de Agrawal *et al.* (2023) explora estratégias específicas dentro da inferência ciente de conhecimento, como a recuperação aumentada por KG e o raciocínio aumentado por KG, que utilizam, respectivamente, modelos de geração auxiliada por recuperação, como RAG e RALM, e métodos que simulam processos de raciocínio humano para melhorar a consciência contextual e as capacidades de raciocínio complexo dos LLMs. Além disso, discute-se o método de geração controlada por conhecimento, que gera conhecimento usando um modelo de linguagem e emprega técnicas para assegurar que as informações geradas estejam alinhadas com fatos, evitando a disseminação de informações falsas. Os autores também abordam o uso de grades de proteção (*guardrails*)

em IA gerativa, como restrições para controlar a geração de saídas dos LLMs, garantindo a sua utilização segura e protegida.

Pan *et al.* (2024) exploram a integração de LLMs e KGs para combinar as suas forças e atenuar as suas fraquezas a partir de três estruturas principais. A primeira, “LLMs aprimorados por KG”, envolve a integração de KGs durante o pré-treinamento e a inferência dos LLMs para enriquecer os modelos com conhecimento explícito, melhorando a precisão factual e a interpretabilidade. A segunda, “KGs aprimorados por LLMs”, utiliza LLMs para enriquecer KGs, facilitando tarefas como incorporação e expansão de KGs. A terceira, “LLMs + KGs combinados”, busca a sinergia entre as tecnologias para otimizar a representação do conhecimento e o raciocínio. Os desafios incluem a natureza opaca dos LLMs, a estrutura estática dos KGs e a complexidade de implementação. Esse enfoque combinado visa avançar a interpretabilidade e adaptabilidade das aplicações, potencialmente aumentando a complexidade computacional e enfrentando desafios de integração, mas oferecendo melhorias significativas na confiabilidade e profundidade para cenários complexos que demandam alto rigor.

No contexto legal, os KGs ajudam a organizar a complexa rede de regulamentos e termos, convertendo textos prolixos em formatos ricos em semântica e legíveis por máquinas. A partir de ontologias, termos e regras são extraídos e relacionados, enriquecendo o grafo com terminologias semanticamente semelhantes e sentenças deônticas, resultando em uma rede que captura os detalhes e interrelações complexas dos documentos legais. A presente pesquisa adota essa estratégia, usando o KG como entrada de dados e enriquecendo-o com metadados para prover fatos sobre as regras.

3.10 Geração aumentada por recuperação (RAG)

A RAG é uma técnica utilizada para melhorar o desempenho de modelos de linguagem pela integração de fontes de conhecimento externas durante o processo de geração. Ela envolve a combinação de um sistema de recuperação, como um banco de dados ou uma coleção de documentos, com um modelo

gerativo, tipicamente um modelo de linguagem baseado em rede neural (He X. *et al.*, 2024).

He X. *et al.* (2024) desenvolvem o modelo G-Retriever, uma solução avançada para tarefas de resposta a perguntas, combinando GNNs, LLMs e RAG. O modelo é especializado em processar grandes grafos textuais do mundo real e é projetado para mitigar o problema de alucinação, garantindo que as respostas sejam baseadas em fatos. O G-Retriever funciona em várias etapas: indexação de grafos com *embeddings* de nós e arestas, recuperação de componentes relevantes para a consulta e construção de um subgrafo otimizado através do algoritmo *Prize-Collecting Steiner Tree* (PCST). Finalmente, o subgrafo é convertido em texto e, com técnicas de *prompting*, o modelo gera respostas precisas e contextualmente apropriadas. Essa abordagem não só aprimora a precisão das respostas, mas também melhora, significativamente, a confiabilidade em tarefas complexas de consulta baseadas em grafos.

Lewis *et al.* (2020) definem o RAG como uma abordagem inovadora em NLP que aprimora significativamente as capacidades dos modelos de linguagem tradicionais, integrando-os com um mecanismo de recuperação externo. Essa integração permite que o RAG acesse e incorpore dinamicamente uma vasta gama de conhecimentos externos, armazenados em bancos de dados como a Wikipedia, nas suas saídas geradas. O RAG consiste em dois componentes principais: o recuperador, frequentemente utilizando um *Dense Passage Retrieval* (DPR) para buscar documentos semanticamente relevantes, e o gerador, tipicamente um modelo de linguagem pré-treinado como BART ou T5, que sintetiza o conteúdo recuperado com a consulta de entrada para gerar respostas ricas em contexto e precisas. O DPR, um componente-chave do RAG, aprimora o processo de recuperação pela sua arquitetura bi-codificador, no qual tanto consultas quanto os documentos são convertidos em representações vetoriais densas. Essa configuração permite uma recuperação mais precisa e baseada em semântica, comparada aos métodos tradicionais baseados em palavras-chave.

Esses métodos são úteis quando não há uma correspondência exata da pergunta com a resposta. No escopo da presente pesquisa, esses métodos

auxiliam na resolução de ambiguidades, quando o termo não corresponde exatamente ao que está no vocabulário, mas estão próximos no espaço vetorial.

3.11 Validação e testes

O processo de validação visa garantir que o artefato desenvolvido aborde efetivamente o problema identificado e atenda aos requisitos pré-definidos. Isso envolve confirmar a correção do artefato, avaliar a sua eficácia na resolução do problema pretendido, verificar a conformidade com padrões, avaliar a usabilidade para o público-alvo, garantir a confiabilidade em diferentes condições, testar métricas de desempenho como eficiência e escalabilidade, bem como examinar o seu impacto sobre os usuários e o domínio. O objetivo é fornecer evidências de que o artefato é adequado para o seu uso pretendido e contribui positivamente para o problema que foi projetado para resolver (WIERINGA, 2014). Esta seção discute os métodos para validar os artefatos do CFR2SBVR.

3.11.1 Validação e avaliação

Na DSR, Wieringa (2014) distingue entre validação e avaliação os processos essenciais para medir diferentes aspectos de um artefato. A validação busca assegurar que o artefato esteja corretamente alinhado com seu projeto pretendido e cumpra com os requisitos especificados, respondendo à questão: “Estamos construindo a coisa certa?”. Esse processo examina a correção interna e a completude do artefato em relação às suas especificações. As técnicas utilizadas na validação incluem prototipagem, verificação de modelo, simulação, revisão por pares e inspeções, todas visando verificar a correção interna e a completude do artefato.

Por outro lado, a avaliação investiga a eficácia e o impacto do artefato no contexto prático do mundo real, medindo o seu desempenho, usabilidade e sucesso geral na resolução do problema para o qual foi projetado. Este processo responde à pergunta: “Estamos construindo a coisa de maneira correta?”. As técnicas de avaliação, como testes de usabilidade, testes de campo, pesquisas e análises comparativas, coletam feedback do uso no mundo real para avaliar o desempenho e o impacto do artefato. Embora validação e avaliação tenham

objetivos distintos, ambos são processos orientados a objetivos e iterativos, que são fundamentais para garantir que o artefato não apenas atenda às especificações de projeto, mas também seja eficaz na prática, assegurando tanto a solidez técnica quanto a eficácia prática.

Esta pesquisa implementa o processo de validação e mesmo quando alguma abordagem ou técnica de avaliação for discutida, ainda será no âmbito do seu uso para a validação dos artefatos.

3.11.2 Métricas

As métricas para avaliação e validação variam com base na natureza do artefato e nos objetivos do projeto. As métricas típicas de validação incluem precisão, sensibilidade, completude, consistência, conformidade e taxa de erro, que avaliam a aderência do artefato às especificações e sua correção. As métricas de avaliação tipicamente incluem eficácia, eficiência, usabilidade, satisfação do usuário, escalabilidade, confiabilidade e impacto, que medem o desempenho do artefato em alcançar o seu propósito pretendido, experiência do usuário e efeito geral no domínio alvo. Essas métricas ajudam a garantir que o artefato atenda às especificações técnicas e às necessidades do usuário, contribuindo para o seu sucesso e eficácia nas aplicações do mundo real (WIERINGA, 2014).

3.11.3 Métodos

Haj *et al.* (2021a) apresentam um processo de validação para transformação automática de regras de negócio textuais em um modelo. Inicialmente, os dados são preparados a partir de três conjuntos de declarações de regras de negócio, cada um contendo declarações claras e sem ambiguidades. Em seguida, é realizada uma extração manual para estabelecer uma base comparativa, identificando elementos como conceitos de substantivos, verbos, definições, sinônimos e relações lógicas, de acordo com o padrão SBVR. Após isso, os autores executam uma extração automática dos mesmos elementos. A comparação entre os resultados automáticos e manuais foca na precisão dos elementos extraídos, utilizando métricas como *True Positive* (TP), *False Positive* (FP), *False Negative* (FN), precisão, sensibilidade e F1-Score, para avaliar o desempenho da abordagem automática. A análise dos resultados destaca as

áreas de sucesso e as necessidades de melhoria do processo, embora não detalhe um ciclo iterativo para aprimoramento contínuo da metodologia ou dos artefatos gerados.

Chen *et al.* (2023) exploram o uso de LLMs, como GPT-3.5 e GPT-4, para a modelagem de domínio em UML automatizada a partir de descrições textuais. Utilizando técnicas de engenharia de *prompts*, como aprendizado *zero-shot*, *N-shot* e *prompting* de cadeia de pensamento (*Chain of Thought* - CoT), os autores buscam aprimorar o desempenho desses modelos em gerar diagramas de classes. A avaliação é feita por um conjunto de dados ouro, contendo dez exemplos diversificados de um curso de programação orientada a modelos, cobrindo áreas como transporte, saúde e educação. Cada exemplo, acompanhado de uma descrição textual e um modelo de domínio de referência criado por especialistas, foca em elementos como classes, atributos e relações, omitindo componentes mais complexos, como operações e interfaces.

Wei *et al.* (2022) desenvolvem o método CoT, que melhora a capacidade de raciocínio dos LLMs, ao incluir passos intermediários, orientando o modelo a seguir uma sequência lógica para resolver problemas complexos. Em contraste com os métodos tradicionais de *prompting*, que usam pares simples de entrada e saída para tarefas de resposta direta, o CoT permite que o modelo de linguagem decomponha problemas em subetapas, facilitando a resolução de tarefas como aritmética, raciocínio de senso comum e manipulação simbólica. Isso torna o CoT particularmente eficaz em modelos com mais de 100 bilhões de parâmetros, no qual ele gera ganhos significativos de precisão e facilita a generalização para tarefas fora do domínio de treinamento.

Os resultados dos experimentos mostram que o CoT supera o *prompting* tradicional nas tarefas que exigem múltiplos passos de raciocínio, como o benchmark GSM8K para problemas matemáticos e o StrategyQA para inferências de senso comum. Além disso, o CoT oferece vantagens em interpretabilidade, permitindo que os pesquisadores entendam o caminho de raciocínio do modelo e identifiquem erros. Contudo, ele exige um grande poder computacional, especialmente nos modelos de larga escala, e a sua aplicação

em modelos menores é limitada, pois não exibe os mesmos ganhos de desempenho.

Wyner e Peters (2011) detalham um processo de validação comparando os resultados da transformação automática com um padrão ouro criado manualmente. Esse padrão ouro é constituído por anotações de especialistas em textos regulatórios, marcando elementos como obrigações e permissões. A metodologia automatizada processa os mesmos textos e as suas saídas são avaliadas contra o padrão ouro usando métricas de precisão e sensibilidade, indicando a exatidão e a abrangência da ferramenta automatizada. Com base nesta comparação, são identificados pontos fracos e fortes, permitindo refinamentos na metodologia. Esse processo de validação é iterativo, continuando até que os resultados desejados de precisão e sensibilidade sejam atingidos. Min *et al.* (2023) discutem métodos para gerar dados usando LLMs, que podem ser úteis para aumentar conjuntos de dados de treinamento ou criar dados sintéticos para tarefas com exemplos rotulados limitados.

Huang *et al.* (2024) investigam o uso de modelos LLMs para avaliar a performance de outros LLMs, conceito conhecido como *LLM-as-a-Judge*. Os autores analisam quatro modelos de julgamento afinados (JudgeLM, PandaLM, Auto-J e Prometheus), que, apesar de eficazes em domínios específicos, apresentam limitações em termos de generalização, imparcialidade e capacidade de avaliação detalhada. Para superar essas limitações, eles propõem o método CascadedEval, que combina esses modelos afinados com o GPT-4, de forma seletiva. O CascadedEval utiliza uma métrica de confiança: quando o modelo afinado está menos confiante na sua avaliação, o GPT-4 assume o julgamento, mantendo a precisão e reduzindo custos com APIs ao limitar o uso do GPT-4 para casos complexos. Dessa forma, o método aproveita a especialização dos modelos afinados junto à robustez do GPT-4, sem a necessidade de afinar um novo modelo.

Zheng *et al.* (2023) analisaram a viabilidade de usar LLMs, como o GPT-4, como juízes (*LLM-as-a-Judge*) para avaliar a qualidade das respostas de assistentes virtuais em interações abertas e multeturnos, superando as limitações dos *benchmarks* tradicionais. Foram desenvolvidos, para tanto, dois novos

benchmarks, o MT-Bench, com perguntas multeturnos para avaliar a capacidade de diálogo e instrução dos modelos, e o Chatbot Arena, uma plataforma *crowdsourced* na qual os usuários escolhem a resposta preferida entre dois modelos em tempo real. Esses benchmarks demonstraram que LLMs como juízes conseguem replicar preferências humanas com alta concordância, atingindo até 80% de acordo com avaliações humanas, o que sugere que podem ser uma alternativa escalável e eficiente para medir o alinhamento dos modelos com as expectativas dos usuários.

Apesar dos resultados promissores, a pesquisa identificou limitações e vieses nos LLMs ao atuar como juízes, incluindo o viés de posição (preferência por respostas apresentadas em uma determinada ordem), viés de verbosidade (preferência por respostas mais longas) e dificuldades com questões de matemática e raciocínio complexo. Para mitigar esses problemas, técnicas como inversão de posições, uso de respostas de referência e *prompts* com raciocínio em cadeia (CoT) foram empregadas, melhorando a precisão das avaliações. Os *benchmarks* mostraram que, embora o GPT-4 e o GPT-3.5 superem outros modelos como LLaMA e Vicuna, uma combinação de métricas tradicionais e avaliação de preferências humanas é necessária para uma avaliação abrangente das capacidades dos modelos.

Wataoka, Takahashi e Ri (2024) alertam sobre a técnica *LLM-as-a-Judge*, que usa LLMs para avaliar respostas de sistemas de diálogo de forma autônoma, o que permite análises rápidas e escaláveis sem depender, exclusivamente, de avaliadores humanos. No entanto, essa técnica enfrenta desafios, especialmente o viés de auto preferência, em que o modelo tende a favorecer respostas que ele próprio gerou ou com as quais está mais familiarizado. Esse viés foi identificado, por exemplo, no GPT-4, que avalia as suas respostas mais favoravelmente do que as avaliações humanas indicariam. Para medir e entender esses vieses, métricas baseadas em conceitos de justiça algorítmica, como igualdade de oportunidade, foram propostas para tornar as avaliações mais justas e reduzir as influências de preferência interna do modelo.

Além do viés de auto preferência, LLMs também apresentam outros tipos de vieses, incluindo o viés de posição, em que as respostas em certas posições são

preferidas, o viés de verbosidade, que favorece respostas mais longas, e o viés de perplexidade, que prioriza respostas de baixa perplexidade. Outros vieses, como demográfico e de estilo, também podem influenciar as avaliações, refletindo desigualdades presentes nos dados de treinamento. Para mitigar esses problemas, estratégias como a rotação de posições nas respostas, uso de múltiplos modelos e ajustes nos prompts de entrada são adotadas, permitindo uma avaliação mais equilibrada e imparcial.

Esta pesquisa usa uma combinação desses métodos que é apresentado no Capítulo “7 VALIDAÇÃO”.

3.12 Resumo

As abordagens formais, como o SBVR, melhoram a eficiência das organizações em cumprir obrigações regulatórias, ao definirem vocabulários de negócios e regras, de forma clara. A utilização de ontologias, especialmente a FIBO no domínio financeiro, auxilia na desambiguação e estruturação do conhecimento, minimizando erros de interpretação. Além disso, técnicas de NPL, reforçadas por LLMs, engenharia de *prompt* e grafos de conhecimento aprimoram a interpretação e transformação de regulamentos escritos em NL. Porém, a validação dos resultados é um problema da mesma magnitude da execução, que exige sofisticação para aferir a qualidade.

4 DISCUSSÃO DAS SOLUÇÕES

4.1 Padrões para comparação das abordagens de transformação

Esta seção propõe e apresenta padrões (*patterns*) para classificar e comparar abordagens existentes de transformação de textos regulatórios em linguagem controlada. Essa comparação auxilia a identificação sistematizada de aspectos das soluções analisadas permitindo o estabelecimento de um conjunto de requisitos para a solução elaborada na presente pesquisa.

Padrões costumam ser descritos como soluções para problemas que se repetem. Eles têm o objetivo de auxiliar outros a aprender com as soluções existentes e possibilitar a aplicação dessas soluções em contextos semelhantes. Geralmente, um padrão é nomeado, o que funciona como um identificador conceitual para simplificar discussões acerca das informações que ele abrange, segundo Alexander (1979):

Esses padrões em nossas mentes são, mais ou menos, imagens mentais dos padrões no mundo: são representações abstratas das próprias regras morfológicas que definem os padrões no mundo. No entanto, em um aspecto, eles são muito diferentes. Os padrões no mundo simplesmente existem. Mas os mesmos padrões em nossas mentes são dinâmicos. Eles têm força. Eles são geradores. Eles nos dizem o que fazer; eles nos dizem como nós devemos, ou podemos, gerá-los; e eles nos dizem também que, sob certas circunstâncias, devemos criá-los. Cada padrão é uma regra que descreve o que você tem que fazer para gerar a entidade que ele define. (ALEXANDER, 1979, p. 181, tradução do autor deste trabalho).

Os padrões aqui propostos foram derivados das soluções estudadas, que variam bastante nas suas abordagens. Entretanto, há um conjunto comum de problemas que a maioria das abordagens tenta resolver. Para auxiliar a comparação, elencou-se um conjunto de sete padrões que representam problemas comuns no processo de transformação. Os elementos descritos de cada padrão foram inspirados em Gamma *et al.* (1995). Cada padrão está definido como uma tupla: Identificador, Nome, Contexto, Problema, Solução, Consequência e Exemplo.

P1. Processamento de documento

Contexto: Documentos variam em formato e estrutura. Muitas vezes essas formas têm carga semântica para os trechos, sendo suas conexões implícitas.

Além da forma, é comum que trechos façam explicitamente referência a outras trechos.

Problema: Definir ou adotar uma ontologia para decompor um ou mais documentos em trechos preservando sua semântica na estrutura do documento.

Solução: Identificar e extrair os trechos do texto e relacioná-los de acordo com a sua estrutura. Buscar por referências explícitas a outros trechos e conectá-las.

Consequências: A incorreta representação do documento pode ocultar ou modificar o significado do trecho podendo levar a uma interpretação equivocada do seu significado. (LOUIS *et al.*, 2023)

Exemplo: Na CFR os subparágrafos e/ou parágrafos, constituem os trechos indivisíveis e são subordinados a seção / subseção, parte / subparte, capítulo / subcapítulo, título e geralmente fazem referência a outros trechos.

P2. Extração de regras das sentenças

Contexto: Nem todos os trechos constituem ou contêm regras.

Problema: Identificar os trechos que contêm regras.

Solução: Uso de NLP e algoritmos para identificar quais trechos representam regras e classificá-las em comportamentais ou de definição.

Consequência: A correta classificação de um trecho tem impacto direto na qualidade da transformação.

Exemplo: Na CFR um trecho pode conter o texto do regulamento, que é a parte principal onde normalmente se encontra o que é requerido, permitido ou proibido para aquele regulamento, por outro lado também pode conter definições, exceções e isenções, penalidade e fiscalização, além de outras subseções, parágrafos e subparágrafos.

P3. Extração de conceitos

Contexto: Os conceitos presentes nos textos podem ser conceitos gerais (hiperônimos), próprios ou dependentes do domínio.

Problema: Identificar e extrair os conceitos dos trechos e classificá-los quanto ao seu tipo (substantivo, verbo, característica, individual, geral, quantidade, papéis, estados etc.), e posteriormente associá-los a outros conceitos.

Solução: Uso de NLP marcando partes-da-fala é comumente utilizado para identificar entidades que podem ser relacionadas ou usadas para criar dicionários de termos.

Consequência: A interpretação de um trecho depende da correta interpretação das entidades presentes no texto. Essa interpretação é ainda mais crítica para conceitos do domínio, que podem ser confundidos com conceitos gerais e terem significados completamente distintos.

Exemplo: Na CFR título 17, capítulo 2, parte 230, seção 135, subseção a, encontram-se termos como anúncio genérico e companhias de investimento, com significado particular para o contexto do regulamento.

P4. Relacionamento entre os conceitos

Contexto: O mesmo conceito pode ser representado de várias formas.

Problema: Identificar conceitos que estão relacionados com outros conceitos, seja por sinonímia ou por generalização.

Solução: Identificar e relacionar conceitos com seus sinônimos e/ou generalização.

Consequência: O não entendimento de sinônimos ou generalizações podem levar a interpretações diferentes da mesma regra.

Exemplo: Na CFR título 17, capítulo 2, parte 230, seção 135, o termo anúncio é sinônimo de nota, circular, carta ou comunicação. Da mesma forma, o termo valores imobiliário refere-se a uma ampla gama de instrumentos financeiros, tais como ações, títulos, opções e fundos mútuos. Um outro exemplo é entidade regulada que se refere a bancos, companhias de investimentos e corretoras.

P5. Desambiguação de conceitos

Contexto: Um termo pode ter mais de um significado dependendo do seu contexto.

Problema: Identificar o contexto e relacionar o termo ao seu significado nesse contexto.

Solução: A identificação de um termo específico a um domínio e o uso de dicionário do domínio para associar ao significado correspondentes.

Consequência: Não interpretar um conceito de acordo com seu domínio pode levar a uma interpretação equivocada da regra.

Exemplo: Na CFR título 17, capítulo 2, parte 230, seção 100, subseção a parágrafo um, define que o termo “comissão” se refere a SEC e no parágrafo dois que “ato” se refere a lei de valores imobiliários de 1933. Para a interpretação dos trechos relacionados a esse regulamento é necessário entender esses termos.

P6. Extração de fatos

Contexto: Fatos são trechos com afirmações sobre conceitos que são verdadeiras ou falsas (mas não ambas).

Problema: Identificar e extrair fatos dos trechos.

Solução: Uso de NLP e algoritmos para identificar quais trechos representam fatos e anotá-los.

Consequência: Um fato é um elemento fundamental na definição do vocabulário e das regras de negócios. Ele serve como um bloco de construção para criar regras de negócios e restrições. A incorreta identificação e interpretação levam a regras inconsistentes ou incorretas.

Exemplo: No CFR, os fatos são frequentemente apresentados como regulamentações, requisitos, definições ou outras declarações que se baseiam em autoridade legal e são destinadas a ser fatuais e executáveis. O trecho do Título 17, Capítulo II, Parte 240, Regra 240.10b-5 (Emprego de Dispositivos Manipulativos e Enganosos) declara que “... uso de práticas manipulativas ou enganosas em conexão com a compra ou venda de títulos é ilegal ...”.

P7. Transformação das regras para linguagem controlada

Contexto: Uma linguagem controlada reduz ou elimina a ambiguidade característica da linguagem natural.

Problema: Transformar o trecho que representa uma regra de definição ou comportamental para um formato não ambíguo e interpretável por humanos e máquinas.

Solução: O uso de algoritmos e o mapeamento para a sintaxe da linguagem controlada são geralmente empregados.

Consequência: O resultado tem como objetivo reduzir ou eliminar as incertezas sobre a interpretação. Erro na transformação irá criar uma regra equivocada.

Exemplo: Para o trecho 240.10b-5: “*It shall be unlawful for any person, directly or indirectly, by the use of any means or instrumentality of interstate commerce, or of the mails, or of any facility of any national securities exchange to employ any device, scheme, or artifice to defraud*” é transformada para “**It is prohibited for any person to use any means or instrumentality of interstate commerce, or of the mails, or of any facility of any national securities exchange to employ any device, scheme, or artifice with the intent to defraud**”. Em que os termos em sublinhado são conceitos já definidos no dicionário terminológico.

Quadro 7 - Requisitos

Requisito	Descrição	Componentes
R1 - Representação do documento legal (P1)	Identificar e extrair os trechos do texto e relacioná-los de acordo com a estrutura do texto e referências.	6.2.1 KG para CFR e FIBO
R2 - Extração de regras das sentenças (P2, P5)	Identificar os trechos dos textos legais que contêm regras.	6.2.3 Anotações semânticas – Algoritmo “extract / classify elements”
R3 - Dicionário de conceitos (P3 e P4)	Identificar conceitos específicos do domínio para criar dicionário de termos e relacioná-los aos seus sinônimos e/ou generalização.	6.2.3 Anotações semânticas – Algoritmo “extract / classify elements”

Requisito	Descrição	Componentes
R4 - Regras não ambíguas (P5, P7)	As regras em NL devem ser transformadas para uma linguagem controlada, interpretável por humanos e máquinas.	6.2.3 Anotações semânticas – Algoritmo “taxonomy classification and templates” 6.2.4 Transformação
R5 - Catálogo de regras	As regras devem ser armazenadas em um catálogo.	6.2.5 Catálogo de regras
R6 - Rastreabilidade	As regras do catálogo devem ser rastreáveis até sua origem no texto legal e vice-versa.	6.2.3 Anotações semânticas – Algoritmo “extract / classify elements”
R7 - Ferramenta de apoio	Incluir ferramentas de apoio para a realização das transformações.	6 FERRAMENTAS DE SUPORTE

Fonte: Elaborado pelo autor.

Esses padrões são utilizados para comparar as abordagens e resumem os requisitos importantes a serem considerados ao transformar textos em linguagem natural para linguagem controlada. O Quadro 7 apresenta os requisitos que foram identificados como necessários ao realizar a comparação das abordagens existentes e os relaciona com os componentes do CFR2SBVR.

Os identificadores P1 a P7, em cada requisito, indica a origem do requisito em relação ao padrão discutido. Esses requisitos representam as principais necessidades das partes interessadas que estão preocupadas com transformação dos textos legais, tais como auditores, gestores de governança e avaliadores de riscos. Os requisitos identificados foram elencados com base nos resultados da avaliação das abordagens e das observações empíricas do domínio de GRC.

Os requisitos R1 a R6 são identificados como funcionais, enquanto o R7 é destacado como um requisito de infraestrutura, uma vez que a ausência de ferramentas adequadas é uma causa comum de falha em diversas metodologias. No entanto, esses requisitos não abordam todos os desafios relacionados à gestão da conformidade. Tais desafios englobam a necessidade de interpretar normas abertas para aplicá-las em processos de negócios específicos, a coordenação entre as várias disciplinas envolvidas na

conformidade e a adaptação às mudanças constantes nas leis e nos processos empresariais. Esses desafios evidenciam a complexidade de garantir a conformidade regulatória e a necessidade de sistemas e fluxos de trabalho que deem suporte à interpretação legal, mapeiem normas para processos empresariais e gerenciem alterações na legislação e nas operações de negócios (BOELLA, 2013). Contudo, a catalogação e a formalização das regras simplificam o mapeamento e a interpretação das normas, além de facilitar a comunicação e o dinamismo da lei, pela automação do processo. Assim, a abordagem adotada nesta pesquisa contribuí para a GRC, bem como para os objetivos empresariais de permanecerem em conformidade com certo custo ideal e exposição mínima a riscos.

4.2 Abordagens de transformação de NLP para SBVR

Compara-se a seguir os trabalhos para transformação de regras de negócio para SBVR encontrados na literatura pesquisada. Com base nos padrões apresentados na seção anterior, as abordagens dos autores estudados foram resumidas no Quadro 8 e as que mais influenciaram esta pesquisa são discutidas em detalhes a seguir.

Quadro 8 - Comparação das abordagens que transformam regras de negócio para SBVR

Autores	Resultado	Método	Padrões implementados	Depende de	Nível automação
Abi-Lahoud <i>et al.</i> (2013)	Business vocabulary and rules	Baseado em regras	1-7		Manual
Bajwa <i>et al.</i> (2017)	Business vocabulary and rules	NLP + baseado em regras	2, 3, 6, 7	Diagrama de classes UML	Semi
Chittimalli <i>et al.</i> (2020)	Business vocabulary and rules	NLP + baseado em regras	2-7		Auto
Haj <i>et al.</i> (2021a)	Business vocabulary and rules	NLP + baseado em regras	2-7		Auto

Autores	Resultado	Método	Padrões implementados	Depende de	Nível automação
Joshi e Saha (2020)	Extração das regras e grafo para QA	NLP + baseado em regras	1-7		Auto
Polonio, 2018	Business vocabulary and rules	Baseado em regras	1-7	BPMN, UML	Manual
Roychoudhury <i>et al.</i> (2017)	Business vocabulary and rules	NLP + Baseado em regras	2-4, 6, 7	Dicionário	Semi
Skersys <i>et al.</i> (2022)	Business vocabulary and rules	Baseado em regras	1-3, 6, 7	Diagramas BPMN	Semi
Nossa abordagem	Business vocabulary and rules	NLP + Modelo probabilístico	1-7	FIBO	Auto

Fonte: Elaborado pelo autor.

A pesquisa bibliográfica indicou outros autores que também podem contribuir para esta pesquisa. Eles serão analisados posteriormente durante o refinamento do capítulo “5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR”, são eles: Bhattacharyya *et al.* (2017), Bajwa *et al.* (2011), Chittimalli *et al.* (2019a), Danenas *et al.* (2020), Dragoni *et al.* (2018), Haj *et al.* (2018), Haj *et al.* (2019), Haj *et al.* (2021b), Haj *et al.* (2021c), Haj *et al.* (2021a), Ramzan *et al.* (2014), Skersys *et al.* (2018), Skersys *et al.* (2021), Skersys *et al.* (2018), Skersys *et al.* (2021), Selway *et al.* (2015) e Wyner e Peters (2011).

Abi-lahoud (2013) - Interpreting Regulations with SBVR

Abi-lahoud *et al.* (2013) interpretam regulamentações dos EUA utilizando o SBVR, partindo de um conceito comprovado pelo *Governance, Risk, and Compliance - Technical Center* (GRCTC) na Irlanda. Eles detalham o protocolo para decompor regulamentações em partes gerenciáveis para análise por SMEs, utilizando ferramentas de software especializadas. O objetivo é melhorar a acessibilidade e precisão de textos regulatórios complexos para um melhor

entendimento de conformidade e governança. Esse protocolo foi base para esta pesquisa e atende a todos os padrões discutidos, embora a abordagem seja completamente manual.

No processo desenvolvido pelos autores, os SMEs são os responsáveis pela interpretação, estruturação e validação do texto legal e sua transformação em SBVR. O protocolo consiste em:

1. Seleção e leitura de texto - SMEs selecionam e leem o texto legal para entender o conteúdo, contexto e implicações;
2. Identificação de modalidades - SMEs identificam modalidades como obrigações, permissões e proibições no texto. Para cada modalidade identificada no texto regulatório, o protocolo exige a adição da palavra-chave da modalidade relevante, a estilização do verbo em inglês no estilo SBVR, a identificação e estilização de conceitos de substantivo ou frases verbais envolvidas, e a incorporação desses elementos no formato estruturado SBVR;
3. Extração e definição de vocabulário - SMEs extraem o vocabulário de negócios relevantes e definem termos para refletir seu significado dentro do contexto legal. Para cada conceito de substantivo listado, inicia-se o enriquecimento identificando suas características específicas e outros elementos definidores;
4. Estruturação de regras - SMEs estruturam as regras identificando cláusulas condicionais e resultados e usam o SE do SBVR para expressar cada regra;
5. Estilização SBVR - SMEs aplicam a estilização SBVR aos verbos e conceitos de substantivos identificados;
6. Formalização - SMEs formalizam as regras usando uma linguagem formal baseada em SBVR para garantir precisão e legibilidade por máquina;
7. Validação - SMEs validam as representações SBVR para garantir precisão e reflexo da intenção do texto legal;

8. Revisão e refinamento - SMEs revisam as regras SBVR para completude, precisão e conformidade com o texto legal original e refinam conforme necessário.

Os autores colocam como trabalhos futuros o desenvolvimento de ontologias e modelos de dados para a indústria financeira utilizando vocabulários SBVR para aprimorar a automação, reduzindo a necessidade de intervenção manual de especialistas. Eles sugerem que essa abordagem poderia tornar as regulamentações mais acessíveis e compreensíveis, integrando a expertise dos especialistas com a precisão do SBVR na representação do conhecimento, reduzindo as incertezas e imprecisões, facilitando o compartilhamento e expansão do entendimento sobre os requisitos regulatórios e proporcionando aos reguladores uma visão mais clara de como as normas são percebidas pelos regulados.

Bajwa and Shahzada (2017) - Automated generation of OCL constraints: NL based approach vs pattern based approach

Em continuidade ao trabalho de Bajwa *et al* (2011), a abordagem de Bajwa and Shahzada (2017) para a geração automatizada de restrições *Object Constraint Language* (OCL) a partir de especificações em NL, utiliza o SBVR como um passo intermediário. A abordagem envolve a análise linguística do texto em NL, incluindo marcação de PoS e análise baseada em regras, para extrair elementos básicos do SBVR (padrões P2, P3, P6 e P7). Esses elementos são então transformados em sentenças OCL usando transformações de modelo. A abordagem tem algumas limitações, incluindo sua restrição ao domínio do modelo de classe UML, os nomes do vocabulário usados nas restrições devem ser consistentes com os nomes do vocabulário usados no modelo de classe UML e há suporte apenas para certos tipos de sentenças OCL. Os autores não lidam com a representação do documento (padrão P1), desambiguação e generalização de conceitos (padrões P4 e P5).

Chittimalli *et al.* (2020) - An approach to mine SBVR vocabularies and rules from business documents

Chittimalli *et al.* (2020) ampliam a abordagem de Bhattacharyya *et al.* (2017) ao mapear essas intenções para um formato de regra formalizado, que é adequado para processamento automatizado e verificação em sistemas de negócios. A abordagem não supervisionada de Chittimalli *et al.* (2020) para extrair vocabulários e regras SBVR de documentos de negócios é dividida em quatro partes: extração de sentenças (padrão P2), extração de entidades, extração de fatos e mineração de regras (padrões P3, P6 e P7). A abordagem utiliza técnicas de NLP, incluindo um analisador de árvore de dependência, para identificar e extrair informações relevantes dos documentos. O método foi testado em dois conjuntos de dados: o exemplo do Anexo G da especificação SBVR EU-Rent²⁹ e o documento *Reserve Bank of India (RBI) – Know Your Customer KYC*³⁰, mostrando resultados promissores em termos de precisão e sensibilidade. Os autores sugerem que essa abordagem pode reduzir o esforço manual necessário na interpretação de documentos de negócios e melhorar a precisão da extração de regras. Uma limitação significativa da abordagem é que ela não lida com a característica de correferência. O analisador SpaCy usado no sistema não permite correferência. Também a identificação de sinônimos e relações de generalização foram ignoradas (padrões P4 e P5). Diferente das demais abordagens, a forma como as sentenças são classificadas como sentenças de regra ou ruído (padrão P2) usando um modelo de linguagem N-gram, porém os autores tratam as sentenças como independentes (padrão P1).

Haj *et al.* (2021a) - The semantic of business vocabulary and business rules: An Automatic Generation From Textual Statements

Haj *et al.* (2021a) apresentam um método automático para transformar regras de negócios em NL para SBVR utilizando NLP para analisar linguisticamente os textos de negócios, incluindo análises lexicais, sintáticas e semânticas para

²⁹ <https://www.omg.org/cgi-bin/doc?formal/19-10-06.pdf>

³⁰ <https://www.rbi.org.in/CommonPerson/english/scripts/notification.aspx?id=2607>

entender a estrutura gramatical e o significado da linguagem (padrão P2). O método enfatiza a extração de um dicionário terminológico (padrões P3, P4 e P5) e formulações semânticas do texto, levando à identificação de regras de negócios (padrão P6). Essas regras são mapeadas para elementos SBVR por meio de correspondência de padrões que capturam uma variedade de estruturas de frases e cláusulas, tanto explícitas quanto implícitas, incluindo aquelas com condições, negações e quantificações (padrão P7). Entretanto, o método tem como principal objetivo obter o dicionário terminológico, concentrando-se em regras que fornecem informações ou restrições sobre conceitos. Embora seja eficaz no tratamento de regras explícitas e bem estruturadas, enfrenta limitações ao lidar com regras complexas, ambíguas ou implícitas que requerem conhecimento profundo do domínio. O método também exclui regras que expressam processos de negócios ou cálculos, dependendo de algoritmos para transformação, o que pode limitar sua flexibilidade e adaptabilidade. Apesar dessas limitações, o método apresentou eficácia com uma pontuação F1-Score média de 87%, indicando sua capacidade de capturar o significado das declarações de regras de negócios. Os autores não tratam a questão da representação da hierarquia das cláusulas no documento (padrão P1), tratando todas como independentes.

Joshi e Saha (2020) - A Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations

Joshi e Saha (2020) convertem regulamentações do CFR em um formato estruturado e legível por máquinas. Primeiro, utiliza técnicas de aprendizado profundo para identificar e extraír termos e definições relevantes e relacioná-los (padrões P3, P4), assim como regras do conteúdo textual do CFR (padrão P2, P6). Esses elementos são então organizados em um KG enriquecido semanticamente, destacando as relações e dependências entre diferentes regulamentações (padrão P1). Esse grafo é um artefato fundamental, possibilitando o raciocínio automatizado sobre as regulamentações, o que pode apoiar esforços de conformidade, análise legal e o desenvolvimento de aplicações regulatórias. Esse processo foi aplicado ao Título 48 da CFR para

ilustrar o seu valor prático. Esse arcabouço é um passo importante na presente pesquisa, pois demonstra o uso de uma ontologia para representar a CFR, embora não tendo como objetivo transformar os regulamentos em regras SBVR (padrão P7), mas prover uma base para um sistema de perguntas e respostas. Os autores também endereçam a ambiguidade dos termos relacionando-os entre si (padrão P5), criando um dicionário de termos a partir do próprio corpus. A validação dos artefatos é realizada por SMEs.

Os autores exploram diversas técnicas de NLP para analisar textos regulatórios e contratuais. Utilizam métodos baseados em regras com expressões regulares, POS e caracteres curinga e técnicas avançadas como *Recurrent Neural Network* (RNN) e LLM. Nos experimentos E1 e E2, eles empregam o modelo Word2Vec³¹ com a camada de incorporação treinada em conjuntos de dados específicos, enquanto no E3 a camada de incorporação é reiniciada e as camadas internas são congeladas para capturar semelhanças entre regulamentos e contratos. O LLM foi ajustado em duas fases: pré-treinamento com tarefas de mascaramento e predição de sentenças, seguido de um ajuste fino em dados de regulamentos. Esse modelo, aplicado a dados contratuais desconhecidos, demonstrou alto desempenho, confirmando similaridades linguísticas e taxonômicas entre regulamentos e contratos e destacando-se nas categorias de obrigação, permissão e proibição devido à sua arquitetura de transformador (*transformer*) que permite uma compreensão profunda do estilo de escrita das cláusulas.

Polonio (2018) - Método para obtenção de SBVR com definições bilíngues

Polonio (2018) aborda o desenvolvimento de um método para a obtenção de definições bilíngues em SBVR aplicado ao setor de seguros no Brasil, motivado pelo crescimento desta indústria e pelas mudanças regulatórias subsequentes à Lei Complementar Nº 126 de 2008. Este cenário regulatório em transformação, aliado à abertura do mercado a competidores nacionais e internacionais, evidencia a necessidade de sistemas de informação robustos e adaptáveis, que

³¹ <https://en.wikipedia.org/wiki/Word2vec>

suportem a dinâmica do setor e facilitem a conformidade regulatória. O trabalho propõe a utilização do SBVR juntamente com técnicas de conversão para ontologia OWL2 criando um conjunto consistente de regras de negócio, essenciais para a governança corporativa e a gestão de riscos. O SBVR atuará como base de uma língua franca permitindo transformações entre diversos tipos de linguagens sem perda de significado.

Esse método transforma vários tipos de documentos, tais como: diagramas de BPMN, UML, OCL e documentos textuais. Para os documentos textuais, o autor aplica o protocolo de Abi-lahoud (2013) para obter as regras. Esse é o mesmo protocolo adotado por esta pesquisa. Por fim, os conceitos são traduzidos e o vocabulário e regras de negócio são convertidos para OWL2.

Roychoudhury *et al.* (2017) - From natural language to SBVR model authoring using structured english for compliance checking

Roychoudhury *et al.* (2017) apresentam uma abordagem semiautomatizada para transformar texto jurídico em NL para SE para verificação de conformidade no setor financeiro. O processo envolve várias etapas, começando com a detecção de sentenças e extração de proposições (padrões P2, P6) usando a *Clause-Based Open Information Extraction* (ClausIE)³², uma ferramenta de código aberto, seguido pela seleção e modificação de proposições com base em um modelo de domínio e dicionário. Finalmente, as proposições são transformadas para a sintaxe SE desejada (padrão P7). A abordagem utiliza Aprendizado de Máquina (ML) / NLP para extrair o modelo de domínio e dicionário do texto regulatório (padrão P3), auxiliando na identificação de conceitos, definições, sinônimos e relações especializadas (padrão P4). Apesar de oferecer uma maneira promissora de transformar texto legal em um formato estruturado para verificação de conformidade, a abordagem tem limitações, como a dependência do ClausIE e necessidade de intervenção manual para a transformação. No entanto, a abordagem fornece uma ferramenta para especialistas do domínio

³²<https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/clausie>

criarem regras regulatórias em SE sem precisar entender as especificações formais subjacentes, simplificando assim o processo de verificação de conformidade no setor financeiro. No entanto, os especialistas do domínio precisam conhecer a notação de SE para usar o editor SE fornecido para editar sugestões ou adicionar informações adicionais, como sinônimos e definições. Os autores consideram as cláusulas independentes (padrão P1) e não abordam a questão da desambiguação de termos (padrão P5).

Na abordagem dos autores, há uma transformação de NL para SE de forma semiautomática e de SE para SBVR automática. No primeiro passo da transformação de NL para SE, o texto legal é dividido em um conjunto de sentenças. Este processo permite a invocação opcional de rotinas de pré-processamento que podem auxiliar na limpeza do texto, como a remoção de cabeçalhos e rodapés. Para essa detecção de sentenças foram utilizadas heurísticas provenientes de uma ferramenta. No passo seguinte, são extraídas proposições n-árias de cada sentença obtida no texto usando a ferramenta ClausIE. Por fim, o especialista do domínio tem a capacidade de selecionar as proposições para consideração futura usando uma noção de “informatividade” baseada nas ocorrências das menções nas proposições. A “informatividade” de uma proposição é calculada em termos da proporção do número total de tokens da proposição para o número total de menções.

$$\text{Informatividade} = \frac{\text{Número de menções}}{\text{Número total de tokens}}$$

Suponha que temos a proposição: “Customer must provide valid_ID”. Se “client” e “valid_ID” são menções reconhecidas do modelo de domínio, então temos 2 menções. A proposição tem 4 tokens (Customer, must, provid e valid_ID), a proporção de informatividade seria:

$$\text{Informatividade} = \frac{2}{4} = 0,5$$

Portanto, a informatividade da proposição é 0,5 e quanto mais próximo de 1, mais informativa é a proposição, pois indicaria que uma proporção maior de tokens na proposição corresponde a menções no modelo de domínio. Na

transformação de SE para SBVR, cada elemento de SE é mapeado para elementos correspondentes no SBVR, mas os autores não apresentam esse mapeamento.

Skersys et al. (2022) - Transforming BPMN Processes to SBVR Process Rules with Deontic Modalities

As regras de negócio podem estar codificadas em um modelo intermediário. Para este cenário, a metodologia Skersys et al. (2022) para converter diagramas de BPMN em regras SBVR utiliza uma abordagem que incorpora modalidades deôntica para definir regras normativas dos processos de negócios. O processo começa com a identificação dos elementos-chave do BPMN, como tarefas, eventos e fluxos de sequência (padrão P1), e seu mapeamento para os conceitos correspondentes do SBVR. Esse mapeamento preserva a semântica do processo de negócio na transição de representações gráficas para verbais. A lógica deôntica é então aplicada a esses elementos para especificar obrigações, permissões e proibições, levando à definição formal de regras de negócios em NL estruturada (padrão P2). A metodologia emprega tecnologia de transformação de modelo para modelo, utilizando um conjunto de regras de transformação e algoritmos para automatizar o processo de conversão e garantir precisão e consistência na tradução de BPMN para SBVR (padrão P7). Ela exige o uso de elementos de *pool/lane* do BPMN para definir participantes. As convenções de nomenclatura para elementos de modelo e condições em gateways são críticas para gerar regras de processo válidas. Essas limitações exigem uma preparação e estruturação cuidadosas dos modelos BPMN para garantir uma transformação bem-sucedida. A abordagem concentra-se em um subconjunto de conceitos do BPMN que são relevantes para a formação de regras de processo, o que limita o escopo dos processos de negócio que podem ser traduzidos.

Na abordagem dos autores, a transformação garante que cada conceito em BPMN seja mapeado para um conceito equivalente em SBVR (padrões P2, P3),

proporcionando uma transposição sem perda de informação, mantendo a integridade do nível de detalhe original.

A transformação de modelo geral (T), que representa um mapeamento entre a estrutura do BPMN de entrada (PP_i) e a estrutura de saída SBVR (RP_i), define para cada regra uma transformação específica (T_i), com $T_i: PP_i \rightarrow RP_i$, com i variando de 1 a 9, sendo:

1. **Um fluxo de sequência (PP1):** Envolve etapas básicas do processo conectadas por um fluxo de sequência, com a conclusão de uma etapa seguindo a outra em uma ordem especificada. O padrão de regra SBVR (RP_1) traduz isso em uma regra que afirma ser obrigatório que a atividade ou evento subsequente ocorra após o anterior.
2. **Gateways exclusivos divergentes e/ou convergentes (PP2):** Lida com pontos de decisão que o processo diverge em diferentes caminhos com base em determinadas condições (*gateways exclusivos*). As regras SBVR correspondentes (RP_2) especificam ser obrigatório que as atividades ou eventos ocorram após o *gateway*, dependendo das condições que são atendidas.
3. **Gateways inclusivos divergentes e/ou convergentes (PP3):** Semelhante aos *gateways exclusivos*, os *gateways inclusivos* permitem que múltiplos caminhos sejam tomados, todavia mais de um caminho pode ser tomado simultaneamente. As regras SBVR (RP_3) refletem isso ao afirmar ser obrigatório que todas as atividades ou eventos subsequentes possíveis ocorram após o *gateway inclusivo*, com base nas condições.
4. **Gateway baseado em evento divergente (PP4):** Usado quando o fluxo do processo é determinado por eventos em vez de atividades. A regra de transformação (RP_4) dita que é obrigatório que as atividades orientadas a eventos ocorram em resposta aos eventos desencadeadores.
5. **Gateways paralelos divergentes e/ou convergentes (PP5):** Usado quando as atividades podem ocorrer simultaneamente, representadas por *gateways paralelos*. As regras SBVR (RP_5) afirmam ser obrigatório que

todas as atividades concorrentes sejam concluídas para que o processo prossiga.

6. **Evento de limite bloqueante / não bloqueante (PP6):** Inclui eventos que interrompem ou não o fluxo atual. As regras SBVR (RP6) refletem isso ao afirmar ser permitido ou obrigatório que certas atividades ocorram quando esses eventos acontecem.
7. **Duas variações semanticamente equivalentes de uso de objeto de dados com: (a) associação de dados e (b) associação (PP7):** Descreve como os objetos de dados são usados dentro do processo, seja com associação direta de dados ou uma associação mais geral. As regras SBVR (RP7) especificam as obrigações ou permissões em relação ao manuseio desses objetos de dados.
8. **Duas variações semanticamente equivalentes de uso de armazenamentos de dados com: (a) associação de dados e (b) associação (PP8):** Semelhante aos objetos de dados, descreve o uso de armazenamentos de dados e como eles estão associados a atividades ou eventos no processo. As regras SBVR (RP8) definem as condições necessárias para que os dados sejam recuperados ou armazenados.
9. **Composto pelas variações de “participante como uma caixa-preta” e “participante como uma caixa-branca” com fluxos de mensagens de entrada e saída (PP9):** Lida com a interação entre diferentes participantes no processo, em que os detalhes internos do processo podem ou não ser visíveis (caixa-preta vs. caixa-branca). As regras SBVR (RP9) declaram as permissões ou obrigações para enviar e receber mensagens entre os participantes.

Embora o experimento realizado pelos autores tenha mostrado que as regras e algoritmos de transformação (padrão P7) foram eficazes e cobriram todo o fluxo de trabalho especificado, o estudo reconhece várias limitações em sua metodologia que poderiam impactar a validade dos resultados. A validade de construção, que assegura que a abordagem seja aplicada corretamente e compreendida uniformemente, enfrentou riscos de interpretação errônea dos

conceitos dentro do domínio do problema. No entanto, a automação das regras e algoritmos de transformação ajudaram a mitigar esses riscos. A validade interna também foi uma preocupação, com a possibilidade de erro humano e subjetividade influenciando o processo. A seleção manual dos modelos de origem para o experimento e o número limitado de modelos (32) podem comprometer a generalização dos resultados. Além disso, a ausência de um consenso sobre práticas de modelagem padronizadas também foi identificada como fator que poderiam afetar negativamente os resultados.

Considerações adicionais

Nenhuma das abordagens estudadas usa a FIBO, embora ela exista desde 2008. O não uso de uma ontologia como FIBO, tem a desvantagem de lidar apenas com os termos presentes no corpus em questão, reduzindo a probabilidade de reutilização. Como foi visto, a FIBO desempenha um papel essencial na definição de conceitos de negócios na indústria de serviços financeiros. Sua padronização pela OMG tem o propósito de fornecer um recurso padrão para uso operacional e definições de negócios, facilitando a interoperabilidade semântica e o gerenciamento eficiente de informações no setor financeiro (AMARAL *et al.*, 2021). A manutenção da FIBO conta com o envolvimento de mais de 300 empresas e instituições do setor financeiro (EDCM, 2024).

Os autores Chittimalli *et al.* (2020) e Haj *et al.* (2021a) têm como parte de seus objetivos extrair o dicionário do texto, enquanto os demais dependem de outras fontes para obter as definições dos termos ou não descrevem como são obtidos. A qualidade do dicionário tem influência direta na qualidade da transformação. O uso de uma ontologia desenvolvida e reconhecida para o domínio melhora a qualidade da tradução (FORD *et al.*, 2016).

Embora as abordagens variem, todas apresentam características semelhantes, como a transformação baseada em regras. A prevalência do uso de NLP é para decompor os elementos fundamentais do texto e fornecê-los com parâmetros

para os algoritmos utilizados na transformação. Os autores relatam resultados promissores, mas a rigidez das abordagens baseadas em regras sugere que esses resultados podem não ser replicáveis em outros conjuntos de dados devido à limitada capacidade de generalização.

Outra característica dos trabalhos é o tratamento do fragmento de texto extraído para transformação. Eles são frequentemente tratados como fragmentos independentes, o que não representa textos legais (AGNOLONI; FRANCESCONI, 2011). As exceções são: Joshi e Saha (2020), que desenvolveu uma ontologia para representar a CFR e os metadados de modalidade e similaridade; e Dragoni *et al.* (2018), com a sua ontologia estrutural leve para um documento genérico, mas sem a intenção de atender as características dos textos legais.

Cabe destacar que nenhum dos autores disponibilizou seus códigos, conjuntos de dados ou outros artefatos que possam ser utilizados para reproduzir suas pesquisas.

4.3 Preocupações éticas e legais

Sansone e Sperlí (2022) identificam que a aplicação de inteligência artificial no domínio jurídico deve respeitar princípios como transparência, não discriminação e privacidade desde a concepção. Como desafios destacam a adesão a princípios éticos, providos por entidades como a Comissão Europeia, e a complexidade de analisar casos jurídicos anteriores, especialmente ao monitorar mudanças nas leis e regulamentos. Essas questões são mais relevantes quando aplicadas a tarefas de identificação de frases mais relevantes em documentos jurídicos, nos quais o conceito de relevância é derivado da interpretação jurídica; na criação de perfis ou classificação para fornecer dados probabilísticos sobre reincidência criminal; na avaliação de pequenas reivindicações, como está sendo estudado em lugares como a Estônia; e na previsão de atividades criminosas em áreas específicas de uma cidade.

Para Louis *et al.* (2023), uma das principais considerações éticas é a precisão e justiça (*fairness*) do sistema. Em contextos legais, qualquer forma de viés ou

imprecisão na recuperação e interpretação de textos legais pode levar a resultados injustos. É essencial que o sistema seja projetado e treinado de maneira a minimizar vieses, sejam eles inerentes aos dados de treinamento ou introduzidos pelos próprios algoritmos. Isso envolve a seleção e curadoria cuidadosa de conjuntos de dados de treinamento, bem como o desenvolvimento de algoritmos que sejam transparentes e explicáveis. Outra preocupação ética é a privacidade e confidencialidade dos dados legais. Documentos legais frequentemente contêm informações sensíveis. Garantir que essas informações sejam manuseadas de forma segura e em conformidade com as leis de proteção de dados é de suma importância. Isso inclui a implementação de medidas de segurança robustas para prevenir o acesso não autorizado e garantir que o sistema adira aos padrões legais em relação à privacidade de dados. O potencial para o uso indevido da tecnologia também é uma questão ética significativa. Há um risco de que esses sistemas possam ser usados para fins antiéticos ou ilegais, como vigilância ou discriminação. É importante estabelecer diretrizes e regulamentações claras que governem o uso da tecnologia para prevenir tal uso indevido.

Além disso, os autores consideram a questão da responsabilidade. Quando sistemas são usados na tomada de decisões legais, deve ficar claro quem é responsável pelas decisões tomadas com base na saída do sistema. O papel da tecnologia deve ser de auxiliar, não substituir, o julgamento humano, e deve haver mecanismos para revisar e contestar as recomendações do sistema. Por fim, deve-se considerar o impacto da tecnologia na profissão jurídica e no acesso à justiça. Embora essas tecnologias possam tornar a pesquisa jurídica mais eficiente, há uma preocupação de que elas também possam levar à redução da necessidade de profissionais jurídicos ou criar barreiras para aqueles que não têm acesso a tal tecnologia.

Min et al. (2023) discutem questões éticas relacionadas aos LLMs no contexto de suas limitações e desafios. Eles destacam preocupações relacionadas ao viés e à justiça (*fairness*), que são considerações éticas importantes no desenvolvimento e implantação desses modelos. O viés nos LLMs pode se manifestar de várias formas, como viés de gênero, racial ou cultural, que muitas

vezes são reflexos dos vieses presentes nos dados de treinamento. Esses vieses podem levar a resultados injustos ou discriminatórios quando os modelos são usados em aplicações do mundo real. Por exemplo, um modelo de linguagem que aprendeu associações tendenciosas de gênero pode gerar texto que reforça estereótipos ou discrimina contra certos grupos.

Os autores sugerem que abordar preocupações éticas, incluindo viés e justiça (*fairness*), é uma direção importante para pesquisas futuras no campo do NLP. Eles enfatizam a necessidade de desenvolver métodos para identificar, mitigar e avaliar vieses nos LLMs para garantir que esses modelos sejam usados de maneira responsável e ética.

Longo *et al.* (2024) discutem a Inteligência Artificial Explicável (XAI), entendida como portadora de métodos e técnicas utilizados para tornar as decisões e o funcionamento dos sistemas de IA transparentes e compreensíveis para os seres humanos. O objetivo do XAI é fornecer explicações claras e compreensíveis para o comportamento e as saídas dos modelos de IA permitindo que os usuários confiem e interajam efetivamente com as tecnologias de IA. Isso é particularmente importante em aplicações nas quais as decisões de IA têm consequências significativas, como saúde, finanças e sistemas autônomos. Ao tornar os sistemas de IA mais explicáveis, o XAI visa melhorar sua responsabilidade, justiça (*fairness*) e uso ético. Os autores identificam vários desafios em XAI, entre eles, a criação de explicações para novos tipos de IA, como modelos gerativos, por exemplo os LLMs, e sistemas de aprendizagem distribuída; a melhoria dos métodos atuais de XAI, como métodos de atribuição e algoritmos de aprendizagem baseados em conceitos; o esclarecimento do uso de conceitos em XAI, como a relação entre XAI e confiabilidade; a avaliação de métodos e explicações de XAI; e o apoio à centralidade humana das explicações, tornando-as compreensíveis para os humanos e personalizando-as para os usuários.

No setor financeiro, os autores destacam que XAI é importante para garantir a conformidade regulatória, o gerenciamento eficaz de riscos e a construção de confiança em decisões baseadas em IA. A XAI proporciona transparência em decisões de crédito, ajudando a abordar vieses e justificar resultados. Também

aprimora a detecção de fraudes, esclarecendo a lógica por trás das transações sinalizadas. Para estratégias de investimento, a XAI favorece intuições sobre os fatores que influenciam decisões baseadas em IA melhorando o processo de tomada de decisão para gestores de carteira. Por fim, a XAI é essencial para manter a confiança do cliente, aderir a padrões regulatórios e gerenciar riscos de forma eficaz na indústria financeira.

A presente pesquisa não trata diretamente da questão ética e legal, mas vai de encontro a ela indiretamente ao buscar a precisão na transformação da regra em NL para sua contrapartida em SE preservando o seu significado e a rastreabilidade. Essas duas qualidades não garantem o uso ético ou legal, mas promovem a transparência, essencial para qualquer avaliação ética ou legal.

4.4 Resumo

Os padrões de comparação entre diferentes abordagens de transformação foram elaborados para auxiliar a comparação dos métodos estudados e fornecer o conjunto mínimo de requisitos para esta pesquisa. Foram exploradas várias abordagens de transformação que utilizam NLP para converter textos regulatórios escritos em NL em modelos formalizados, frequentemente incluindo tarefas como: análise e interpretação de NL, extração de elementos relevantes como termos e regras, e suas subsequentes transformações em um vocabulário controlado e regras de negócio estruturadas.

A precisão e clareza na transformação dos textos são necessárias para garantir que erros de interpretação não levem a não conformidades que, por sua vez, resultariam em riscos legais e financeiros para as organizações. As abordagens discutidas e as técnicas de NLP têm grande potencial de resolver os desafios de GRC, proporcionando uma base sólida para a automação.

Não há necessidade de reinventar a roda, porém há sempre espaço para melhorias. Assim, as abordagens estudadas foram discutidas e suas contribuições, desafios e limitações foram destacadas, apontando a necessidade de melhorias contínuas nas tecnologias e métodos. A nossa proposta para esta pesquisa visa cobrir algumas dessas lacunas e atualizar algumas das

tecnologias usadas a fim de melhorar tanto a precisão como a eficiência, principalmente no que diz respeito à forma como o texto é representado, adotando o uso de KG para melhor representar as conexões entre as partes dos textos e seus componentes. Ao reconhecer a importância da sua interpretação, técnicas de desambiguação são empregadas para garantir a correta interpretação e alinhá-la com as definições do mercado. Quanto ao quesito tecnologia, o uso de LLMs aprimoraram as técnicas de NLP, melhoraram a precisão e a capacidade de generalização de toda a abordagem proposta.

5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR

Este capítulo traz uma visão geral da solução na seção “5.1 Introdução”, em seguida, são apresentados os principais artefatos da solução na seção “5.2 Principais artefatos”, na seção “5.3 Cenários de validação” descreve os cenários que serão usados para validação da solução no capítulo “7 VALIDAÇÃO” e a seção “5.4 Resumo” faz um apanhado geral da solução.

5.1 Introdução

A extração de conhecimento de documentos legais é uma área de pesquisa ativa, mas ainda limitada quanto à extração automática de regras e políticas de documentos regulatórios. As técnicas tradicionais de NLP e de recuperação de informações (LOUIS et al., 2023; ANH et al., 2023; SUN et al., 2022), tais como o modelo de saco de palavras (*bag of words*) ou modelo vetorial, são insuficientes para automatizar o processo de análise. Essas técnicas não conseguem capturar as relações semânticas entre os diversos elementos legais presentes na estrutura hierárquica dos documentos regulatórios ou legais. É necessário combinar diversas técnicas para abordar essa questão (JOSHI; SAHA, 2020).

O problema pode ser descrito como uma transformação de texto para um modelo CNL no padrão SBVR-CE. Wimmer e Burguenö (2013) abordam as transformações do tipo *model-to-text* (M2T) e *text-to-model* (T2M), reformulando o teste dessas transformações como um problema equivalente a M2M por meio de um metamodelo genérico para textos. Esse metamodelo organiza o conteúdo textual em elementos de modelo estruturados, o que permite que as transformações M2T e T2M - que geralmente trabalham com texto não estruturado - possam ser testadas de maneira semelhante às transformações M2M.

Os autores propõem validar a transformação de T2M utilizando um arcabouço baseado em contratos, especificamente o método “Tracts”, que originalmente foi desenvolvido para transformações de M2M. A ideia central é transformar os artefatos textuais em modelos estruturados, utilizando um metamodelo genérico

que organiza o texto em elementos de modelo, como “Repository”, “Folder”, “File” e “Line”. Assim, a transformação T2M pode ser validada de forma semelhante às transformações M2M, em que tanto o texto de entrada quanto o modelo de saída são representados de maneira estruturada.

Ao aplicar T2M à linguagem natural, essa representação estruturada facilita a conversão de textos não estruturados ou semiestruturados, como documentos legais, em modelos que destacam entidades, relacionamentos e conceitos-chave. Essa abordagem é especialmente benéfica para áreas como conformidade legal e financeira, em que regulamentações precisam ser analisadas sistematicamente e vinculadas a requisitos operacionais. Ao tratar o conteúdo textual como dados estruturados, esse método permite que ferramentas de MDE suportem tarefas como verificação de conformidade, análise de risco e gestão documental em domínios que tradicionalmente dependem de informações textuais complexas e não estruturadas (WIMMER; BURGUENÖ, 2013).

A abordagem proposta por Wimmer e Burguenö (2013) apresenta uma limitação: o metamodelo utilizado é bastante simplificado, sendo projetado principalmente para lidar com código-fonte e textos semi-estruturados em *UML-based Specification Environment* (USE). Consequentemente, ele pode não capturar completamente a complexa semântica dos textos legais em NL, como o CFR. Esta pesquisa propõe uma transformação de textos em NL para um modelo semiestruturado no padrão SBVR-SE. Ainda assim, a proposta dos autores para validar transformações abre uma nova linha de pesquisa que pode ser explorada em trabalhos futuros.

O protocolo de interpretação de regras baseado em SBVR adotado nesta pesquisa, conforme descrito por Abi-Lahoud et al. (2013), propõe uma abordagem estruturada para interpretação de textos regulatórios servindo de base para determinar as principais tarefas a serem executadas. O protocolo consiste em:

1. **Leitura do texto:** Inicialmente, o texto regulatório é lido cuidadosamente para entender seu conteúdo e contexto.

2. **Identificação de modalidades:** Modalidades-chave, como obrigações, proibições etc., são identificadas no texto. Essas modalidades são necessárias para capturar a natureza deôntica das regras legais.

3. **Processamento de cada modalidade:**

- **Adição de palavras-chave de modalidade:** Palavras-chave de modalidade relevantes são adicionadas ao texto.
- **Identificação do verbo em inglês:** O verbo em inglês no qual a modalidade é aplicada é identificado.
- **Estilização do verbo:** Este verbo é então estilizado usando o estilo de verbo SBVR.
- **Identificação de conceitos de substantivo/frases verbais:** Conceitos de substantivo ou frases verbais que desempenham papéis no verbo são identificados.
- **Completando o conceito de verbo modificado SBVR:** Se os papéis do verbo são desempenhados por conceitos de substantivo, o conceito de verbo modificado SBVR é completado estilizando os conceitos de substantivo identificados. Esses conceitos de substantivo estilizados são adicionados à lista de conceitos de substantivo.
- **Estilização de frases verbais:** Se os papéis do verbo são desempenhados por frases verbais, cada frase verbal é estilizada identificando verbos em inglês, conceitos de substantivo SBVR e palavras-chave. Cada frase verbal é então adicionada à lista de conceitos de verbo de suporte.

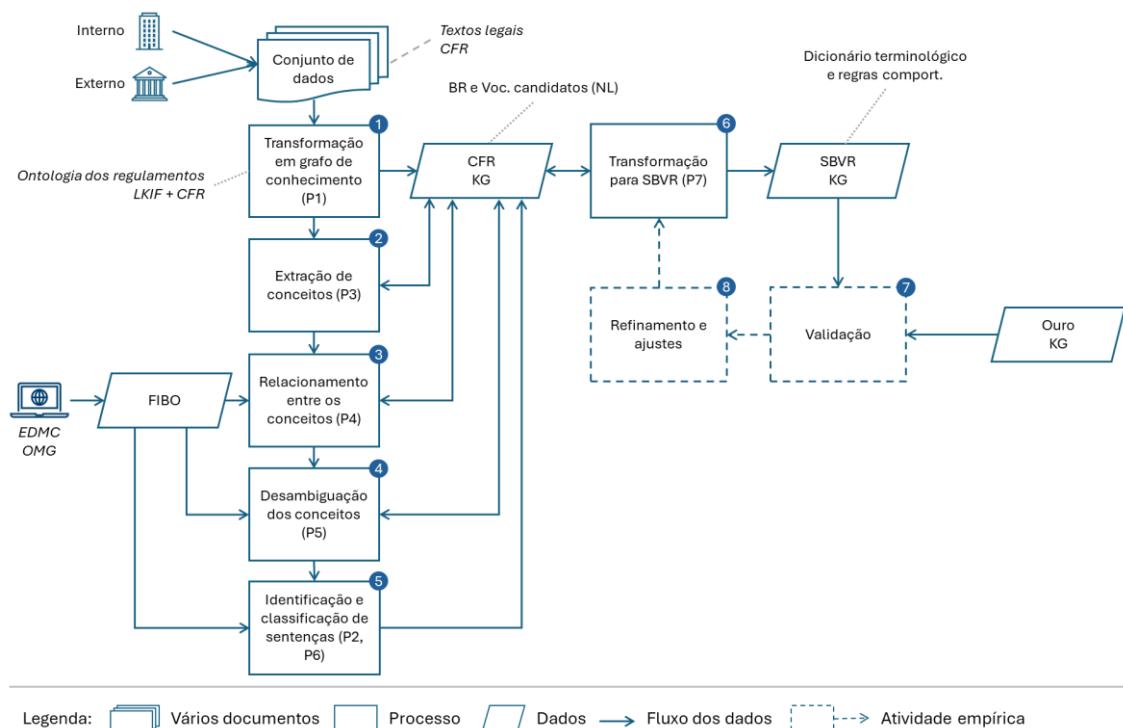
4. **Processamento de cada conceito de substantivo na lista de conceitos de substantivo:**

- **Enriquecimento de conceitos de substantivo:** A característica de cada conceito de substantivo é identificada e enriquecida relacionando com sinônimos, generalização e a ontologia.

- Identificação de outros elementos de definição:** Outros elementos de definição são identificados para esses conceitos de substantivo.

Esse protocolo foi usado para criar o conjunto de dados ouro (WYNER e PETERS, 2011), uma vez que foi projetado para uso manual.

Figura 7 - Arcabouço para o processo de transformação de NL para SBVR.



Fonte: Elaborado pelo autor.

Para a versão automatizada, foram adicionadas etapas inspiradas no trabalho de Haj *et al.* (2021a), por tratar das nuances e complexidades dos textos legais, incluindo a anotação semântica, o uso de KG para capturar conexões textuais e os padrões para identificação e mapeamento das declarações para SBVR. O arcabouço da presente pesquisa para o processo de transformação de NL para SBVR é apresentado na Figura 7.

Em uma visão geral, o passo 1 do arcabouço tem como objetivo a leitura do texto, apresentado na seção “5.2.1 Conjunto de dados”, e a captura da estrutura do documento e suas interconexões, transformando o texto em nós do KG, criando o artefato da seção “5.2.2 KG para CFR e FIBO”. Essa etapa também

leva em consideração as referências cruzadas frequentes em textos legais, que demandam um entendimento dos documentos interligados.

O processo de anotação semântica, passos 2, 3 e 4 é discutido na seção “5.2.3 Anotações semânticas”, cujo principal objetivo é extrair e relacionar os elementos e o passo 5 classificá-los. Todos esses passos têm como objetivo enriquecer os metadados das regras, conforme apresentado na seção “5.2.4 Regras candidatas”.

O sexto passo abrange a transformação do vocabulário e das regras para o formato SBVR, processo no qual os termos e as regras candidatas são transformados para o vocabulário e as regras em SE, conforme a especificação SBVR e discutido na seção “5.2.5 Transformação para SBVR”. Esse processo resolve desafios como a complexidade de frases longas, múltiplas cláusulas e ambiguidades de termos (WYNER; PETERS, 2011) e cria o artefato final “5.2.7 Catálogo de regras”. Os demais passos referem-se à validação da transformação (7) e os seus refinamentos e ajustes (8), detalhados na seção “5.3 Cenários de validação” e no capítulo “7 VALIDAÇÃO”.

A apresentação sequencial do arcabouço é apenas para fins ilustrativos e seu desenvolvimento é apresentado no capítulo “6 FERRAMENTAS DE SUPORTE”. A escolha das abordagens de Abi-Lahoud *et al.* (2013) e Haj *et al.* (2021a) se deu pela completude das abordagens, que atendem a todos os padrões, discutidos na seção “4.1 Padrões para comparação das abordagens de transformação”. Embora as limitações das abordagens discutidas tenham sido endereçadas, entender textos legais frequentemente requer conhecimento contextual mais amplo, não apenas da própria lei, mas também do sistema legal e dos princípios da lei (NGUYEN *et al.*, 2022). Dito isso, a abordagem proposta não está isenta de falhas.

O objetivo desta pesquisa é a automação do processo; no entanto, reconhece-se a importância da supervisão humana (HUANG *et al.*, 2024), conforme apontado também por Louis *et al.* (2023), Min *et al.* (2023) e Longo *et al.* (2024). Para incorporar a validação por especialistas no assunto (SMEs) e tornar o processo semiautomático, um passo adicional poderia ser inserido antes da

conversão para SBVR (etapa 7). Nesse momento, as regras ainda preservam as características do texto original e podem ser facilmente rastreadas até sua fonte, o que facilita a validação pelos SMEs.

O CFR2SBVR contribui para resolução de desafios, como os descritos por Roychoudhury (2017), incluindo a dificuldade para especialistas do domínio extrair regras de grandes volumes de documentos, a complexidade na mineração das regras sem software de apoio e a relutância dos especialistas em adotar linguagens formais. Além disso, a transformação desses textos em modelos regulatórios conceituais é complexa, exigindo ferramentas adequadas e enfrentando problemas de adoção devido à necessidade de edição de modelos como o SBVR. Com essa abordagem, os SMEs podem interagir com as regras e vocabulários no seu estado mais próximo do original (regras candidatas) e validá-las. Esse conjunto de dados pode ser utilizadas para criar treinamentos para outros modelos de ML (MIN *et al.*, 2023). Na próxima seção são detalhados os principais artefatos do arcabouço.

5.2 Principais artefatos

Nesta seção são apresentados os principais artefatos da solução CFR2SBVR.

5.2.1 Conjunto de dados

A publicação de materiais jurídicos online por governos e administrações evoluiu de apenas fornecer acesso para oferecer esses materiais em formatos legíveis por máquina, que aderem aos princípios de dados abertos vinculados³³. Esses princípios incluem a nomeação de recursos com identificador uniforme de recursos (URI), afirmações claras sobre as relações entre recursos ou entre recursos e valores de dados, e a flexibilidade para modificar facilmente essas relações e recursos. A fusão da publicação de documentos jurídicos, como legislação, regulamentações, casos judiciais e decisões administrativas, juntamente com o uso de ontologias OWL/RDF e *thesaurus* SKOS, abre novas

³³ <http://www.w3.org/DesignIssues/LinkedData.html>

oportunidades para aprimorar tanto a acessibilidade quanto a compreensão das informações jurídicas (CASELLAS, 2012; CASELLAS *et al.*, 2011).

Nesta pesquisa, o conjunto de dados são três seções dos regulamentos do título 17, capítulo II, parte 275 da CFR que se refere ao *Rules and Regulations, Investment Advisers act of 1940*. Esses dados serão transformados para SBVR pela solução do arcabouço proposta.

Esses regulamentos estão em formato PDF e XML. O formato XML está documentado no e-CFR³⁴ XML User Guide³⁵ explicando a sua estrutura hierárquica: capítulos têm subcapítulos, cada subcapítulo tem partes, cada parte tem uma subparte, cada subparte tem uma seção e cada seção tem uma subseção. Os documentos podem ser obtidos no site do e-CFR ou por API, sendo duas APIs particularmente úteis, a `structure/{date}/title-{title}.json` que retorna a estrutura de um título (*title*, *subtitle*, *chapter*, *subchapter*, *part*, *subpart*, *section* e *appendix*) e a `versioner/v1/full/{date}/title-{title}.xml` que retorna todo o texto em formato XML. Ambas as APIs estão disponíveis para experimentação no site para desenvolvedores do e-CFR³⁶ e a parte 275 em RDF-Turtle³⁷ disponíveis no site da FRO³⁸.

5.2.2 KG para CFR e FIBO

KG, descrito conceitualmente na seção “3.9 Grafos de conhecimento”, é utilizado de modo bastante amplo. No escopo desta pesquisa, KG é “um grafo de dados destinado a acumular e transmitir conhecimento do mundo real, cujos nós representam entidades de interesse e cujas arestas representam potencialmente diferentes relações entre essas entidades” (HOGAN *et al.*, 2021). De acordo com os autores, o conhecimento refere-se ao “conhecimento explícito”, que é definido como “algo que é conhecido e pode ser escrito”. O conhecimento pode consistir em declarações simples (por exemplo, “Brasília é a capital do Brasil”) ou

³⁴ <https://www.ecfr.gov/>. Acesso em: 15/mar./2025.

³⁵ <https://github.com/usgpo/bulk-data/blob/master/ECFR-XML-User-Guide.md>. Acesso em: 15/mar./2025.

³⁶ <https://www.ecfr.gov/developers/documentation/api/v1>. Acesso em: 15/mar./2025.

³⁷ <https://www.w3.org/TR/turtle/>. Acesso em: 15/mar./2025.

³⁸ <https://finregont.com/fro/cfr/CFR-2012-title17-vol3-part275.ttl>. Acesso em: 15/mar./2025.

declarações quantificadas (por exemplo, “todas as capitais são cidades”). Declarações simples podem ser representadas como arestas no grafo de dados, enquanto representações mais expressivas, como ontologias ou regras, são necessárias para declarações quantificadas, a fim de permitir que métodos dedutivos infiram e acumulem mais conhecimento. Nesse contexto, o conhecimento explícito do mundo real são os regulamentos da CFR.

Os autores discutem diferentes tipos de modelos de grafos usados em grafos de conhecimento: os grafos rotulados com arestas direcionadas (*directed graphs*), são a forma mais básica, com os nós representando entidades e as arestas direcionadas e rotuladas representando relações binárias entre essas entidades; os grafos heterogêneos, atribuem tipos a nós e arestas, proporcionando uma representação mais estruturada; os grafos de propriedades, estendem os *directed graphs* permitindo que nós e arestas tenham propriedades, oferecendo maior flexibilidade na modelagem de dados; e os conjuntos de grafos, permitem o gerenciamento de múltiplos grafos, facilitando a organização de dados de várias fontes ou representando diferentes perspectivas dos dados. Esses modelos fornecem a base para a construção e manipulação de grafos de conhecimento em diferentes cenários de aplicação.

O uso de grafos de conhecimento em áreas como processamento de texto legal permite representar relações complexas e informações interconectadas de forma estruturada. A escalabilidade e flexibilidade dos grafos de conhecimento, junto com a capacidade de integrar diversas fontes de dados, os tornam valiosos para gerenciar e interpretar informações legais, em que é fundamental entender a interação de várias leis, casos e princípios. O objetivo é registrar o conhecimento dos regulamentos por meio da captura de termos-chave, regras, resumos de tópicos, relações entre termos variados, terminologias semanticamente parecidas, sentenças deônticas e fatos e regras com referências cruzadas.

Joshi e Saha (2020) discutem que a natureza semiestruturada desses documentos cria desafios significativos no processo de identificação e referência cruzada de seções relevantes para questões específicas. As buscas por palavras-chave, embora disponíveis, tendem a resultar em um grande volume de correspondências, exigindo uma análise manual detalhada para discernir

entre informações pertinentes e não pertinentes. Além disso, a estrutura organizacional e a indexação em nível relativamente elevado dentro das seções regulatórias complicam a tarefa de localizar e comparar disposições relevantes entre diferentes seções e títulos. Por esse motivo, automatizar o processamento do conhecimento embutido em documentos regulatórios, que normalmente incluem definições, jargões, regras e referências cruzadas, acelera significativamente os processos de agências governamentais, empresas e especialistas jurídicos que precisam se referir a esses regulamentos.

Para representar os regulamentos na CFR, optou-se pelo grafo de propriedades, que será preenchido usando a *Financial Regulation Ontology* (FRO), um arcabouço de código livre criado usando a linguagem OWL. FRO é um componente da Compliance Semântica®, uma marca registrada da Jayzed Data Models Inc., projetada para instituições financeiras. Esse arcabouço é baseado em dois padrões da indústria: a FIBO e o LKIF. FIBO abrange vários aspectos financeiros como fundos, clientes, títulos e derivativos, enquanto LKIF aborda questões legais, incluindo regras e formulários da SEC. O FRO é importante para a conformidade em vários subsetores financeiros, incluindo seguros, bancos e fundos. Notavelmente, o FRO engloba o texto completo das leis e regulamentos dos EUA pertinentes ao setor bancário e de gestão de investimentos.

Jayzed Data Models inc. (2021) detalha o processo de importação do texto das leis e regulamentos dos EUA para o FRO. Isso inclui fontes como o CFR e o *United States Code* (USC)³⁹. O processo envolve vários passos:

- 1. Definição e armazenamento de dados e regulamentações financeiras:** Dados financeiros, regulamentos, relatórios e metadados são armazenados de forma uniforme usando OWL. CFR e USC são processadas desde a extração até o carregamento na ontologia. O *Government Publishing Office* (GPO)⁴⁰ fornece versões em XML do CFR,

³⁹ Esta é a compilação e codificação oficial dos estatutos federais gerais e permanentes dos EUA. Ele organiza as leis por assuntos amplos em 50 títulos e é usado por profissionais do direito e legisladores como referência para estatutos federais

⁴⁰ O GPO é uma agência federal do ramo legislativo do governo dos EUA. É responsável por produzir, publicar e distribuir as publicações do governo dos EUA, incluindo leis, regulamentos e outros documentos federais.

e o *Office of the Law Revision Counsel* (OLRC)⁴¹ codifica a lei e a publica em formato XML. O FRO importa os títulos 12 (Banks and Banking) e 17 (Commodity and Securities Exchanges) do CFR, e os Títulos 12 (Banks and Banking) e 15 (Commerce and Trade) do USC.

2. **Utilização de ontologias de referência:** FRO integra duas principais ontologias de referência de domínio - a FIBO e o LKIF. FIBO abrange conceitos da indústria financeira, enquanto LKIF modela regras legais encontradas em legislação e regulamentos. Essas ontologias de referência são preenchidas com o texto completo das leis relevantes dos EUA, como o CFR e o USC. Alinhando e estendendo essas ontologias, FRO facilita consultas e raciocínio em diferentes domínios.
3. **Projeto de classes de ontologia:** O processo envolve o projeto de classes de ontologia e a extração, transformação e carregamento de dados de arquivos XML que contêm textos de leis e regulamentos.
4. **Definição de hierarquias de classes e restrições:** FRO utiliza, a linguagem de consulta RDF (cf. seção “3.9 Grafos de conhecimento”), em particular o *Resource Description Framework Schema* (RDFS) para descrever taxonomias de classes e propriedades. Isso inclui a definição de subclasses, propriedades de objeto e restrições de classe.
5. **Armazenamento de metadados em triplos:** Metadados, mapeamento e informações de linhagem são armazenados dentro da ontologia na forma de triplos. Este formato é necessário para consultar e raciocinar sobre os dados usando SPARQL (cf. seção “3.9 Grafos de conhecimento”).

Embora a ontologia FRO esteja disponível online, apenas uma parcela do CFR está acessível gratuitamente: a parte 275 do título 17 e as partes 252, 217, 225 do título 12. Para os propósitos desta pesquisa, esses fragmentos são representativos. O processo para transformar a CFR em KG pouco contribui para o objetivo da presente pesquisa, podendo ser revisitado em trabalhos futuros. O

⁴¹ Este escritório faz parte da câmara dos representantes dos EUA e é responsável pela preparação e publicação do USC. O OLRC codifica as leis promulgadas pelo congresso, garantindo que sejam organizadas e apresentadas de forma clara e coerente.

FRO está disponível sob a licença GLP-3.0⁴² e o FIBO sob a licença MIT⁴³, ambas licenças de software livre.

5.2.3 Anotações semânticas

Para esta pesquisa, define-se como anotação semântica o processo de adicionar metadados a textos em NL para resolver ambiguidades e melhorar sua compreensão. Isso envolve associar cada elemento do texto com significados e interpretações específicas para garantir que eles sejam claros, inequívocos e semanticamente consistentes (ASHAFAQ e BAJWA, 2021).

A anotação semântica tem como objetivo preencher a lacuna entre a linguagem natural e as representações formais, tornando mais fácil para pessoas e máquinas entenderem e processarem os textos. No contexto desta pesquisa, essa etapa é responsável por garantir que a inferência sobre o texto legal contém todos os dados necessários para reduzir as possibilidades de interpretações. Também auxilia SMEs a entenderem como a solução transformou o texto original. Ela foi proposta por Omrane *et al.* (2011) e adotada como parte da presente solução.

O processo envolve a vinculação de entidades ontológicas aos termos usados para mencioná-las nos textos. Essa técnica é essencial para entender a aplicação de conceitos em documentos de negócios, as regras associadas a eles e como esses conceitos e regras evoluem com as atualizações das políticas. O processo de anotação semântica destaca menções de entidades ontológicas, como conceitos, instâncias e papéis em documentos, relacionando essas entidades aos termos usados nos textos. O foco da anotação semântica neste contexto está no vocabulário usado para expressar regras estruturais.

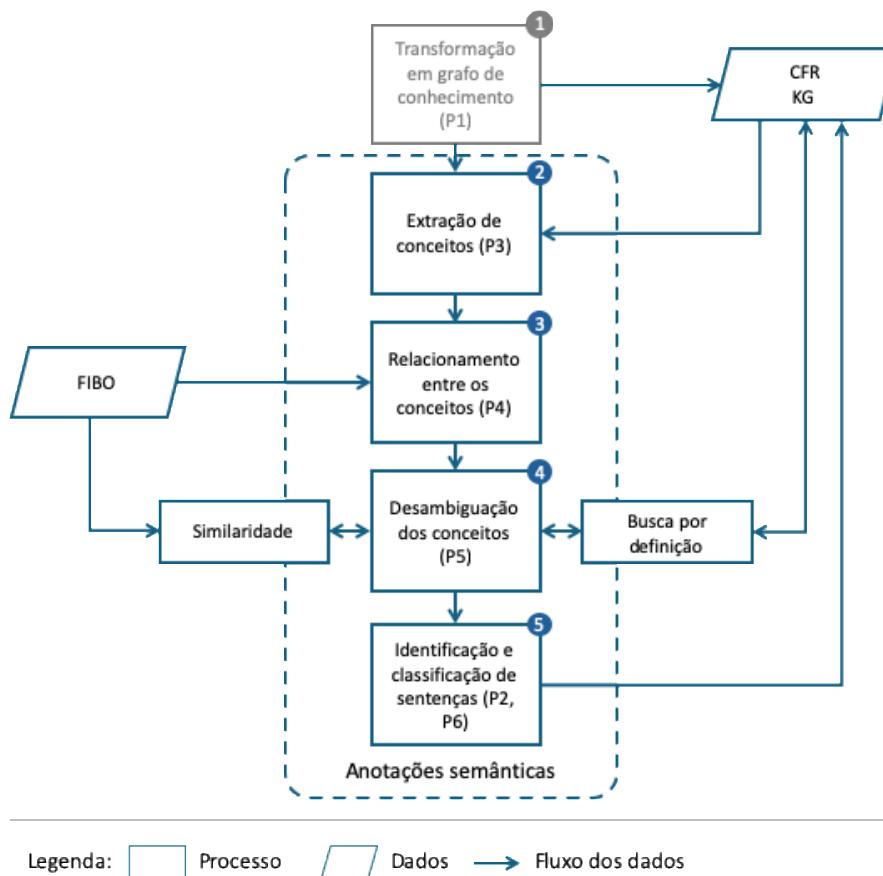
Essa abordagem não necessariamente envolve a codificação de informações complexas, como a estrutura morfológica dos termos. Em vez disso, concentra-se em vincular termos aos seus significados conceituais, criando uma ponte entre a linguagem usada em documentos e os conceitos ontológicos

⁴² <https://www.gnu.org/licenses/gpl-3.0.html.en>. Acesso em: 15/mar./2025.

⁴³ <https://opensource.org/license/mit>. Acesso em: 15/mar./2025.

subjacentes. Isso permite uma compreensão mais detalhada e precisa, além da gestão eficaz de regras de negócios. Um processo semelhante é adotado por Joshi e Saha (2020) em seu arcabouço que utiliza Word2Vec para resolver ambiguidades contextuais por representações vetoriais das palavras, extraindo termos semanticamente semelhantes para enriquecer o KG. Através da resolução de entidade relacionam-se múltiplas instâncias de textos pela referência à mesma entidade da ontologia, mesmo se a entidade for descrita de formas diferentes.

Figura 8 - Anotação semântica



Fonte: Elaborado pelo autor.

Haj *et al.* (2021a) incluem análises lexicais, sintáticas e semânticas para entender a estrutura gramatical e o significado da linguagem para extração do dicionário terminológico e formulações semânticas do texto. Os autores desenvolveram um algoritmo para extração de conceitos de substantivos para identificar e categorizar substantivos e frases substantivas (substantivos

compostos). Inicialmente, ele filtra cada palavra nas declarações para encontrar substantivos, ignorando palavras de outras classes. Substantivos de uma única palavra são identificados e adicionados à lista de conceitos de substantivos. Em seguida, ele examina combinações de palavras para identificar substantivos compostos e adiciona-os à lista como conceitos de substantivos compostos. O resultado é uma coleção de todos os conceitos de substantivos encontrados nas declarações de regras de negócios.

Os autores identificam, com base na extração, padrões nas declarações que são mapeados para declarações SBVR. Esse é o mesmo princípio aplicado nesta pesquisa, mas ao invés de criar nossos padrões, foi utilizado a taxonomia de Witt (2012) que são as declarações SBVR e um conjunto de instruções para o LLM extrair, classificar as declarações e transformá-las com base nos *templates* (semelhante aos padrões de Haj *et al.*, 2021a).

Em linhas gerais, a Figura 8 representa as tarefas de anotação semântica do arcabouço (cf. Figura 7) e ilustra a abordagem aqui utilizada que conta com a capacidade de NLP dos LLMs para extração e relacionamento dos conceitos, utilizando engenharia de *prompt* e adicionando o resultado como propriedades, nós e relações no KG do CFR.

A estratégia de extração adotada foi a classificação das declarações em regras operativas e tipos de fatos, seguido da extração de termos substantivos próprios e comuns e símbolos verbais. Essa estratégia é derivada do manifesto de regras de negócio⁴⁴ que diz no artigo 3:

[...] 3.1. As regras constroem-se sobre factos, e os factos sobre conceitos que são expressos por termos.

3.2. Os termos expressam conceitos de negócio; os factos constituem asserções sobre estes conceitos; as regras restringem e suportam estes factos. [...]

Dessa forma, os termos, nomes e símbolos verbais estão ligados as declarações de regras operativas e fatos. As regras operativas e fatos compartilham o mesmo

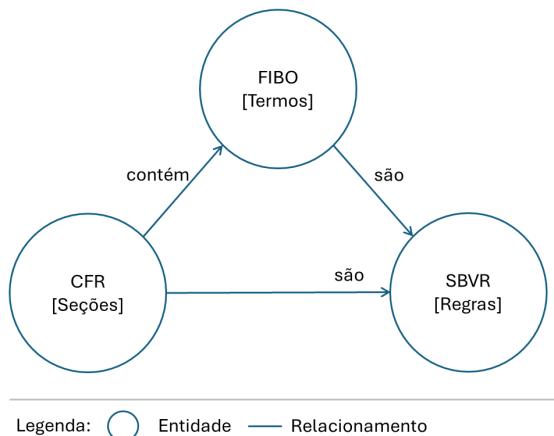
⁴⁴ <https://www.businessrulesgroup.org/brmanifesto.htm>. Acesso em: 15/mar./2025.

vocabulário. Como melhorias futuras, uma ligação do tipo “fato suporta regra” poderá ser introduzida.

Cada conceito extraído (2) é classificado como pertencente ao domínio em questão ou geral, buscando o conceito no grafo obtido da FIBO e estabelecendo sua relação (3). Os conceitos que não são encontrados na FIBO são aproximados por um algoritmo de representação vetorial densa e uma busca por similaridade é realizada na FIBO (4). Esse processo é semelhante ao usado por Joshi e Saha (2020) para resolver desambiguação. Se o conceito não é encontrado, com auxílio de um LLM e um mecanismo de busca para o grafo, faz-se a busca pela definição, que pode estar presente no KG (4); se encontrada, essa definição fará parte do vocabulário SBVR. Essa técnica é semelhante às tarefas 3 - Recuperação de lei estatutária e 4 - Corpus de dados de resposta a questões jurídicas do COLIEE (NGUYEN, 2022). Por fim, o texto é classificado em termos, fatos e regras e é posteriormente essa classificação é refinada seguindo a taxonomia de Witt (2012) (5), com o auxílio de um LLM.

Os KGs estão relacionados como ilustrado na Figura 9, O grafo CFR contém todas as seções e parágrafos da CFR. Os termos FIBO foram armazenados no grafo FIBO e posteriormente transformados em enunciados SBVR, juntamente com definições pelo processo de anotação semântica, o resultado foi armazenado no grafo SBVR. Os parágrafos do grafo CFR têm termos associados aos termos FIBO e aos termos SBVR, já os enunciados SBVR estão associadas somente aos termos no grafo SBVR. Dessa forma, nenhuma extensão à especificação SBVR foi necessária e ainda é possível encontrar uma declaração do grafo SBVR pelos termos da FIBO.

Figura 9 – KGs conceitual



Fonte: Elaborado pelo autor.

5.2.4 Regras candidatas

Resultado intermediário entre o texto legal e a linguagem controlada, as regras candidatas são o mesmo que os nós no KG do CFR com uma propriedade que indica que são regras. Os parágrafos da CFR podem conter outros textos. Esse dado foi criado pelo processo apresentado na seção “5.2.3 Anotações semânticas” e é usado para determinar se o nó do grafo deve ser transformado ou não para uma regra SBVR. A principal propriedade da regra candidata é que ela preserva a característica do texto original e pode ser rastreada à sua origem. Haj *et al.* (2021a), Ashfaq e Bajwa (2021) e Bajwa *et al* (2017) utilizam estratégia semelhante, transformando o texto para um modelo e, posteriormente, de um modelo para outro modelo.

5.2.5 Transformação para SBVR

O processo de transformação consiste em recuperar as regras candidatas e seus relacionamentos com a estrutura do documento (capítulo, seções etc.), juntamente com a definição dos conceitos associados e submeter a inferência de um modelo LLM para a transformação para SBVR.

Um modelo SBVR é formado por um ou mais vocabulários e pode incluir conjuntos de regras baseadas nos conceitos definidos nos vocabulários. A representação textual de modelos SBVR possui uma aparência semelhante à de uma linguagem natural, enquanto simultaneamente são especificações

fundamentadas no *Meta Object Facility* (MOF)⁴⁵ (OMG, 2019), tornando-as interpretáveis por computadores.

Abi-Lahoud *et al.* (2013), Bajwa *et al.* (2017), Chittimalli *et al.* (2020), Haj *et al.* (2021a), Joshi e Saha (2020), Roychoudhury *et al.* (2017) e Skersys *et al.* (2022) lidam com transformações texto-para-modelo ou modelo-para-modelo, tendo como alvo o modelo SBVR. Nas referências bibliográficas estudadas, nenhum autor usou essas técnicas como aplicadas na presente pesquisa.

5.2.6 Recuperador de regulamentos e termos

Na Figura 7 os processos se relacionam com os KGs pelas consultas e operações de escrita. Para as operações triviais, o mecanismo de consulta do banco de dados orientado a grafos é suficiente para atender às necessidades de localizar regulamentos ou termos. Porém, como denotado na seção “5.2.2 KG para CFR e FIBO” e ilustrado na Figura 8, quando a busca exata não localiza o termo, uma busca por similaridade é realizada. Se essa busca ainda não encontrar o conceito, outra busca por similaridade (busca por definição) é realizada, dessa vez, buscando uma definição nos regulamentos.

Para termos que não possuem uma correspondência direta, a técnica de busca por similaridade emprega o mecanismo de RAG (cf. 3.9 Grafos de conhecimento). O conteúdo do KG é transformado em vetores numéricos (*embeddings*), que são então armazenados nas propriedades do próprio KG. Durante uma consulta, a pergunta é também convertida em vetor e sua similaridade é buscada dentro deste espaço vetorial. Os nós que são recuperados são usados pelo LLM para estabelecer a correspondência.

5.2.7 Catálogo de regras

O catálogo de regra (requisito R5 - cf. seção “4.1 Padrões para comparação das abordagens de transformação”) é o repositório onde as regras transformadas serão armazenadas. Para esta pesquisa o próprio KG é o repositório. O repositório contém o livro de regras SBVR, um modelo de conteúdo que inclui

⁴⁵ <https://www.omg.org/mof/>. Acesso em: 15/mar./2025.

tanto o dicionário terminológico quanto orientações comportamentais, contendo regras de negócios.

Segundo a especificação SBVR (OMG, 2019), as formas de empacotar conteúdo para publicação incluem o dicionário terminológico e/ou o livro de regras. O dicionário terminológico contém uma coleção de representações, como designações ou definições de conceitos, além de outras especificações relacionadas a esses conceitos. O livro de regras completo inclui orientação comportamental e regras de negócios, combinando esses elementos com o dicionário terminológico. Resultante desta pesquisa será um livro de regras completo com uma representação informal. A representação formal é aquela em que todas as palavras estão anotadas de acordo com uma notação que pode ser mapeada para o SBVR, ou seja, cada elemento da linguagem é rigorosamente definido e segue uma estrutura formalizada. Já a representação informal é aquela em que nem todas as palavras estão anotadas, podendo incluir termos que não seguem uma notação formal completa. A representação informal pode conter elementos menos estruturados, mas ainda claros para a comunidade discursiva. Essas duas formas de representação são mutuamente exclusivas.

O processo de transformação CFR2SBVR, visa identificar e transformar o maior número de elementos, mas somente uma revisão por SMEs pode garantir sua completude, porém, é assumido que os elementos na CFR são entendidos pela comunidade discursiva, no entanto, não é possível deduzir que os elementos são implicitamente conhecidos e não precisariam de uma representação formal no livro de regras SBVR, segundo a especificação SBVR (OMG, 2019).

Os elementos SBVR no KG serão triplas obedecendo a ontologia SBVR adaptada para esta pesquisa. O predicado “sbvr:referenceSupportsMeaning” é usado para relacionar os elementos SBVR ao parágrafo. Com essa associação, o livro de regras SBVR atende aos requisitos de rastreabilidade (requisito R6) e permite que os SMEs auditem o processo de transformação.

5.2.8 Ontologia SBVR

A especificação SBVR (OMG, 2019) não prescreve uma ontologia, de fato e por curiosidade, a especificação versão 1.5 da SBVR menciona a palavra ontologia(s) seis vezes, quatro vezes nas referências, uma vez nas marcas registradas sobre o FIBO e uma vez na definição de quantidade (OMG, 2019 p.93), porém, dedica o capítulo 23 (OMG, 2019 p.195-218) sobre como o SBVR utiliza o MOF e XMI para representar semanticamente o conteúdo do SBVR em modelos legíveis por máquina. O capítulo aborda a transformação das entradas terminológicas do SBVR (Cláusulas 7 a 21) para o metamodelo SBVR MOF, que é serializado no metamodelo SBVR XMI. Isso inclui como representar conceitos substantivos e verbais, suas designações e formulações, além de especializações, generalizações, tipos de conceitos, definições e formulações semânticas formais. Também inclui declarações de necessidade, vocabulários, linguagens, namespaces e URIs, bem como notas, exemplos, fontes e descrições.

No entanto, as cláusulas 1.4 e 1.5 da especificação SBVR (OMG, 2019) deixam claro que o conteúdo do SBVR, incluindo o dicionário terminológico e o livro de regras, não é um metamodelo diretamente utilizável como modelo de dados, modelo de mensagens ou modelo de informações de negócios. Ele é uma documentação de significados, como conceitos e proposições, e expressa essas ideias em linguagem natural, similar a um dicionário de negócios ou manual de políticas. Para usar esses conteúdos em sistemas de TI, é necessária uma transformação para que o modelo de conteúdo SBVR possa ser aplicado a modelos de dados, informações ou raciocínio sobre negócios.

Esta pesquisa não tem como objetivo se aprofundar na definição de uma ontologia para SBVR, ficando para trabalhos futuros, porém, de forma pragmática, foi necessário adaptar o metamodelo SBVR XMI para uma ontologia que permita usá-la no modelo de dados dos KGs. A adaptação foi limitada aos elementos do modelo MOF usados na transformação são: “sbvr:Term”, “sbvr:Name”, “sbvr:Fact” (tipo “sbvr:SententialForm”) e “sbvr:Rule”.

Essa adaptação foi inspirada por Hinkelmann *et al.* (2018), que discute que a função tradicional de um metamodelo é fornecer a estrutura e as regras que guiam a construção de modelos, definindo quais elementos e relações podem ser usados. No caso de uma ontologia sendo usada como metamodelo, ela vai além de simplesmente definir a estrutura: ela também formaliza a semântica dos conceitos do domínio, permitindo uma interpretação precisa e não ambígua, tanto para humanos quanto para máquinas. Ao representar a semântica de uma linguagem de modelagem em uma ontologia, os autores garantem que os modelos criados sejam consistentes e compreensíveis por sistemas automatizados, ao mesmo tempo que oferecem uma camada de representação gráfica que facilita a compreensão por parte dos humanos.

O SBVR fornece o metamodelo na forma XSD. A ontologia foi adaptada convertendo o esquema XML para o formato RDF, mapeando elementos e tipos do XSD em classes e propriedades RDF. Primeiro o algoritmo define e vincula *namespaces* comumente usados em arquivos XSD, incluindo os vocabulários XML Schema, XMI e SBVR. Em seguida, o algoritmo analisa o arquivo XSD, processa seus elementos “xs:element” e “xs:complexType”, e gera as classes, propriedades e relações de subclasse correspondentes no RDF. Além disso, atributos e elementos filhos (como sequências e escolhas) dentro dos tipos complexos também são traduzidos para RDF. O grafo RDF é então serializado em formato RDF/turtle. Essa conversão permitiu que a estrutura definida no XSD fosse representada em um formato compatível com a web semântica. O algoritmo e a ontologia estão disponíveis na integra no repositório de códigos⁴⁶.

A conversão foi o ponto de partida, foram adicionados comentários “rdfs:comment” para melhorar a documentação e corrigidos problemas como a remoção de declarações autorreferenciais incorretas de “rdfs:subClassOf” e a substituição do prefixo incomum “xs” pelo prefixo padrão “xsd” para os tipos de dados do XML Schema. Todas as instâncias de “xsd:IDREF” foram eliminadas, pois seu uso não é comum em RDF ou OWL. As propriedades “xmi:id” e “xmi:Extension” foram removidas, já que são geralmente usadas para referenciar

⁴⁶ Ontologia completa disponível no repositório de códigos na pasta “data/sbvr-dtc-19-05-32-ontology-v1.ttl”. A transformação do esquema SBVR (XSD) para RDF foi realizada pelo script “src/sbvr_xsd_to_rdf.py”

elementos e adicionar recursos personalizados a modelos de metadados, como XML/UML. Também foram substituídas as referências a “xmi:Any”, uma vez que tipos de dados como “xsd:anyType” ou “xsd:string” são mais apropriados para contextos da web semântica. Nomes de classes foram normalizados para *PascalCase* e as propriedades para *camelCase*. Em alguns casos, existiam nomes que eram classes e propriedades, por exemplo “sbvr:Rulebook”, nesse caso, de acordo com a especificação, “rulebook” é uma classe e a ontologia foi ajustada. Essas alterações visaram aumentar a conformidade com os padrões RDF/OWL.

Quadro 9 - Ontologia SBVR adaptada

```

sbvr:Term a owl:Class ;
rdfs:label "term" ;
rdfs:comment "A verbal designation of a general concept that is in a given subject field" ;
rdfs:subClassOf skos:Concept .

sbvr:exactMatch a rdf:Property ;
rdfs:label "exact match" ;
rdfs:comment "Links an sbvr:term to a skos:Concept that has the same meaning. When used to connect terms in a similarity search, means the term has at least a 99% match",
"This is the exactMatch property or class within the SBVR ontology." ;
rdfs:domain sbvr:Term ;
rdfs:range skos:Concept ;
rdfs:subPropertyOf skos:exactMatch .

sbvr:closeMatch a rdf:Property ;
rdfs:label "close match" ;
rdfs:comment "Links an sbvr:term to a skos:Concept that has the closest meaning. When used to connect terms in a similarity search, means the term has at least a 80% match",
"This is the closeMatch property or class within the SBVR ontology." ;
rdfs:domain sbvr:Term ;
rdfs:range skos:Concept ;
rdfs:subPropertyOf skos:closeMatch .

# Usage
INSERT {
cfr-sbvr: TotalAsset sbvr:Term,
sbvr:NounConcept,
```

```

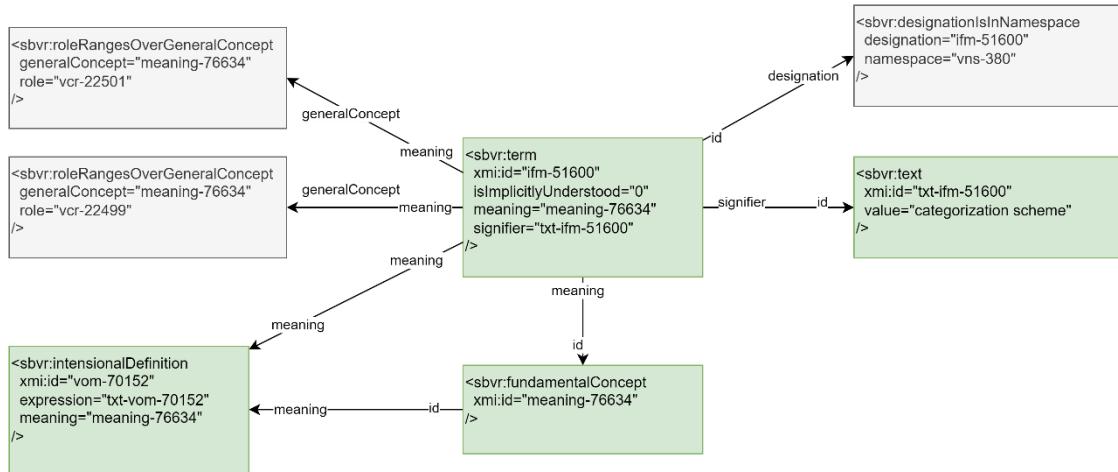
sbvr:IntensionalDefinition ;
sbvr:isImplicitlyUnderstood "false"^^xsd:boolean ;
sbvr:designationIsInNamespace cfr-sbvr:SBVR_CFR ;
sbvr:signifier "total asset" ;
sbvr:Statement " The total assets as shown on the balance sheet of the investment adviser or another relevant person, or the consolidated balance sheet of the adviser with its subsidiaries."
sbvr:closeMatch
<https://spec.edmcouncil.org/fibo/ontology/SEC/Securities/SecuritiesClassification/AssetClass>,
<https://spec.edmcouncil.org/fibo/ontology/FND/Accounting/AccountingEquity/FinancialAsset>,
<https://spec.edmcouncil.org/fibo/ontology/FND/OwnershipAndControl/Ownership/TangibleAsset>,
<https://spec.edmcouncil.org/fibo/ontology/FND/OwnershipAndControl/Ownership/Asset>,
<https://spec.edmcouncil.org/fibo/ontology/FBC/ProductsAndServices/ClientsAndAccounts/AccountAsAnAsset>,
<https://spec.edmcouncil.org/fibo/ontology/FND/Accounting/AccountingEquity/PhysicalAsset>
.
}

```

Fonte: Elaborado pelo autor.

No entanto, o projeto (*design*) do esquema XSD SBVR não foi pensado para ser utilizado como ontologia e a conversão resultou em classes e propriedades, cuja conexão (predicado) é uma classe como “sbvr:RoleRangesOverGeneralConcept”, que poderia ser melhor representada como “sbvr:Role” “sbvr:rangesOver” “sbvr:GeneralConcept”, em que o predicado seria uma propriedade RDFS.

Figura 10 - Associação dos elementos no modelo de conteúdo SBVR



Fonte: Elaborado pelo autor.

Como ilustrado na Figura 10, os predicados foram definidos considerando a associação de identificadores, seria necessário identificá-los, seguindo a especificação SBVR e reclassificá-los como propriedades RDFS e ajustar seus nomes, adicionar restrições e revisar as subclasses, ficando isso para trabalhos futuros. Porém, esse problema de projeto não inviabiliza o seu uso, já que as principais classes estão presentes e a redundância dos predicados é tolerável.

Outra adaptação feita à ontologia é que a especificação SBVR não oferece recurso para conectar um elemento com outro para representar a sua relação aproximada com elementos de outro vocabulário, por exemplo, a ontologia MVF Termos e Definições⁴⁷ têm predicados como “hasNarrowerEntry”, “hasSemanticReference”, “hasBroaderEntry” que auxiliam a determinar o quanto próximo uma definição está de outra. Em SKOS, os predicados “exactMatch” e “closeMatch” capturam essa relação, embora de forma binária. Para a ontologia adaptada, foram incorporados aos predicados “skos:exactMatch” e “skos:closeMatch”, com o uso exemplificado no Quadro 9. O “skos:exactMatch” corresponde a uma pontuação mínima de 99% na busca por similaridade de termos no KG FIBO e KG CFR_SBVR e o “skos:closeMatch” pontuação mínima de 80%. O uso desses predicados cria a associação dos termos SBVR aos

⁴⁷ <https://spec.edmcouncil.org/idmp/ontology?query=https://www.omg.org/spec/MVF/ISO1087-VocabularyForTermsAndDefinitions/>. Acesso em: 15/mar./2025.

termos FIBO. O valor da pontuação mínima de 80% e o limite de 5 resultados foram arbitrários.

Para trabalhos futuros, a definição de objetos para representar os graus de similaridade e predicados para expressá-las poderão auxiliar na compreensão dessa relação.

Quadro 10 - Inserção de um termo no KG

s	p	o
cfr-sbvr:TotalAsset	rdf:type	sbvr:Term
cfr-sbvr:TotalAsset	sbvr:signifier	total asset
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:AssetClass
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:FinancialAsset
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:TangibleAsset
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:Ownership/Asset
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:AccountAsAnAsset
cfr-sbvr:TotalAsset	sbvr:closeMatch	fibo:PhysicalAsset
cfr-sbvr:TotalAsset	sbvr:isImplicitlyUnderstood	"false"^^< http://www.w3.org/2001/XMLSchema#boolean >
cfr-sbvr:TotalAsset	rdf:type	sbvr:NounConcept
cfr-sbvr:TotalAsset	rdf:type	sbvr:IntensionalDefinition
cfr-sbvr:TotalAsset	sbvr: <u>Statement</u>	The total assets as shown on the balance sheet of the investment adviser or another relevant person, or the consolidated balance sheet of the adviser with its subsidiaries.

Fonte: Elaborado pelo autor.

O comando de inserção apresentada no Quadro 9, cria 13 triplas, apresentadas no Quadro 10. As triplas seguem a ontologia adaptada, o termo é do tipo “sbvr:Term” ou “sbvr:Name” dependendo da classificação do tipo de conceito.

5.3 Cenários de validação

De acordo com Wieringa (2014), na DSR as etapas de validação e avaliação são críticas. Nelas, o artefato projetado é examinado para garantir que atenda aos requisitos estabelecidos e aborde efetivamente o problema identificado. Essas etapas envolvem avaliar a interação do artefato com o contexto para observar os efeitos produzidos. Considerações-chave incluem compensações para

diferentes artefatos, sensibilidade em vários contextos e se os efeitos satisfazem os requisitos. Esse processo de validação é essencial para confirmar que o projeto do artefato não é apenas teoricamente sólido, mas também viável praticamente e capaz de alcançar os resultados pretendidos no contexto do problema.

A validação e a avaliação têm propósitos distintos. Na validação é verificado se o artefato cumpre seu propósito pretendido, atende aos requisitos pré-definidos e é adequado ao contexto do problema para o qual foi projetado. Trata-se de garantir a “correção do projeto”. Durante a validação, perguntas como “Estamos construindo a coisa certa?” e “Este artefato faz o que deveria fazer?” são abordadas. A validação frequentemente ocorre antes e durante a fase de implementação para garantir que o projeto do artefato esteja correto e funcionará conforme o esperado.

A avaliação, por outro lado, trata de avaliar a eficácia e o desempenho do artefato após sua implementação. Envolve medir o impacto real do artefato em seu ambiente operacional, determinar se resolve o problema para o qual foi destinado a resolver e quanto bem o faz. Perguntas como “Quão bem-estamos construindo?” e “Quão bem o artefato atua no contexto do mundo real?” são feitas durante a avaliação. A avaliação olha para o resultado e inclui a avaliação do artefato contra critérios como eficiência, eficácia e satisfação do usuário após ter sido implantado no contexto real (WIERINGA, 2014). O escopo desta pesquisa é o ciclo de projeto, que não contempla a etapa de avaliação, somente a validação, que é discutida a seguir.

Cenários de validação foram criados para validar se o artefato funciona conforme o esperado em condições o mais próximas possível de seu ambiente operacional. Os cenários de validação estão estruturados da seguinte maneira:

1. **Projeto do cenário:** O cenário de validação é projetado para simular as condições sob as quais o artefato ou sistema será usado. Isso inclui definir o contexto, os casos de uso esperados e os critérios para operação bem-sucedida.

2. **Seleção de parâmetros:** Parâmetros relevantes que influenciarão o resultado do cenário são identificados. Isso pode incluir condições ambientais, interações do usuário, entradas de dados e outras variáveis que afetam o comportamento do sistema.
3. **Criação de casos de teste:** Com base nos parâmetros, o caso de teste é desenvolvido. Cada caso especifica as entradas ou ações que serão executadas e os resultados esperados. Os resultados esperados têm como referência o conjunto de dados padrão-ouro (de referência). Esse conjunto de dados é definido na seção “3.11 Validação e testes”.
4. **Monitoramento e coleta de dados:** Durante a execução dos cenários de validação, os dados são coletados sobre como o sistema se comporta, incluindo quaisquer desvios do comportamento esperado.
5. **Refinamento iterativo:** Muitas vezes, os cenários de validação levam a um ciclo de refinamento. Se o artefato não funcionar conforme o esperado, o projeto pode ser modificado e o cenário reexecutado até que os resultados desejados sejam alcançados.
6. **Documentação:** Registros detalhados do processo de validação, observações, dados coletados e análises são mantidos. Essa documentação é importante para entender o desempenho do sistema e para fazer quaisquer ajustes necessários.

Para um conjunto de cenários:

1. **Execução:** Os casos de teste são então executados em um ambiente controlado, onde o comportamento do artefato pode ser observado e registrado. Isso pode ser um ambiente físico, uma simulação ou uma combinação de ambos.
2. **Análise:** Os resultados dos casos de teste são analisados para determinar se o artefato atende aos critérios de sucesso definidos. Isso inclui a busca por falhas ou deficiências no projeto.

3. **Resultados:** Por fim, os *stakeholders* podem revisar os resultados dos cenários de validação para confirmar que o sistema atende às suas necessidades e expectativas antes de ser implantado. Nesta pesquisa não há partes interessadas; a revisão será feita pelo autor.

Seguem os cenários de validação do CFR2SBVR.

5.3.1 Extração dos elementos

Cenário de validação para a extração das declarações e elementos das seções da CFR.

- Contexto: O texto da CFR está disponível e adequado para processamento.
- Entradas: Seções 275.0-2, 275.0-5 e 275.0-7
- Casos de teste:
 - Regra candidata: Declarações são extraídas do texto e representam regras operativas ou tipos de fatos;
 - Elementos termos substantivos comuns e próprios e símbolos verbais são identificados;
 - As declarações e elementos são identificados corretamente (maior que 80%) em relação ao conjunto de dados ouro.
- Refinamento: Múltiplas interações de refinamento do *prompt* para extração.
- Monitoramento, coleta de dados e documentação: Registro do resultado dos casos de testes e do cenário estão no repositório de código no notebook “code/src/chap_7_validation_elements_extraction.ipynb”.

5.3.2 Classificação das regras

Cenário de validação para a classificação dos elementos extraídos.

- Contexto: Os elementos extraídos estão disponíveis.
- Entradas: Declarações do tipo regras operativas, tipos de fatos, termos (comuns e próprios) e símbolos verbais.
- Casos de teste:
 - Regras operativas: São classificadas em relação ao topo da taxonomia de Witt (2012);
 - Elementos termos substantivos comuns e próprios e símbolos verbais são identificados;
 - Os elementos são classificados corretamente (maior que 80%) em relação ao conjunto de dados ouro.
- Refinamento: Múltiplas interações de refinamento do *prompt* para extração.
- Monitoramento, coleta de dados e documentação: Registro do resultado dos casos de testes e do cenário estão no repositório de código no notebook “code/src/chap_7_validation_rules_classification.ipynb”.

5.3.3 Transformação para CNL

Cenário de validação para a transformação dos elementos classificados em CNL.

- Contexto: Os elementos classificados estão disponíveis.
- Entradas: Elementos classificados com tipo e subtipo: regras operativas, tipos de fatos, termos e nomes.
- Casos de teste:
 - Elementos: Transformadas de acordo com os *templates* de Witt (2012);

- Os elementos são transformados corretamente (maior que 80%) de acordo com os algoritmos de similaridade semântica.
- Refinamento: Múltiplas interações de refinamento do *prompt* para transformação.

Monitoramento, coleta de dados e documentação: Registro do resultado dos casos de testes e do cenário estão no repositório de código no notebook “code/src/chap_7_validation_rules_transformation.ipynb”.

5.4 Resumo

A solução para a transformação de NL para SBVR, como discutido nas seções anteriores, compreende vários passos, sendo os principais a transformação do texto original em um KG, que preserva as relações existentes no texto legal e permite enriquecê-las com a identificação de conceitos-chave e sua associação com sua definição na FIBO, bem como a identificação de regras comportamentais e de definição, nomeada de anotação semântica. Esses dados enriquecidos podem ser utilizados por SMEs para validar o texto legal antes ou após a sua transformação, tornando o processo transparente, auditável e minimizando os riscos de más interpretações. Por fim, com o auxílio de um modelo de IA avançado em técnicas de NLP e os dados enriquecidos, a transformação do regulamento para SBVR é realizada, usando a engenharia de *prompt* para instruir o modelo e um conjunto de metadados associados ao regulamento para reduzir a probabilidade de alucinações. No capítulo “6 FERRAMENTAS DE SUPORTE” discute-se como a solução de software é desenvolvida.

6 FERRAMENTAS DE SUPORTE

Neste capítulo, são apresentadas as ferramentas e a infraestrutura de software desenvolvidas para dar suporte à solução proposta de transformação de textos em linguagem natural para modelos SBVR. Com base no método de trabalho delineado anteriormente, foi necessário criar um conjunto de ferramentas que implementassem os principais artefatos definidos no arcabouço teórico, além de possibilitar a coleta de dados e a validação dos resultados obtidos.

Na seção “6.1 Ferramentas e infraestrutura”, são descritas as ferramentas utilizadas e construídas para esta pesquisa. Isso inclui a arquitetura geral do sistema discutida na seção “6.1.1 Arquitetura do sistema”, explicando como os diversos componentes interagem para realizar as tarefas propostas. Também é apresentada a infraestrutura estabelecida para suportar a execução dos cenários de validação, garantindo que o ambiente reflita as condições necessárias para a aplicação e teste da solução, conforme descrito na seção “6.1.2 Infraestrutura para validação”.

Em seguida, na seção “6.2 Implementação dos principais componentes”, cada um dos artefatos é discutido em detalhe. A subseção “6.2.1 KG para CFR e FIBO” explica como o grafo foi construído e utilizado para representar os regulamentos e conceitos financeiros. Na “6.2.2 Orquestrador” é explicado o algoritmo que orquestra a execução dos demais algoritmos e a seção “6.2.3 Anotações semânticas”, detalha-se o processo de enriquecimento dos dados por meio da identificação e vinculação de conceitos-chave. Em “6.2.4 Transformação”, é apresentado o processo de conversão das regras candidatas para o modelo SBVR utilizando técnicas de IA. Por fim, na seção “6.2.5 Catálogo de regras”, discute-se o repositório final das regras transformadas e como ele permite a rastreabilidade e auditoria do processo.

6.1 Ferramentas e infraestrutura

As ferramentas são os algoritmos desenvolvidos em Jupyter Notebooks com kernel Python 3.11⁴⁸, em que o LLM ChatGPT 4o versão gpt-4o (OPENAI, 2024) é o LLM responsável pela inferência e a persistência é realizada no sistema de arquivos e no banco de dados de grafo. Os algoritmos serão discutidos nas seções seguintes. Os *notebooks* foram organizados com as seguintes seções:

- **Título:** O título e quais algoritmos o *notebook* implementa;
- **Google Colab:** Código para adaptar a execução em Google Colab;
- **Imports:** Bibliotecas utilizadas pelo *notebook*;
- **Settings:** Obtenção da configuração (URLs, diretórios padrão etc.) e configuração de *logging*;
- **Checkpoints:** Configuração dos *checkpoints*. São arquivos *json* que contém os dados referentes a uma execução. Ele é usado para transferir dados de uma execução para outras. Os *notebooks* devem ser executados em ordem: (1) “chap_6_semantic_annotation_elements_extraction.ipynb”, onde o arquivo de *checkpoint* é definido; seguido do (2) “chap_6_semantic_annotation_rules_classification.ipynb”, que reutiliza o *checkpoint* criado na última execução do *notebook* anterior; o (3) “chap_6_nlp2sbvr_transform.ipynb” também reutiliza o *checkpoint* criado pelo primeiro *notebook*; por fim o (4) “chap_6_nlp2sbvr_elements_association_creation.ipynb” que utiliza o resultado do último *checkpoint* e insere as triplas no KG. Os *notebooks* “chap_7_validation_*.ipynb” utilizam os *checkpoints* como entrada para os cenários de validação.
- **General functions and data structures** (opcional): Contém as funções e definições de estruturas de dados utilizados no *notebook*;
- **Datasets** (opcional): Dados que serão usados no processamento do *notebook*;

⁴⁸ As bibliotecas adicionais utilizadas estão referenciadas no arquivo requirements.txt no repositório de códigos.

- **Prompt engeneering** (opcional): Contém o código e discussão sobre os *prompts* utilizados no *notebook*;
- **Execution**: Contém o código de processamento do *notebook*.

Os *notebooks* de validação (“chap_7_validation_*.ipynb”) abrigam os códigos para a validação dos algoritmos, o *notebook* “chap_6_create_kg.ipynb”, inicializa o KG e o “chap_6_cfr2sbvr_modules.ipynb” abriga os módulos e códigos compartilhados entre os *notebooks*. Para mais informações sobre os *notebooks* consulte o “APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO”.

O LLM é um componente sensível para o experimento, sua escolha tem impacto direto nos resultados. Wang *et al.* (2024) propõe um *benchmark* para avaliar a performance de modelos de linguagem, como o ChatGPT, em cenários reais. A avaliação, baseada em mais de 1.800 casos fornecidos por usuários, mostra que os LLMs são eficazes em tarefas objetivas, mas têm dificuldades em lidar com contextos subjetivos e complexos, como criatividade e aconselhamento pessoal. Nas tarefas objetivas, o ChatGPT se saiu melhor que os demais em questões factuais.

Chang *et al.* (2024) avalia 18 modelos de LLMs em diferentes tarefas e organiza a avaliação em três dimensões: o que avaliar (tarefas como análise de sentimentos e geração de linguagem), onde avaliar (benchmarks como GLUE e MMLU), e como avaliar (métodos automáticos e humanos). O GPT-4 foi o modelo de melhor desempenho geral, superando o ChatGPT-3.5 e o GLM-130B em raciocínio matemático, geração de código, factualidade, tradução multilíngue e classificação de textos.

O ChatGPT 4o apresenta melhorias significativas em relação as versões anteriores no processamento de linguagem natural. Uma das principais inovações é a capacidade de gerar saídas estruturadas, como JSON, diretamente do modelo, simplificando a integração com sistemas que exigem dados em formatos predefinidos, reduzindo a necessidade de pós-processamento. Além disso, o GPT-4o apresenta uma janela de contexto maior, permitindo lidar com até 128.000 tokens, em comparação com os 4.096 tokens do GPT-3.5 turbo, por exemplo. Isso possibilita lidar com textos mais longos e

mais complexos de forma eficiente. O modelo também oferece maior precisão e entendimento multimodal, especialmente em tarefas de compreensão de texto e integração de imagens e dados (OPENAI, 2024).

O banco de dados para os KGs utilizado foi o AllegroGraph versão 8.2.1. Ele foi escolhido por ser uma plataforma de banco de dados orientada a grafos e documentos e incluir suporte aos padrões da W3C, como RDF, RDF-Star e SPARQL e a *embeddings* que permite armazenar vetores gerados a partir de LLMs, como o GPT da OpenAI. Esses vetores são armazenados no banco de dados vetorial junto com o texto original e as relações RDF correspondentes, permitindo consultas por similaridade semântica (FRANZ INC., 2024).

Durante a execução dos processos, foi utilizado uma estrutura de dados do tipo dicionário para armazenar os resultados intermediários e servir de banco de dados para os processos. Foi desenvolvido um módulo simples (*checkpoint*) para manipular essa estrutura de dados. Para trabalhos futuros e para melhorar a escalabilidade da solução, esse componente pode ser substituído por um banco de dados de documentos (no-SQL). Para a interface web de inspeção dos dados, os *checkpoints* foram importados para o DuckDB⁴⁹, um banco de dados SQL OLAP em processo, que dispensa um servidor para gerenciamento. A interface web foi desenvolvida em Python utilizando o framework Streamlit⁵⁰.

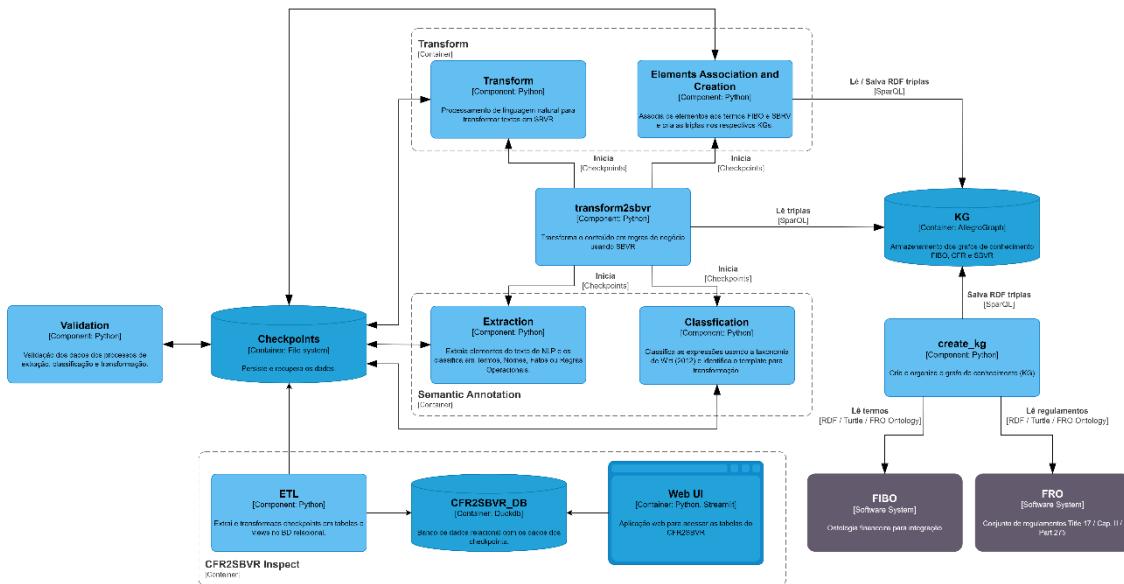
A infraestrutura e ambiente de execução para experimento foram o Google Colab (GOOGLE, 2024) com tipo de ambiente de execução padrão CPU, com aproximadamente 12GB de memória RAM e 100GB de espaço em disco. O CFR2SBVR não tem exigências específicas de hardware, uma vez que tanto o banco de dados como o LLM são executados em ambiente de nuvem do tipo *Software as a Service* (SaaS). O uso de CPU e memória são para a execução dos algoritmos em Python. O experimento também pode ser executado em infraestrutura local com hardware equivalente, o ambiente de execução de notebooks e banco de dados dispõe de versões instaláveis.

⁴⁹ <https://duckdb.org/docs/stable/>. Acesso em: 15/mar./2025.

⁵⁰ <https://docs.streamlit.io/>. Acesso em: 15/mar./2025.

Todos os arquivos utilizados no experimento estão disponíveis no repositório de códigos⁵¹ sob a pasta “code”.

Figura 11 - Arquitetura do CFR2SBVR



Fonte: Elaborado pelo autor.

6.1.1 Arquitetura do sistema

A Figura 11 ilustra a arquitetura do sistema. Há quatro componentes principais, (1) o *create_kg* – responsável por carregar para o banco de dados os dados da FIBO, no formato RDF / Turtle, criando o gráfico de conhecimento FIBO e carregar os dados da FRO, também no formato RDF / Turtle, criando o grafo FRO e os *embeddings* para as triplas do FIBO e os índices necessários para a busca por similaridade; (2) o *transform2sbvr* – responsável por orquestrar os principais componentes, iterando pelas seções do grafo CFR e invocando os componentes para anotação semântica e transformação dos elementos; (3) o “semantic annotation”, responsável pela extração e classificação dos elementos; e (4) o “Transformation”, responsável pela transformação das declarações em NLP para SBVR. A implementação dos componentes é discutida na seção “6.2 Implementação dos principais componentes”.

⁵¹ Repositório do autor: <https://github.com/asantos2000/master-degree-santos-anderson>. Acesso em: 15/mar./2025.

O componente “CFR2SBRV Inspect” é um componente adicional para visualização dos dados dos *checkpoints* relacionados aos diferentes processos (extração, classificação, transformação e validação), esse componente não está envolvido no processo de transformação e seu objetivo é servir de ferramenta adicional de visualização / interação para os SMEs.

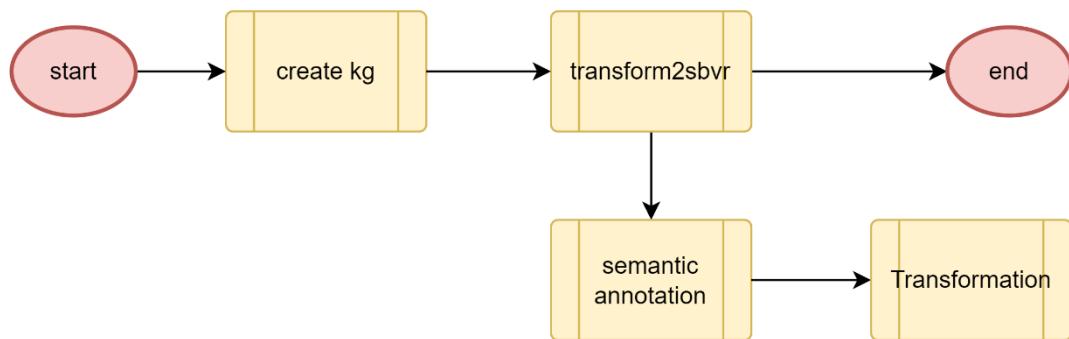
6.1.2 Infraestrutura para validação

A infraestrutura para a validação dos algoritmos é a mesma utilizada para a execução. Os componentes foram organizados em Jupyter Notebooks, um para cada processo principal. Foram desenvolvidos notebooks (code/src/chap_7_validation_*.ipynb) para validação, onde os cenários são executados conforme os cenários da seção “5.3 Cenários de validação” e protocolos descritos no capítulo “7 VALIDAÇÃO”.

6.2 Implementação dos principais componentes

A transformação de regras de negócio contidas nos regulamentos da CFR para o livro de regras no padrão SBVR requer um conjunto de processos. Como justificado no capítulo “4 DISCUSSÃO DAS SOLUÇÕES” e discutido no capítulo “5 TRANSFORMAÇÃO DE REGULAMENTOS PARA SBVR”, esse conjunto de processos, chamado de CFR2SBVR, é uma combinação de algoritmos determinísticos, que são implementados em linguagem de programação convencional e seu resultado varia apenas com a variação de suas entradas de dados; e não-determinístico, baseado em instruções em NL (*prompts*) passadas para o LLM juntamente com as entradas a serem processadas e, pela sua implementação probabilística, seu resultado varia com o mesmo conjunto de instruções e entrada de dados.

Figura 12 - Visão geral da transformação de CRF para SBVR.



Fonte: Elaborado pelo autor.

A implementação do CFR2SBVR, apresentado na Figura 12, foi dividido em quatro algoritmos:

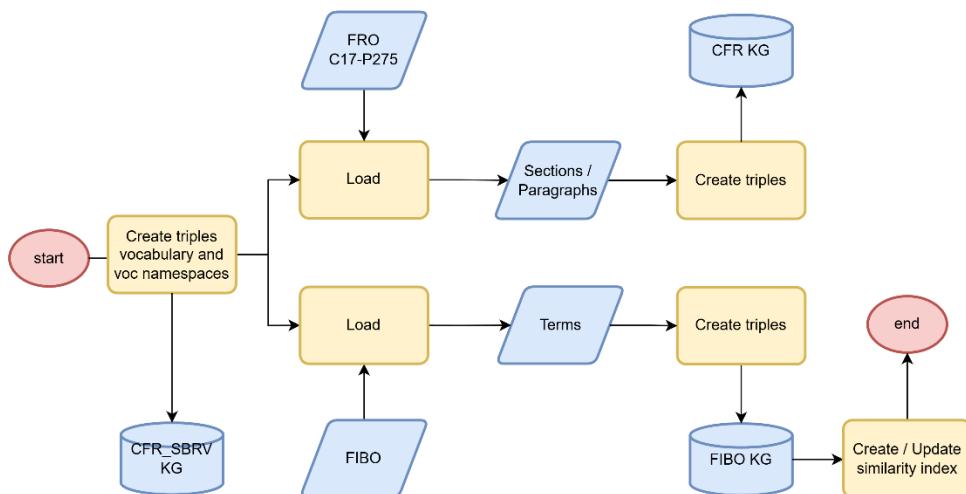
1. “create kg”: Responsável por carregar o KG com os conteúdos existentes da ontologia FIBO (KG FIBO) e os parágrafos da CFR (KG CFR). Explicado na seção “6.2.1 KG para CFR e FIBO”;
2. “transform2sbvr”: Responsável pela orquestração da transformação. Ele recupera as seções a serem processadas pelos demais processos. Essa tarefa é apresentada na seção “6.2.2 Orquestrador”;
3. “semantic annotation”: Sua principal função é realizar a anotação semântica, que extrai os enunciados para os termos, nomes, fatos e regras operativas, chamados de elementos, e os classifica seguindo a taxonomia de Witt (2012) e obtém o *template* para a transformação, criando metadados que serão utilizados na etapa seguinte. Explicado na seção “6.2.3 Anotações semânticas”;
4. “Transform”: Esta tarefa é encarregada de realizar a transformação das regras de definição e comportamentais. Utiliza um *template* para a transformação e o resultado, juntamente com os metadados, são persistidos no KG CFR-SBVR. A seção “6.2.4 ” apresenta este processo.

Os algoritmos são apresentados em fluxogramas e os principais passos são discutidos nessa seção. Os algoritmos não-determinísticos são validados usando metodologia descrita no capítulo “7 VALIDAÇÃO”. Todos os dados e algoritmos podem ser encontrados no repositório de códigos desta pesquisa⁵².

6.2.1 KG para CFR e FIBO

O KG descrito na seção “5.2.2 KG para CFR e FIBO” é obtido pelo componente “create kg”. Esse algoritmo faz parte do processo de extração.

Figura 13 - Componente “create kg”



Fonte: Elaborado pelo autor.

O algoritmo de carregamento dos dados da FIBO e CFR é ilustrado na Figura 13, ele também é o algoritmo de inicialização de um novo repositório no banco de dados e das triplas de vocabulário e vocabulário *namespaces*. Os algoritmos de carga são independentes, a tarefa de “Load” lê o arquivo rdf/turtle com as triplas das seções e parágrafos da CFR. Cada seção e parágrafo é inserido no grafo CFR seguindo a ontologia FRO. O processo é semelhante para FIBO, com a adição da tarefa de criação e atualização do índice para busca por similaridade. O modelo usado para criar os *embeddings* é o “text-embedding-3-small” da OpenAI. Para FIBO, o índice é o predicado “label” dos tipos “Class” e

⁵² <https://github.com/asantos2000/master-degree-santos-anderson>. Acesso em: 15/mar./2025.

“NamedIndividual”, de acordo com a sua ontologia⁵³. Para CFR os predicados são “hasParagraphText” e “hasParagraphEnumText” para o tipo “CFR_Paragraph” da ontologia⁵⁴.

Figura 14 - Consulta SPARQL para termos FIBO.

```

1 PREFIX fibo: <https://spec.edmcouncil.org/fibo/ontology/master/2024Q2/QuickFIBOProd#>
2 PREFIX fro-cfr: <http://finregont.com/fro/cfr/Code_Federal_Regulations.ttl#>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
5
6 SELECT ?term ?definition
7 FROM fibo:FIBO_Graph
8 WHERE {
9   { ?term a owl:NamedIndividual ;
10     skos:definition ?definition . }
11   UNION
12   { ?term a owl:Class ;
13     skos:definition ?definition . }
14 }
```

Fonte: Elaborado pelo autor.

As triplas são inseridas em grafos nomeados⁵⁵. Todas as triplas de termos do grafo FIBO podem ser acessadas em SPARQL referenciando o nome do grafo na cláusula “from”, como ilustrado na Figura 14.

Figura 15 - Consulta SPARQL para seção da CFR.

```

1 PREFIX fro-cfr: <http://finregont.com/fro/cfr/Code_Federal_Regulations.ttl#>
2 PREFIX fro-leg-ref: <http://finregont.com/fro/ref/LegalReference.ttl#>
3
4 SELECT ?section ?section_seq ?section_num ?section_subject ?section_citation ?section_notes ?divide ?
      divide_seq ?paragraph_enum ?paragraph_text
5 WHERE {
6   ?section a fro-cfr:CFR_Section ;
7   fro-leg-ref:hasSequenceNumber ?section_seq ;
8   fro-cfr:hasSectionNumber ?section_num ;
9   fro-cfr:hasSectionSubject ?section_subject .
10 OPTIONAL {?section fro-leg-ref:refers_toNote ?section_notes} .
11 OPTIONAL {?section fro-cfr:hasSectionCitation ?section_citation} .
12
13 ?divide fro-leg-ref:divides ?section ; # rdf:type fro-cfr:CFR_Paragraph
14   fro-leg-ref:hasSequenceNumber ?divide_seq ;
15   fro-cfr:hasParagraphText ?paragraph_text ;
16   fro-leg-ref:hasSequenceNumber ?paragraph_seq .
17 OPTIONAL {?divide fro-cfr:hasParagraphEnumText ?paragraph_enum} .
18 }
19 ORDER BY ?section num ?section ?divide seq
```

⁵³ FIBO Ontology: <https://spec.edmcouncil.org/fibo/ontology/master/2024Q3/prod.fibo-quickstart.ttl>. Acesso em: 15/mar./2025.

⁵⁴ CFR Ontology: https://finregont.com/fro/cfr/Code_Federal_Regulations.ttl. Há um erro de grafia no nome da classe CFR_Paragraph, pelo rótulo deveria ser CFR_Paragraph. Acesso em: 15/mar./2025.

⁵⁵ W3C Named Graphs: <https://www.w3.org/2009/07/NamedGraph.html#named-graphs>. Acesso em: 15/mar./2025.

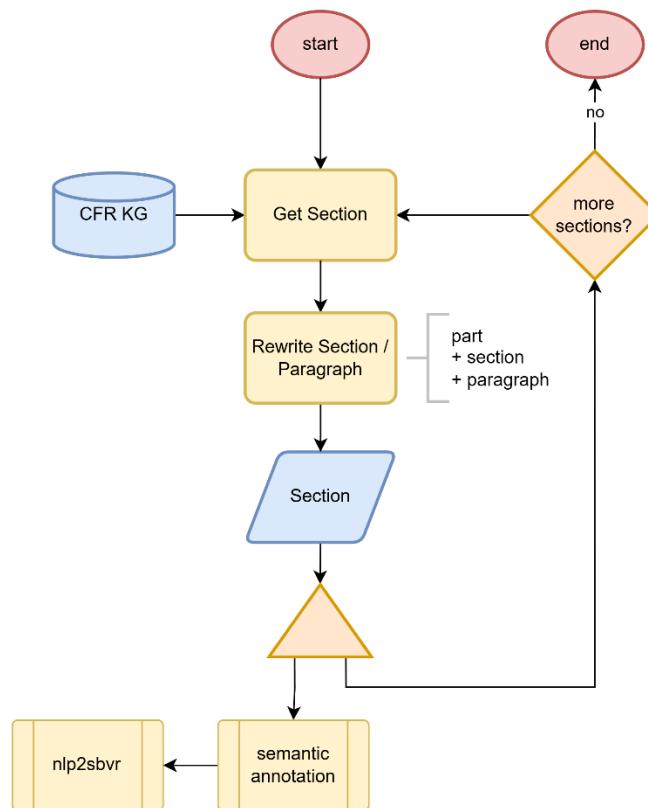
Fonte: Elaborado pelo autor.

De forma semelhante, as triplas do grafo CFR podem ser acessadas com a consulta SPARQL ilustrada na Figura 15. Como discutido na seção “5.2.2 KG para CFR e FIBO”, o grafo foi criado utilizando o arquivo RDF/turtle criado pela Jayzed Data Models inc. (2021).

6.2.2 Orquestrador

O algoritmo “transform2sbvr” é responsável por coordenar o processo de transformação, ele recupera do banco de dados as seções e os parágrafos e reescreve para adicionar contexto do que é parte, seção e parágrafos com objetivo de auxiliar o processo de anotação semântica e transformação. O algoritmo é ilustrado na Figura 16.

Figura 16 - Algoritmo orquestrador – “transform2sbvr”



Fonte: Elaborado pelo autor.

Os parágrafos reescritos incluem metadados da ontologia, tais como:

- section_num – Identificador da seção, exemplo: § 275.0-2;
- section_subject – O nome da seção, exemplo: General procedures for serving non-residents.
- section – A URI que identifica a seção, exemplo: http://finregont.com/fro/cfr/FRO_CFR_Title_17_Part_275.ttl#r-1-4
- section_citation – Indica se a seção faz referência a documentos que originaram o texto da seção, como leis e emendas, exemplo: [65 FR 57448, Sept. 22, 2000]
- section_notes – Se há alguma nota sobre a seção.
- paragraph_enum – Um ou mais parágrafos, indica que os parágrafos a seguir são parte do mesmo parágrafo, são itens de uma enumeração.
- paragraph_text: texto do parágrafo.

Quadro 11 - Problema com o KG CFR

Texto original

(a) For purposes of Commission rulemaking in accordance with the provisions of Chapter Six of the Administrative Procedure Act (5 U.S.C. 601 et seq.) and unless otherwise defined for purposes of a particular rulemaking proceeding, **the term small business or small organization** for purposes of the Investment Advisers Act of 1940 shall mean an investment adviser that:

Fragmento em XML

<P>

(a) For purposes of Commission rulemaking in accordance with the provisions of Chapter Six of the Administrative Procedure Act (5 U.S.C. 601

<I>et seq.</I>

) and unless otherwise defined for purposes of a particular rulemaking proceeding, **the term**

<I>small business</I>

or

<I>small organization</I>

for purposes of the Investment Advisers Act of 1940 shall mean an investment adviser that:

</P>

Fragmento em RDF/Turtle carregado no grafo:

```
fro-cfr-t17-p275:r-1-9-2
rdf:type fro-cfr:CFR_Paragraph ;
fro-cfr:hasParagraphEnumText “small business” ;
fro-cfr:hasParagraphEnumText “small organization” ;
fro-cfr:hasParagraphText “(a) For purposes of Commission rulemaking in accordance with the provisions of Chapter Six of the Administrative Procedure Act (5 U.S.C. 601 et seq.) and unless otherwise defined for purposes of a particular rulemaking proceeding, the termor” ;
fro-cfr:hasParagraphText “for purposes of the Investment Advisers Act of 1940 shall mean an investment adviser that:” ;
fro-leg-ref:divides fro-cfr-t17-p275:r-1-9 ;
fro-leg-ref:hasSequenceNumber 2 ;
fro-ref:hasSourceInstance <http://finregont.com/fro/cfr/CFR-2012-title17-vol3-part275.ttl#r-1-9-2> ;
```

Fonte: Elaborado pelo autor.

Esse algoritmo não foi usado no experimento devido ao problema que comprometeu o seu uso e a precisão dos resultados. Durante os experimentos foram encontrados erros nos textos da CFR-FRO processados pela Jayzed Data Models inc. (2021). Um erro comum é a quebra de parágrafos (predicado fro-cfr:hasParagraphEnumText) em duas ou mais partes, muitas vezes separando o identificador do parágrafo (exemplo (a), (b), (1), etc.) do restante do texto. Em outros casos, os parágrafos de uma seção não correspondem aos da eCFR, exemplo seção § 275.0-6, na FRO contém os parágrafos (a) a (f) e na eCFR os parágrafos (a), (b) e (c), sendo que (f) e (c) são iguais⁵⁶. Uma análise mais detalhada identificou que o texto da FRO está desatualizado e corresponde a versão de abril de 2019.

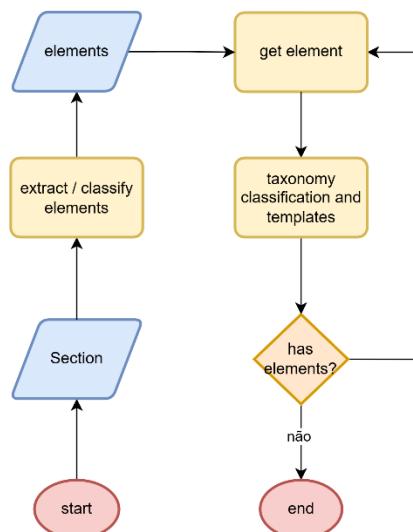
Também foi identificado na recuperação dos textos das seções no KG CFR que no texto da CFR, por exemplo, alguns termos destacados em itálico foram omitidos durante a transformação, como ilustrado Quadro 11. Os termos “small business” e “small organization” são descritos no texto original do regulamento como: “the term **small business** or **small organization**”. No fragmento XML,

⁵⁶ Comparação entre a versão de 02/04/2019 e a versão de 02/08/2024:
<https://www.ecfr.gov/compare/current/to/2019-04-02/title-17/chapter-II/part-275>. Acesso em: 15/mar./2025.

esses termos são formatados em itálico (*<I>*, *</I>*), mas não aparecem no arquivo RDF importado. No arquivo RDF, o trecho apresenta erro de grafia: “... the termor ...”, onde os três pontos representam os espaços reservados para os termos ausentes. Esse erro fez com que o LLM não conseguisse identificar corretamente os termos.

A solução de contorno foi utilizar o texto original disponível na eCFR e adaptar o *prompt* para que o LLM identificasse os metadados do texto, em vez de recuperá-lo do KG. Para trabalhos futuros, um novo processo de conversão da CFR para FRO poderá corrigir esse problema.

Figura 17 - Algoritmo “semantic annotation”



Fonte: Elaborado pelo autor.

6.2.3 Anotações semânticas

A anotação semântica é realizada pelo algoritmo “semantic annotation”. Ele é responsável pela extração (P2), relacionamento (P3), desambiguação (P4) dos conceitos e identificação e classificação dos enunciados (P6), conforme discutido na seção “5.2.3 Anotações semânticas”. O principal objetivo desse algoritmo é extrair trechos do texto e enriquecê-los com metadados para o processo de transformação. A Figura 17 ilustra o algoritmo, ele recebe uma seção e extraí e classifica os trechos do texto em termos, fatos ou regra, chamados simplesmente de elementos. O resultado desse algoritmo são as regras candidatas, discutidas na seção “5.2.4 Regras candidatas”.

Algoritmo “extract / classify elements”

O algoritmo “extract / classify elements”, faz parte do processo de extração, ele utiliza o LLM para extrair e classificar trechos do texto do documento fornecido (seção) em elementos e registra a sua origem no parágrafo e o escopo da definição.

Ele também extrai os enunciados e elementos de forma atômica. Segundo Witt (2012, p. 214-215), uma regra deve ser atômica, podendo ser completamente violada ou não, uma violação ou conformidade parcial não deve ser possível. Haarst (2013, p.75) recomenda manter as regras atômicas e sugere formas de identificá-las, segundo o autor “or” é um indicador de que a regra deveria ser partida em duas e “and” e “or” introduzem ambiguidade e devem ser reformuladas. Roychoudhury *et al.* (2017) trata esse problema no passo de pré-processamento dividindo o texto em sentenças, embora sua abordagem não teste se a sentença resultante é atômica.

A especificação do SBVR trata essa propriedade da regra como formulação atômica. Uma formulação atômica é uma formulação lógica baseada em um conceito verbal e envolve vinculações de papéis para cada papel desse conceito verbal. Essa formulação expressa que há uma ocorrência específica envolvendo cada papel do conceito verbal, correspondendo a um determinado alvo ao qual cada vinculação de papel se refere, assim, uma regra atômica essencialmente fornece uma declaração lógica precisa e indivisível envolvendo um conceito verbal específico.

Foram testadas duas abordagens: (1) em uma única consulta ao LLM, a extração e classificação de fatos, regras e termos, juntamente com a definição dos termos e suas relações de sinônímia; (2) e a extração dos termos, fatos e regras e, em uma segunda consulta ao LLM, a definição dos termos e a relação entre eles e pôr fim a classificação com a taxonomia. Embora útil, a primeira abordagem visava reduzir o custo executando uma única consulta ao LLM.

Quadro 12 - *Prompt*⁵⁷ para extração de termos – primeira abordagem

You are tasked with extracting elements and **relationships** from a given legal document.

Please follow these steps carefully and ensure all instructions are adhered to:

****Steps**:**

1. **Summarize the document:**

- Summarize the document to understand its purpose and use it to verify if all important terms, term definitions, facts, fact types, and rules are identified in subsequent steps.

2. **Identify Facts, Fact Types, and Rules:**

- **Definitions:**

- ****Fact**:** A specific instance or statement that describes an event or condition without any directive element. Facts often involve relationships between terms or entities. Example: “John works for X Inc.”

- ****Fact Type**:** A general, abstract template that describes potential relationships between terms or entities, serving as a model for generating specific facts. Example: “Person works for Company.”

- ****Rule**:** A statement that governs or constrains some aspect of the business, specifying what must be done or what is not allowed. Rules enforce compliance, limit possibilities, or prescribe specific behaviors in response to business situations. Example: “A customer must provide identification before opening an account.”

- **For each fact, fact type, or rule:**

- ****Extract the Statement**:** Identify the exact statement or phrase from the document representing the fact, fact type, or rule.

- ****Extract Terms**:** List all the terms involved in the statement.

- ****Extract Verb Symbols**:** Identify verbs, verb phrases, or prepositions that connect the terms in the statement.

- ****Classification**:** Classify the statement as either a ****Fact****, ****Fact Type****, or ****Rule****.

- ****Source**:** Note the specific paragraph or section of the document where the statement is found (e.g., “(a)(1)”, “(b)”).

3. **Classify Terms:**

- For each term extracted classify it as either a ****Common Noun**** or a ****Proper Noun****.

⁵⁷ Os prompts destinados a LLMs frequentemente utilizam a linguagem de marcação markdown para organização e apresentação do texto. Essa linguagem é reconhecida pelos LLMs e oferece uma maneira de representar hierarquias, realces e sequências de maneira clara e não ambígua, facilitando a interpretação pelo modelo. No entanto, não foram identificados artigos científicos que validem diretamente essa prática. O suporte à utilização de markdown nesse contexto baseia-se principalmente em experimentos empíricos e na literatura cinza disponível.

4. **Define Terms:**

- For each term:
 - Search the entire document for the term's definition, explanation, or meaning. Also, look in the document summary.
 - If the definition is found, include it.
 - If the definition is not found in the document, use **None**.

5. **Identify Relationships Between Terms:**

- **Types of Relationships**:
- **Synonym**: Terms that can be used interchangeably without changing the meaning.
- **For each pair of terms in the document**:
- Identify if a relationship exists as either "Synonym".
- Only include relationships where both terms are present in the document.

6. **Provide JSON Output:**

- Format your answer as per the output example below.
- **All values are optional**: Include as much information as is available based on the document.
- **Do not include any additional text or explanation outside the JSON structure**.

Output Example:

```
```json
{
 "section": "§ 123.4-5",
 "elements": [
 {
 "id": 1,
 "statement": "A person serves a non-resident investment adviser by furnishing the Commission with process, pleadings, or papers.",
 "terms": [
 {
 "term": "Person",
 "classification": "Common Noun",
 "definition": "An individual or legal entity."
 },
 {
 "term": "Non-resident investment adviser",
 "classification": "Common Noun",
 "definition": null
 }
]
 }
]
}
```

```

},
...
],
“verb_symbols”: [“serves”, “by furnishing”, “with”],
“classification”: “Fact Type”,
“source”: “(a)”
},
...
],
“terms_relationship”: [
{
“terms”: [
“Principal office”,
“Place of business”
],
“relation”: “Synonym”
},
{
“terms”: [
“Person”,
“Individual”
],
“relation”: “Synonym”
},
...
]
}
...

```

Fonte: Elaborado pelo autor.

O *prompt*, apresentado no Quadro 12, utilizou a técnica CoT e foi enviado para o LLM, juntamente com o texto da seção (documento a ser processado). Há sete funções chave neste algoritmo:

1. A identificação e extração dos elementos. É dado ao LLM a definição de fatos e regras, baseados na definição da SBVR (OMG, 2019), e é pedido para que identifique os enunciados dos textos que correspondem a cada uma das classes;

2. Identificador da origem do elemento no texto. Identificador do parágrafo de onde o enunciado foi extraído. Essa informação será usada para a criação dos vocabulários e rastreabilidade;
3. A identificação dos termos. Para cada fato ou regra é solicitado que os termos sejam identificados.
4. Classificação dos termos em substantivos comum ou próprios. Para cada termo identificado é solicitado que seja classificado em substantivo comum ou próprio.
5. Extração dos verbos e preposições que associam os termos. É solicitado que o LLM crie uma lista dos verbos que relacionam os termos nos fatos ou regras.
6. Extração da sua definição. Para cada termo identificado busca-se no texto a sua definição;
7. Termos sinônimos: O LLM é instruído a identificar termos que são sinônimos no texto;

Quadro 13 – Fragmento do resultado do algoritmo “extract / classify elements”.

```
{
 "section": "§ 275.0-7",
 "elements": [
 {
 "id": 1,
 "statement": "An investment adviser that has assets under management of less than $25 million is considered a small business for the purposes of the Investment Advisers Act of 1940.",
 "terms": [
 {
 "term": "Investment adviser",
 "classification": "Common Noun",
 "definition": None
 },
 {
 "term": "Assets under management",
 "classification": "Common Noun",
 }
]
 }
]
}
```

```

"definition": None
},
{
"term": "$25 million",
"classification": "Proper Noun",
"definition": None
},
{
"term": "Small business",
"classification": "Common Noun",
"definition": None
}
],
"verb_symbols": ["has", "is considered"],
"classification": "Fact Type",
"source": "(a)(1)"
},
...

```

Fonte: Elaborado pelo autor.

Um fragmento do resultado da execução do algoritmo da primeira abordagem é apresentado no Quadro 13. O resultado foi uma identificação de cinco elementos de sete e a extração de 16 termos de 21. Para a extração de definições, o melhor resultado foi a definição de dois termos, mas não os termos chave, e na maioria dos casos o resultado das execuções foi de zero definições.

Na avaliação qualitativa dos resultados, a extração dos fatos e regras foram consistentes entre as execuções, porém os resultados obtidos para o termo não, na maioria das execuções, o algoritmo falhou para extrair a definição do termo, mesmo quando a definição estava clara no texto, como no documento 275.0-7 onde o fragmento diz ... the **term** small business or small organization for purposes of the Investment Advisers Act of 1940 shall **mean** an investment adviser that: ...”, o *prompt* falhou em definir “small business” e “small organization”, que é o principal propósito do documento.

O problema com essa abordagem, discutido na próxima seção, motivou uma segunda iteração na pesquisa. A forma para contornar o baixo desempenho na definição dos termos foi experimentar diferentes formulações, estruturas e estilos

de *prompts* para identificar aqueles que obtivessem os melhores resultados para a tarefa (TONMOY *et al.*, 2024). Pelo mesmo motivo, essa abordagem não classifica os elementos segundo a taxonomia, o que aumentaria ainda mais a sua complexidade.

A segunda abordagem dividiu a tarefa em dois algoritmos:

1. O algoritmo “extract / classify elements” foi dividido em mais duas partes:
  - Parte 1 (funções de 1 a 5), é responsável por identificar e classificar elementos como “Fatos”, “Tipos de Fatos”, “Regras Operacionais” e “Termos” em um documento legal. Ele analisa cada componente, identificando verbos e expressões que conectam termos, e classifica os termos como substantivos próprios ou comuns, além de registrar a origem de cada elemento.
  - Parte 2 (funções 7 e 8) foca na extração de definições, escopo e na identificação de relações de sinônima entre os termos listados. Inicialmente, o sistema resume o documento para verificar se todos os principais elementos foram identificados, busca definições para cada termo, e caso não as encontre, atribui “None”. Após essa etapa, compara os termos para identificar sinônimos no mesmo contexto. Os resultados, com definições e relações de sinônima, são apresentados em um formato JSON estruturado.
2. O algoritmo “taxonomy classification and templates”, ficou responsável pelas novas funções e é dividido em duas partes:
  - Parte 1 classifica os elementos no topo da taxonomia de Witt (2012), sendo elas regras de definição e operacionais (atividade, papel e dados), seguindo da classifica dos

enunciados das regras operativas de acordo com a taxonomia das regras operativas;

- Parte 2 classificações os termos, nomes e fatos na taxonomia das regras de definição.

Cada função constitui passo no processo de transformação e erros em um passo se acumulam para os passos posteriores.

Quadro 14 - *Prompt* do algoritmo “extract / classify elements” parte 1. Segunda abordagem.

You are tasked with extracting elements from a given legal document. Please follow these steps carefully and ensure all instructions are adhered to:

#### # Steps

1. \*\*Summarize the document\*\* to understand its purpose and use it to verify if all important terms,facts, fact types, and rules are identified in subsequent steps.

2. \*\*Identify elements\*\*:

- \*\*About the elements\*\*:

- \*\*Fact\*\*: A specific instance or statement that describes an event or condition without any directive element. Facts often involve relationships between terms or entities. Example: “John works for X Inc.”

- \*\*Fact Type\*\*: A general, abstract template that describes potential relationships between terms or entities, serving as a model for generating specific facts. Example: “Person works for Company.”

- \*\*Operative Rule\*\*: A statement that governs or constrains some aspect of the business, specifying what must be done or what is not allowed. Rules enforce compliance, limit possibilities, or prescribe specific behaviors in response to business situations. Operative rules (otherwise known as normative rules or prescriptive rules) state what must or must not happen in particular circumstances. Operative rules can be contravened: required information may be omitted, inappropriate information supplied, or an attempt may be made to perform a process that is prohibited. Example: “A customer must provide identification before opening an account.”

- \*\*Term\*\*: A word or a group of words that represents a specific concept, entity, or subject in a particular context.

- Terms, Fact, Fact Type, and Operative Rule are statements that should allow only full compliance or full contravention; partial compliance is not possible. The presence of “or” or “and” often suggests the need to separate a statement into two.
- \*\*For each fact, fact type, or rule\*\*:
- \*\*Extract the statement\*\*: Identify the exact statement or phrase from the document representing the fact, fact type, or rule.
- \*\*Give a unique title to the statement\*\*.
- \*\*Extract and classify Terms\*\*:
- \*\*Extract all the terms involved in the statement\*\*: Record the level of confidence in the extraction, ranging from 0 to 1, and provide a brief reason for the confidence score.
- \*\*Classify each term\*\* as either \*\*Common Noun\*\* or \*\*Proper Noun\*\*.
- If a Term contains nouns separated by “and,” “,”, or “or,” split it into two or more terms. For example, “Principal office and place of business” should be split into “Principal office” and “Place of business”.
- \*\*Extract Verb Symbols\*\*: Identify verbs, verb phrases, or prepositions that connect the terms in the statement. Record the level of confidence in the extraction, ranging from 0 to 1, and provide a brief reason for the confidence score.
- \*\*Classification\*\*: Classify the statement as either a \*\*Fact\*\*, \*\*Fact Type\*\*, or \*\*Rule\*\*.
- \*\*Confidence\*\*: Record the level of confidence in the classification, ranging from 0 to 1.
- \*\*Reason\*\*: Provide a brief reason for the classification score.
- \*\*Source\*\*: Note the specific paragraph or section of the document where the statement is found (e.g., “(a)(1)”, “(b)”).

### 3. \*\*Provide JSON Output\*\*:

- Format your answer as per the output example below.
- \*\*All values are optional\*\*: Include as much information as is available based on the document.
- \*\*Do not include any additional text or explanation outside the JSON structure\*\*.

**\*\*Output Example\*\*:**

```
```json
{
  "section": "§ 123.4-5",
  "elements": [
    {
      "id": 1,
      "title": "some title",
      "statement": "A person serves a non-resident investment adviser by furnishing the Commission with process, pleadings, or papers."
    }
  ]
}
```

```

"terms": [
{
  "term": "Person",
  "classification": "Common Noun",
  "confidence": 0.9,
  "reason": "The term is ...",
  "extract_confidence": 0.9,
  "extract_reason": "The term is ..."
},
{
  "term": "Non-resident investment adviser",
  "classification": "Common Noun",
  "confidence": 0.8,
  "reason": "The term is ...",
  "extract_confidence": 0.8,
  "extract_reason": "The term is ..."
},
...
],
"verb_symbols": ["serves", "by furnishing", "with"],
"verb_symbols_extracted_confidence": [0.9, 0.8, 0.7],
"verb_symbols_extracted_reason": ["The verb is ...", "The verb is ...", "The verb is ..."],
"classification": "Fact Type",
"confidence": 0.8,
"reason": "The statement is ...",
"sources": ["(a)"]
},
...
]
}
```

```

## # Notes

1. Level of Granularity: Extract and analyze every potential statement from the document;
2. Contextual Interpretation: Extract explicitly stated facts, fact types, and rules;
3. Scope of Terms: Classify every noun phrase as a term, even if it is peripheral to the main statement;
4. Verb Symbols: Include prepositions and auxiliary phrases;
5. Classification Nuances: Record the level of confidence range from 0 to 1;
6. Section Handling: Strictly tie every element to its section (e.g., § 275.0-7(a)(1));

- |                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> <li>7. Order of Presentation: Follow the sequence of the document strictly;</li> <li>8. Edge Cases: Classify it with a lower level of confidence;</li> <li>9. Output Preferences: Add timestamp and processing notes;</li> <li>10. Formatting Precision: Ensure the JSON adheres strictly to a specific schema (e.g., for use in a system).</li> </ol> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Fonte: Elaborado pelo autor.

A comparação do desempenho das duas abordagens para o algoritmo “extract / classify elements” utilizou apenas uma seção da CFR e as funções de um a oito foram avaliadas. A seção 275.0-7 define o que constitui uma “small business” ou “small organization”.

**Quadro 15 - *Prompt* do algoritmo “extract / classify elements” parte 2. Segunda abordagem.**

|                                                                                                                                                                                                                      |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are tasked with extracting definitions and **relationships** of terms in the terms list searching a given legal document. Please follow these steps carefully and ensure all instructions are adhered to:</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

#### # Steps

1. \*\*Summarize the document\*\* to understand its purpose and use it to verify if all important terms, term definitions, facts, fact types, and rules are identified in subsequent steps.

2. \*\*Define terms\*\*:

- For each term:
  - Search the entire document for the term's definition, explanation, or meaning. Also, look in the document summary.
  - If the definition is found, include it.
  - If the definition is not found in the document, use null.
  - Record the level of confidence in the definition, ranging from 0 to 1.
  - Explain the reason for the confidence level.

3. \*\*isLocalScope\*\*: Is there an indication that the definition is exclusive to this section? Example: “For purposes of this section...”, “as described in this section”, “as defined in this section”. If yes answer only with true. Otherwise, the answer is false.

4. \*\*Identify synonym relationships between terms\*\*:

- For each term in the terms list:
  - Compare it against other terms in the text to find synonyms.

- Ensure both terms exist within the same document context.
- List all valid synonym pairs identified.
- Record the level of confidence in the synonym relationship, ranging from 0 to 1.
- Explain the reason for the confidence level.

5. **Provide JSON Output\*\*:**

- Format your answer as per the output example below.
- **All values are optional\*\*:** Include as much information as is available based on the document.
- **Do not include any additional text or explanation outside the JSON structure\*\*.**

**Output Example\*\*:**

```
```json
{
  "terms": [
    {
      "term": "Person",
      "definition": "A person is a person.",
      "confidence": 0.9,
      "reason": "The definition was ...",
      "isLocalScope": true,
      "local_scope_confidence": 0.9,
      "local_scope_reason": "The scope is ..."
    },
    {
      "term": "Capital",
      "definition": "The total assets of a person.",
      "confidence": 0.8,
      "reason": "The definition is ...",
      "isLocalScope": false,
      "local_scope_confidence": 0.9,
      "local_scope_reason": "The scope is ..."
    },
    ...
  ],
  "relationships": [
    {
      "term_1": "Person",
      "term_2": "Capital",
      "confidence": 0.7
    }
  ]
}
```

```

"relationship": "Synonym",
"confidence": 0.8,
"reason": "The relationship is ...",
},
{
"term_1": "Capital",
"term_2": "Person",
"relationship": "Synonym",
"confidence": 0.5,
"reason": "The relationship is ...",
},
...
]
}
...

```

Fonte: Elaborado pelo autor.

O objetivo do *prompt*, apresentado no Quadro 14, é a extração e classificação de elementos (termos, fatos, tipos de fato e regras operativas) de uma seção. Ele utilizou a mesma técnica CoT e envia para o LLM as instruções, juntamente com o texto da seção. Há quatro funções chave neste algoritmo: (1) a identificação e extração dos elementos; (2) a classificação do tipo de elemento (termo, fato ou regra); (3) A extração dos termos e a classificação em substantivo comum ou próprio; (4) e a extração dos verbos e preposições que associam os termos. O segundo *prompt*, ilustrado no Quadro 15, é responsável pela segunda etapa: (1) a extração dos textos que definem os termos; e (2) a relação entre eles. Este *prompt* relaciona apenas os termos sinônimos. Em trabalhos futuros, a busca por outras relações como hiperônimos pode ser usado na melhoria da transformação dos fatos.

O resultado foi uma identificação de sete elementos, em vez de cinco da primeira abordagem, um incremento de 40% na extração de fatos e regras, extração 21 termos, de 16, um incremento de 31% na identificação de termos e a extração correta das definições, na qual a primeira abordagem falhou. O teste das duas abordagens está no repositório de código e no capítulo “7 VALIDAÇÃO” é apresentado os resultados da segunda abordagem em relação ao conjunto de dados ouro. A implementação dessa função pode ser encontrada no repositório

de código no arquivo "code/src/chap_6_semantic_annotation_elements_extraction.ipynb".

Algoritmo “taxonomy classification and templates”

Esse algoritmo faz parte do processo de classificação, ele classifica individualmente os elementos para controlar a quantidade de dados fornecidos para o LLM e minimizar os erros. Essa abordagem se apoia na distinção entre regras de definição e operativas, permitindo focar a atenção do modelo. Ele classifica os elementos extraídos no processo de extração.

Os elementos são classificados em: regra de definição, que visa estruturar os conceitos estabelecendo declarações que sustentam a interpretação uniforme da linguagem de negócios. Essas regras formalizam termos, criam esquemas de categorização e delineiam relações, utilizando tipos de fatos que representam desde associações simples até relacionamentos complexos; e a regra operativa, que regulam os requisitos práticos nos processos organizacionais, estipulando condições para ações e papéis específicos.

Na taxonomia de Witt (2012), as regras de definição e operativas são divididas em tipos e subtipos e associados a *templates* que provêm consistência na representação e facilitam a interpretação das declarações, assegurando padronização na forma de documentar as regras.

Para ilustrar, considere o termo “Comissão” no enunciado “Comissão é a entidade com a qual processos, petições ou documentos são arquivados e que encaminha esses documentos para as partes nomeadas”. Esse enunciado pode ser formalmente estruturado com o *template* T7 da taxonomia de regras de definição, cujo formato é: “[A|An] <term 1> {of [a |an] <term 2>} |} is by definition {a|an|the} <term 3> <qualifying clause>“. Aplicando esse *template*, o enunciado é transformado em: “A Comissão é definida como a entidade responsável por receber (arquivar) processos, petições ou documentos e encaminhá-los às partes nomeadas.” Esse *template* articula o escopo, o significado e as responsabilidades da “Comissão” dentro do processo organizacional.

Quadro 16 - *Prompt* para classificação dos elementos no topo da taxonomia de Witt (2012).

You are an expert in SBVR (Semantics of Business Vocabulary and Business Rules).

You are working for regulatory bodies, auditors, or process managers.

You will be provided with a list of statements formatted as JSON.

Your task is to classify each statement into one or more Operative Rules types according to the given definitions.

Steps

1. **Summarize statement**: Summarize the given statement to understand its structure and content.

2. **Classify statement**: Classify each Operative Rule statement into one or more of the provided rule types. The **Operative rules** govern actions or constraints that must or must not happen under certain conditions, such as Data Rules, Activity Rules, and Party Rules. types to classify are:

- **Party rules**: A “Party rule” is a type of operative rule that establishes distinctions or constraints involving parties or the roles they perform. To identify a party rule, it is important to recognize its defining characteristics, as these rules often specify who can carry out certain activities, access particular information, or hold specific responsibilities. Party rules may include restrictions on who is permitted to perform specific roles or processes. For example, a rule might state that a person can serve as the pilot in command only if they hold a current command endorsement. Additionally, these rules may enforce role separation to prevent conflicts of interest, such as a requirement that the cabin crew member verifying an aircraft door's disarmed status cannot be the same individual who initially disarmed it. In other cases, party rules may require role binding, ensuring continuity by stipulating that the consultant who signs a quality review report must be the one who conducted the review. Party rules can also govern information access, specifying who is authorized to view, create, or modify certain data. For instance, a rule might state that an employee's leave record can only be accessed by the employee, their supervisor, or a human resources officer. Furthermore, responsibility rules fall under this category by defining accountability for specific actions or obligations, such as requiring the receiving parties in a property transfer to pay the associated stamp duty. These rules can be identified by linking actions or processes (predicates) to subjects, like roles or data, while applying conditions that qualify or limit their application.

- **Data rules**: A “Data rule” imposes constraints or requirements on the data used in transactions, records, or systems. Identifying a data rule involves analyzing its structure, purpose, and type, which include cardinality, content, and update rules. Data cardinality rules govern the presence and multiplicity of data items. These may include mandatory rules requiring data items, such as specifying at least one passenger name in a flight booking confirmation. They can also restrict data, such as ensuring a one-way flight booking does not include a return date, or enforce limits on the number of data instances in a transaction. Data content rules regulate the values within data items. Examples include value set rules, which require a data item to match one of a specified set of valid values, and range rules that constrain a data item's value to within a specific range. Equality rules ensure consistency between related data items, such as requiring that an origin city matches the corresponding booking request. Additionally, uniqueness constraints ensure that a data item's value does not duplicate within a dataset, and consistency rules maintain logical relationships between multiple data points. Lastly, data update rules constrain modifications to existing data. These rules may prohibit updates entirely, restrict the scope of permissible changes (such as maintaining valid state transitions), or enforce monotonic trends like numeric values that can only increase or decrease. To identify a data rule, it is essential to examine the specific constraints or requirements applied to data items, their interrelationships, and their contexts within a given system or transaction. Templates and formalized structures, can aid in distinguishing these rules effectively.
- **Activity rules**: An “Activity rule” is an operative rule designed to constrain the operation of business processes or activities. Identifying an activity rule involves understanding its subcategories, which define how activities are regulated or mandated. The primary types of activity rules are activity restriction rules, activity obligation rules, and process decision rules. Activity restriction rules are used to place limitations on when or under what conditions an activity can occur. For example, time-based restrictions may stipulate that online check-in for a flight can only occur within a specific time window, such as the 24 hours before departure. Similarly, exclusion period rules prohibit activities during certain times, such as a restriction on operating machinery during nighttime hours. Activity pre-condition rules ensure that an event must occur before another activity can take place, such as requiring passengers to complete a security screening before boarding. Activity obligation rules, on the other hand, specify activities that must be performed either within a maximum time after a triggering event or as soon as practical. For instance, acknowledging an order may be mandated to occur within 24 hours of receipt. These rules enforce timely action and compliance with operational requirements. Process decision rules determine actions in response to specific situations. These rules guide devices or processes, such as ensuring that a ticket barrier retains invalid tickets to prevent misuse. To identify an activity rule, it is essential to look for statements that define constraints, obligations, or decision-making criteria for activities. Such rules are often structured using specific templates that articulate the conditions, timeframes, or triggers associated with the activity. Recognizing these elements helps in categorizing activity rules effectively.

2. Assess a **confidence level** for each classification between 0 and 1. Assign confidence scores to each class for the given statement, ensuring that no two classes receive the same score. If one class is assigned a score (e.g., 0.6), the others must have distinct values that are either higher or lower. The scores should reflect the likelihood of each class while avoiding ties.
3. **Explain classification**: You also need to record a confidence level for each classification and provide an explanation for why the classification was made.
4. **Repeat for each statement**: Repeat the process for each statement in the list.
5. **Output format**: Your output must also be in JSON format. It should contain, for each statement:
 - The `doc_id`
 - The `statement_id`
 - The `statement_title`
 - The original `statement_text`
 - The `statement_sources` of the statement
 - A list of classifications (`classification`), each containing:
 - The `type` of the rule.
 - The `confidence` in your classification.
 - An `explanation` detailing why you made the classification decision.

Here is an example of the expected output:

```
```
[
{
 "doc_id": "some doc id",
 "statement_id": "some id",
 "statement_title": "some title",
 "statement_text": "some text",
 "statement_sources": ["some source"],
 "classification": [
 {
 "type": "Activity rules",
 "confidence": 0.9,
 "explanation": "This statement defines ..."
 },
]
},
```

```
{
 "type": "Party rules",
 "confidence": 0.2,
 "explanation": "There is little reference ..."
},
...
]
},
{
 "doc_id": ...,
 "statement_id": ...,
 "statement_title": ...,
 "statement_text": ...,
 "statement_sources": ...,
 "classification": ...
}
]
```
# Notes
- **Detail the Reasoning**: Make sure to provide explanations that justify why a particular rule type was chosen.
- **Confidence Values**: The confidence value should genuinely represent how strongly you believe the classification is correct, with 1 being an absolute match and 0 meaning unlikely.

Make sure that every statement is analyzed thoroughly, and the final justification for each classification is straightforward and adequately supports both the type choice and confidence level.
```

Fonte: Elaborado pelo autor.

O *prompt* no Quadro 16 classifica os elementos do tipo “Operative rule” em regras operativas do tipo atividade, papel ou dados. Os elementos termos, nomes e fatos não precisam dessa primeira classificação pois são, por definição, da categoria “*Definitional*”. Para este *prompt* são fornecidos os enunciados a serem classificados. O resultado é um conjunto de classificações com uma pontuação de confiabilidade que varia de zero a um, em que a proximidade de 1 (um) indica que o LLM está “confiante” na sua classificação. O LLM foi instruído a não atribuir a mesma pontuação de confiança para classificações distintas. O

objetivo é evitar empates e direcionar o LLM a uma análise de nuances que diferencie suas escolhas.

Quadro 17 - *Prompt* para classificação na taxonomia e obtenção do *template* para transformação (parte 1 – regras operativas).

You are an expert in **SBVR (Semantics of Business Vocabulary and Business Rules)**, working for regulatory bodies, auditors, or process managers.

Your task is to classify {element_count} {element_type}(s) from a provided list into one or more **{rule_type} Rule subtypes**, explain each classification in detail, and assign a confidence score ranging from 0 to 1.

Approach:

Use the **{rule_type} Rule subtype definitions**, associated templates, and guidelines to analyze each statement thoroughly, ensuring accurate classification.

Steps

1. **Classify the Statement**:

- For each `statement_text`, determine its rule subtype according to the provided **{rule_type} Rule subtypes** and their corresponding templates.
- Use the provided templates, definitions, and examples to match the statement to the correct subtype.
- If the statement does not align with a high-level type, analyze the sublevels.
- The subtype to be used starts with “subtype: <subtype name>“.
- You should assign a subtype and a template ID, make your best guess, justify your choice, and lower the confidence level if necessary.
- Templates and examples help identify subtypes.

2. **Assign Confidence Level**:

- Assign a confidence score between **0 and 1** for each classification:
- **1** indicates a strong and clear match.
- Lower scores reflect weaker matches due to ambiguities or partial alignment.
- Consider both template alignment and the clarity of the statement's intent when assigning scores.
- Assign confidence scores to each class for the given statement, ensuring that no two classes receive the same score. If one class is assigned a score (e.g., 0.6), the others must have

distinct values that are either higher or lower. The scores should reflect the likelihood of each class while avoiding ties.

3. **Provide an Explanation**:

- Provide a concise yet detailed explanation for the assigned classification.
- Justify the classification by referencing:
 - Template structure.
 - Terminology used in the statement.
 - Specific conditions or context highlighted by the statement.
 - Explicitly map the elements of the statement (e.g., terms, qualifying clauses, verb phrases, conditional clauses, etc.) to template components.

{rule_type} Rule subtypes

{classification_text}

Definitions

- **attribute term**: A term that signifies a non-Boolean property of an entity class (or object class).
- **role term**: A term that signifies the role played by one of the participating parties or objects in a relationship: for example, employer and employee are role terms (with respect to the relationship whereby an organization employs a person), whereas organization and person are not role terms.
- **category attribute term**: A term is usually admin-defined, with some external inputs. They have unique labels (e.g., 'Cash') and may use internal codes. Boolean attributes indicate "Yes" or "No" responses, shown as checkboxes or "Y/N" fields.
- **quantitative attribute**: An attribute on which some arithmetic can be performed (e.g., addition, subtraction) and on which comparisons other than "=" and "<>" can be performed.
- **qualifying clause**: refines a rule's scope or specificity by limiting the subject or other terms to particular subsets or conditions (e.g., "for a return journey" or "that is current").

Output Format:

Each analyzed statement must be provided in JSON format. The structure for each statement is as follows:

```

```json
{
 "doc_id": "The Document ID from the input",
 "statement_id": "The original statement ID",
 "statement_title": "The original statement_title",
 "statement_text": "The original statement_text",
 "statement_sources": "The original statement_sources",
 "classification": [
 {
 "subtype": "Assigned rule subtype, use the title of the section/subsection (e.g., Activity time limit rules)",
 "templates_ids": ["Template ID that matched the statement."],
 "confidence": Confidence Score (0-1),
 "explanation": "Detailed explanation of why this classification was assigned."
 }
]
}

```
---


## Example Output:


```
[{
 "doc_id": "some doc id",
 "statement_id": "some id",
 "statement_title": "some title",
 "statement_text": "some text",
 "statement_sources": ["some source"],
 "classification": [
 {
 "subtype": "Some Subtype Title",
 "templates_ids": ["T123", "T456"],
 "confidence": 0.9,
 "explanation": "This statement ..."
 },
 {
 "subtype": "Another Subtype Title",
 ...
 }
]
}
]
```

```

```

"templates_ids": ["T789"],
"confidence": 0.4,
"explanation": "There are elements ..."
}
]
},
{
"doc_id": "another doc id",
"statement_id": "another id",
"statement_title": "another title",
"statement_text": "another text",
"statement_sources": ["another source"],
"classification": [
{
"subtype": "Subtype Title",
"templates_ids": ["T123"],
"confidence": 0.7,
"explanation": "The clause dictates ..."
}
]
},
...
]
```

Additional Notes:

- Multiple Classifications:

- A statement can have multiple classifications if it aligns with different subtypes. Justify each with appropriate confidence levels.

- Cross-References:

- When a statement refers to another section (e.g., "(a)(1)"), incorporate the referenced section if it is provided or available. If unavailable, indicate this in the explanation and lower the confidence score.

```
---
```

Fonte: Elaborado pelo autor.

A classificação dos termos, nomes, fatos e regras é realizada individualmente. Para cada tipo de elemento é criado um *prompt* customizado pelas variáveis tipo da regra (*Definitional* ou *Operative*), tipo do elemento (regras operativas, termos, nomes ou fatos) e a quantidade de elementos a serem processados. O LLM parece ter um problema para lidar com listas longas, como discutido na seção “Discussão”, o que faz com que elementos na lista não sejam processados. Para lidar com esse problema, os *prompts* foram divididos em lotes de cinco elementos, isso aumentou o número de chamadas para o LLM, mas garantiu a resposta para todos os elementos. O *prompt* é ilustrado no Quadro 17, ele usa alguns exemplos (*few-shots*) para auxiliar o LLM a identificar a classificação mais próxima para os enunciados do tipo de regra. Esse a função recebe os enunciados e os classifica de acordo com a taxonomia de Witt (2012). A taxonomia completa é apresentada no “APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO”.

Para todos os *prompts*, foi solicitado ao LLM que classificasse de acordo com um grau de confiança variando de zero a um e justificasse as suas escolhas. Os resultados desses algoritmos são discutidos no capítulo “7 VALIDAÇÃO” e serão utilizados pelo algoritmo de transformação. A implementação dessa função pode ser encontrada no repositório de código no arquivo “code/src/chap_6_semantic_annotation_rules_classification.ipynb”.

Discussão

Figura 18 - Pseudocódigo do *prompt* para extração de termos – primeira abordagem

```

1 // Summarize the document
2 summary = summarize_document(document_text)
3
4 // Extract the section number if available
5 section_number = extract_section_number(document_text)
6
7 // Split the document into sentences or paragraphs
8 sentences = split_into_sentences(document_text)
9
10 for sentence in sentences:
11     // Classify the expression
12     classification = classify_expression(sentence)
13
14     // Extract the terms
15     terms_list = extract_terms(sentence)
16
17     // Classify the terms and extract their definitions
18     for term_text in terms_list:
19         term_classification = classify_term(term_text)
20         definition = search_definition(term_text, document_text, summary)
21         terms.append(term_text, term_classification, definition)
22
23     // Extract the verb symbols
24     verb_symbols = extract_verb_symbols(sentence)
25
26     // Extract the source in the document
27     source = find_source(sentence, document_text)
28
29     // Add the element to the list
30     elements.append(element_id, expression, terms, verb_symbols, classification, source)
31
32     element_id += 1
33
34 // Add relationships between terms
35 for term1, term2 in combinations(terms, 2):
36     // Check if the terms are in the document
37     if term1 in document_text and term2 in document_text:
38         // Check if the terms are synonyms
39         if are_synonyms(term1, term2, document_text):
40             // Add the relationship
41             terms_relationships.append([term1, term2], 'Synonym')
42
43 // Output the JSON structure
44 return format_json(section_number, elements, terms_relationships)

```

Fonte: Elaborado pelo autor.

No início desta seção foi apresentada duas alternativas para a implementação do algoritmo de extração e classificação dos elementos (“extract / classify elements”) e a melhora no seu resultado quando os algoritmos foram divididos e simplificados. Uma possibilidade para explicar essa melhora é a dificuldade do LLM em lidar com instruções mais longas e complexas, fenômeno que Liu *et al.* (2023) denominaram de “perdido no meio” (*Lost in the Middle*). Os autores investigaram como os LLMs processam contextos longos, avaliando seu desempenho em tarefas de perguntas e recuperação de pares chave-valor. Um

dos achados foi a degradação do desempenho dos modelos quando as informações relevantes estão no meio do contexto de entrada, devido ao efeito de primazia e recência, que beneficia dados localizados no início ou no fim do contexto. Esse efeito é mais intenso em tarefas complexas que exigem raciocínio sobre múltiplos documentos e integração de informações dispersas. Já tarefas simples, como recuperação de pares chave-valor, apresentam menor impacto, ainda que perceptível em contextos longos.

Estudos como os de He Q. (2024) e Sun *et al.* (2024) indicam que os LLMs tendem a cometer mais erros à medida que aumenta a complexidade das instruções. Essa complexidade refere-se a instruções que envolvem múltiplas camadas de restrições, como formatação específica, condições numéricas e diretrizes detalhadas que o modelo deve seguir. Instruções assim frequentemente combinam diversas condições, tornando mais desafiador para os modelos aderirem rigorosamente aos requisitos. Futuramente, a correlação entre a complexidade das tarefas descritas por esses autores e métricas como a ciclométrica pode contribuir para aprimorar a formulação de prompts. Por exemplo, o prompt da primeira abordagem, representado em pseudocódigo na Figura 18, possui uma complexidade ciclomática (MCCABE, 1976) de seis, calculada pela fórmula ($v = e - n + p$), onde e (arestas) é 24, n (nós) é 19, e p (componentes conectados) é um.

A primeira abordagem apresenta um conjunto maior de instruções, focado em percorrer os parágrafos e, para cada palavra encontrada, voltar aos parágrafos em busca de definições. O *prompt* e o texto da seção somam aproximadamente 1.600 *tokens*, com complexidade ciclomática seis. Na segunda abordagem as instruções foram divididas em dois *prompts*, no primeiro, os textos têm cerca de 1.300 *tokens*, e as instruções tratam apenas dos parágrafos, com complexidade ciclomática cinco. O segundo totaliza 800 *tokens* e suas instruções percorrem os parágrafos para cada termo em uma lista, com complexidade ciclomática de dois. Essa mudança no algoritmo, como discutida, trouxe consistência entre as execuções e aumento de corretude.

LLMs são modelos do tipo “caixa-preta”, e explicar seu funcionamento é um tema de várias pesquisas (BHATTACHARJEE *et al.*, 2024; GAT *et al.*, 2023). O

problema da “caixa-preta” está na falta de transparência, dificultando a compreensão de como as decisões são tomadas. Isso gera preocupações em áreas críticas como saúde e finanças, onde é necessário justificar decisões, identificar vieses, corrigir erros e garantir conformidade regulatória. A opacidade dos modelos pode perpetuar preconceitos, provocar desconfiança e dificultar a segurança e equidade, limitando seu uso em contextos que exigem transparência e responsabilidade (GAT *et al.*, 2023).

Ainda que quantidade de instruções e tamanho não sejam os únicos fatores, a complexidade da tarefa intensifica o problema “perdido no meio”. Segundo Liu *et al.* (2023) e He J. *et al.* (2023), tarefas complexas, como QA com múltiplos documentos, exigem que o modelo não apenas localize informações relevantes em contextos longos, mas também as sintetize e relacione coerentemente para gerar uma resposta correta. Isso demanda múltiplas etapas de raciocínio, tornando o modelo mais suscetível ao problema. Isso foi percebido ao solicitar ao modelo que executasse um conjunto de tarefas para uma lista longa de itens.

Esses fatores, mesmo que não conclusivos, sugerem que a redução de instruções é benéfica. Uma possibilidade para trabalhos futuro é o uso de sistemas multiagentes, como o Autogen da Microsoft⁵⁸, que auxiliam na redução da complexidade de cada *prompt* ao criar um conjunto de agentes, cada um com foco específico em um resultado.

Sistemas multiagentes baseados em LLM (LLM-MA) exploram a colaboração de agentes autônomos em tarefas complexas, com aplicação em áreas como desenvolvimento de software, robótica, economia e simulação de políticas. Cada agente tende a ser especializado, e a colaboração permite a solução de problemas complexos em contextos dinâmicos. Esses sistemas são vantajosos para tarefas que requerem coordenação, especialização de funções e soluções emergentes (GUO *et al.*, 2024). Contudo, segundo Guo *et al.* (2024), LLM-MA não é sempre a abordagem ideal para todas as tarefas, sendo intensivos em termos computacionais e podendo introduzir complexidade desnecessária em tarefas simples. Desafios como gerenciamento de comunicação, alucinações e

⁵⁸ <https://microsoft.github.io/autogen>. Acesso em: 15/mar./2025.

coordenação eficiente limitam o uso desses sistemas, enquanto tarefas que exigem alta precisão pode ser mais bem atendidas por modelos determinísticos.

Outro problema com essa abordagem é a inconsistência: o modelo pode dar respostas diferentes para consultas semanticamente equivalentes, comprometendo a confiança do usuário, especialmente em áreas de alto risco como saúde e finanças. Jiang *et al.* (2020) exploraram essa questão, notando que pequenas variações nas probabilidades de *tokens* durante a geração podem resultar em saídas divergentes. Wei *et al.* (2022) observaram que mesmo o CoT, ao organizar o pensamento em etapas intermediárias, pode levar modelos a gerar sequências de raciocínio aparentemente coerentes, mas incorretas. Esse comportamento é mais comum em modelos menores, enquanto em modelos maiores o CoT melhora a precisão dos passos intermediários. Segundo Opsahl-Ong *et al.* (2024), pequenas alterações no prompt afetam o resultado, pois os modelos são altamente sensíveis à estrutura e contexto das instruções. Mudanças mínimas podem levar o modelo a interpretar a tarefa de maneira diferente, impactando resultados em pipelines de múltiplas etapas. Ajustes fino são essenciais para alcançar o desempenho desejado em tarefas específicas.

Um exemplo desse comportamento foi a classificação da sentença: “*If the Secretary certifies that the Commission was served with process, pleadings, or other papers pursuant to paragraph (a)(1) of this section and forwarded these documents to a named party pursuant to paragraph (a)(2) of this section, this certification constitutes evidence of service upon that party.*” A primeira classificação (Quadro 16) teve baixa confiança, pois a sentença admite duas interpretações: regra de papéis ou regra de consistência de dados.

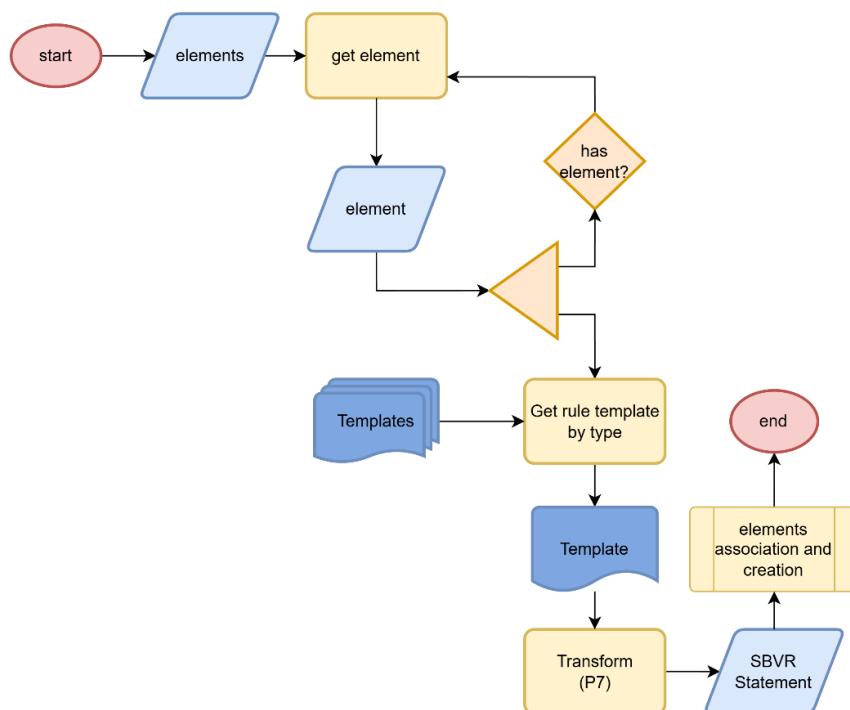
A regra de papéis foca em governança e responsabilidade, destacando os papéis dos indivíduos, como o Secretário, cuja certificação serve como evidência de serviço. Já a perspectiva de consistência de dados considera a relação entre elementos de dados, visando a consistência lógica entre eles, como a certificação e o encaminhamento de documentos. Essa classificação é mais adequada em contextos centrados em dados. Inicialmente, o *prompt* não especificava o foco, o que levou à escolha da perspectiva técnica e causou problemas na segunda classificação, na qual o LLM não conseguiu encontrar

uma categoria adequada na taxonomia. Com a adição da frase “*You are working for regulatory bodies, auditors, or process managers.*” a interpretação mudou para classificação de regra de papéis, aumentando a confiança e a consistência na classificação e classificou como “Regras de responsabilidade” na taxonomia (Quadro 17).

6.2.4 Transformação

Os algoritmos desta seção são responsáveis pela transformação das regras em NLP para SBVR e a persistência no KG CFR-SBVR, completando o processo CFR2SBVR.

Figura 19 - Algoritmo “Transformation”.



Fonte: Elaborado pelo autor.

Algoritmo “transformação para SBVR”

A Figura 19 ilustra o algoritmo “Transformation”, ele transforma as regras em NL, classificadas no processo de classificação, para SBVR-SE (P7). O processo consiste em transformar as declarações extraídas e classificadas pelos algoritmos anteriores usando os *templates* associados, de acordo com a taxonomia de Witt (2012).

Quadro 18 - Instruções para a transformação para SBVR.

Transform each given {element_name} {statement_name} into a structured format by matching it to the specified templates and subtemplates.

Steps

1. **Summarize {statement_name}:** Summarize the given {element_name} {statement_name} to understand its structure and content.
2. **Use Template:**
 - For given expression, use the templates and subtemplates ("Fact Type Form" if element_name in ["Fact Type", "Fact"] else "Rule Form") provided for transformation.
 - Determine the appropriate template or subtemplate based on the structure of the expression.
3. **Replace Placeholders:**
 - Substitute placeholders, such as '<term>', '<verb phrase>', '<conditional clause>', etc., with suitable values as per the expression.
 - For terms and names, the statement_id is the term defined by the statement.
4. **Include Qualifying Details:**
 - Where placeholders, such as '<qualifying clause>', require additional details (e.g., attributes or qualifiers to distinguish meaning), ensure that these are included appropriately as per the respective subtemplate.
5. **Transform into Structured Format:**
 - Once the transformation is complete, ensure it's in the correct template format.
6. **Output as Structured JSON:**
 - For every transformed expression generate a JSON object as per the specified output format.
7. **Review and Validate:**
 - Ensure accuracy in grammar and compliance with logical constructs when performing substitutions.
 - Ensure the generated JSON is in the correct template format.
8. **Assess the Transformation:**
 - Record the confidence level and reason for the confidence score in the JSON object.

{rule_template_formulation}

```

# Provided templates and subtemplates for transformation

{rule_templates_subtemplates}

# Output Format

[
  {
    "doc_id": <doc_id>,
    "statement_id": <statement_id or signifier>,
    "statement_title": <statement_title>,
    "sources": [<source>],
    "statement": <statement or definition>,
    "templates_ids": [<templates_id>],
    "transformed": <transformed_statement>,
    "confidence": <confidence_level>,
    "reason": <reason_for_confidence>
  },
  ...
]

```

- **`doc_id`**: A original identifier of the given document.

- **`statement_id or signifier`**: The original identifier of the given {statement_name}. e.g., '1', 'Person'.

- **`statement_title`**: The title of the given {statement_name}.

- **`sources`**: The original sources of the given {statement_name}.

- **`statement or definition`**: The original text of the given {statement_name}.

- **`templates_ids`**: The template(s) used for the transformation (e.g., T1, T2, etc.)

- **`transformed`**: The transformed statement according to template.

- **`confidence`**: The confidence level of the transformation range from 0 to 1.

- **`reason`**: The reason for the confidence score.

Notes

- Use only the provided templates and subtemplates for transformation.
- If a placeholder within an expression is not applicable or optional, consider whether it should be omitted or replaced by a suitable value.
- Each expression may involve nested levels of substitution as indicated by the subtemplate hierarchy (e.g., a qualifying clause that contains sub-elements).

- Ensure accuracy in grammar and compliance with logical constructs when performing substitutions.

Fonte: Elaborado pelo autor.

O objetivo das instruções do *prompt* no Quadro 18 é transformar as declarações utilizando os *templates* e *subtemplates* obtidos na etapa de classificação. O processo resume a declaração (*{statement_name}*) para identificar sua estrutura e conteúdo, seleciona o *template* associado ao elemento e substitui os *placeholders* por valores adequados (como termos, frases verbais e cláusulas qualificadoras) e gera uma versão transformada da declaração. Por fim, o modelo registra os níveis de confiança e justificativas para a transformação.

Quadro 19 - Exemplo de instruções para utilização do *template* na transformação.

```
## T7

### Subtemplate(s) in use
- S14: qualifying clause

### Rule Form

{A|An} <term 1>
{of {a|an} <term 2>} |
is by definition
{a|an|the} <term 3>
<qualifying clause>.

### Fact Type Form

<term 1> is a category of <term 3>
if "of {a|an} <term 2>" is used: <term 1> is of <term 2>

### Explanation

This template has the following important options and placeholders:
1. <term 1> is the term being defined;
2. "of {a|an} <term 2>" is only required if <term 1> is an attribute term;
3. <term 3> is the hypernym of <term 1>;
```

4. <qualifying clause> is used to state the characteristics that distinguish members of the set signified by <term 1> from members of the sets signified by other hyponyms of <term3>; the options for a <qualifying clause> are listed in subtemplate S14.

S14: qualifying clause

Subtemplate(s) in use

- S15: and-qualifying clause
- S9: verb part
- S13: conditional clause
- S18: either-or-qualifying clause
- S11: object
- S16: both-and-qualifying clause
- S17: or-qualifying clause

Form

```
<qualifying clause> ::=  
{ {that | who} <verb phrase> {<object>| } } |  
<verb part> <object> |  
other than {<object>| [<object>, or]} |  
{ <preposition> {which|whom}|whose } <conditional clause> |  
{ that | who } <verb phrase> { that | if | whether } <conditional clause> |  
<and-qualifying clause> |  
<or-qualifying clause> |  
<both-and-qualifying clause> |  
<either-or-qualifying clause> }
```

Explanation

...

Fonte: Adaptado de Witt (2012).

Esse algoritmo depende do *template* obtido na etapa anterior, ele é usado para suprir o modelo com as instruções específicas na taxonomia de Witt (2012). Os *templates* são apresentados para o modelo como exemplificado no Quadro 19. O *template* principal (T7) define a relação entre termos, categorias e atributos, descrevendo fatos e tipos de fatos com cláusulas qualificadoras. Essas

cláusulas, modeladas pelo *subtemplate* S14, especificam características adicionais ou condições, com formas variadas, como operadores lógicos (“and”, “or”, “both”) e cláusulas condicionais. Outros *subtemplates* complementares, como S5 (funções de conjunto) e S10 (predicados), adicionam variações ao *template*. Os placeholders, como “<term>”, “<qualifying clause>” e “<verb phrase>”, são substituídos por elementos contextuais seguindo regras gramaticais e sintáticas definidas.

Quadro 20 - Instruções para uso dos templates.

How to interpret the templates and subtemplates

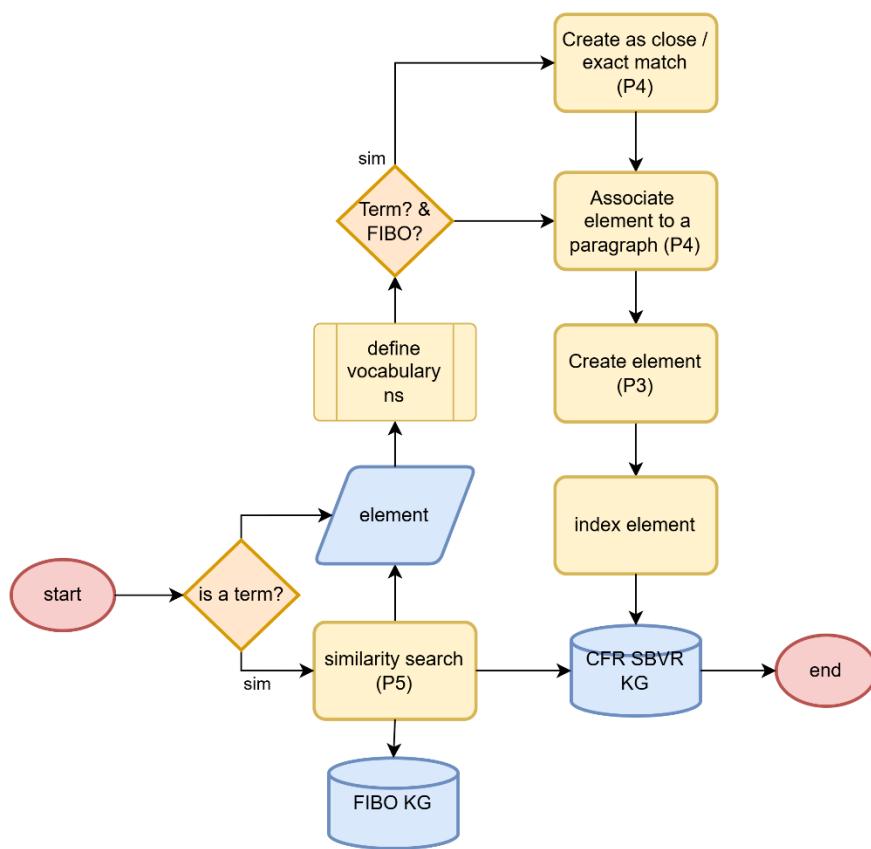
Each formulation is expressed using a template, in which the various symbols have the following meanings:

1. Each item enclosed in “angle brackets” (“<” and “>”) is a placeholder, in place of which any suitable text may be substituted. For example, any of the following may be substituted in place of <operative rule statement subject> (subtemplate):
 - a. a term: for example, “flight booking request”,
 - b. a term followed by a qualifying clause: for example, “flight booking request for a one-way journey”,
 - c. a reference to a combination of items: for example, “combination of enrollment date and graduation date”, with or without a qualifying clause,
 - d. a reference to a set of items: for example, “set of passengers”, with or without a qualifying clause.
2. Each pair of braces (“{” and “}”) encloses a set of options (separated from each other by the bar symbol: “|”), one of which is included in the rule statement. For example,
3. If a pair of braces includes a bar symbol immediately before the closing brace, the null option is allowed: that is, you can, if necessary, include none of the options at that point in the rule statement.
4. Sets of options may be nested. For example, in each of the templates above
 - a. a conditional clause may be included or omitted,
 - b. if included, the conditional clause should be preceded by either “if” or “unless”.
5. A further notation, introduced later in this section, uses square brackets to indicate that a syntactic element may be repeated indefinitely.
6. Any text not enclosed in either “angle brackets” or braces (i.e., “must”, “not”, “may”, and “only”) is included in every rule statement conforming to the relevant template.

Fonte: Adaptado de Witt (2012).

No Quadro 20 é apresentado para o modelo como interpretar os *templates* e *subtemplates* usados na formulação de declarações estruturadas, destacando o significado de suas notações. *Placeholders* entre “<>” representam elementos substituíveis por termos, cláusulas qualificadoras ou referências a conjuntos. Opções entre “{}” oferecem alternativas, e uma barra (“|”) antes do “}” indica que vazío é uma opção válida. Opções podem ser aninhadas, e colchetes “[]” indicam elementos repetíveis. Texto fora dessas notações, como “must” e “may,” é fixo em todas as declarações baseadas no *template*.

Figura 20 - Algoritmo “elements association and creation”



Fonte: Elaborado pelo autor.

A verbosidade dos *prompts*, por causa das instruções detalhadas, causa um aumento no número de *tokens* e por consequência o custo. Esse cenário pode ser melhorado se o modelo for afinado (*fine-tunning*) com as instruções, porém, é necessário um conjunto mínimo de dados para o treinamento, o que não estava disponível na realização desta pesquisa, podendo ser explorado em trabalhos

futuros. O algoritmo completo está no repositório de código na pasta “src/code/labs/chap_6_nlp2sbvr_transform.ipynb”.

Algoritmo “elements association and creation”

O algoritmo “elements association and creation”, apresentado na Figura 20, é responsável pela criação e persistência, no KG CFR-SBVR, das associações entre os elementos e pelo mapeamento dos dados e metadados capturados pelos algoritmos anteriores para a ontologia SBVR adaptada.

Quadro 21 - Triplas para regras de definição.

```
cfr-sbvr:NonresidentGeneralPartner27502 a sbvr:GeneralConcept,
sbvr:IntensionalDefinition,
sbvr:Term ;
sbvr:signifier "Non-resident general partner" ;
sbvr:closeMatch
<https://spec.edmcouncil.org/fibo/ontology/BE/Partnerships/Partnerships/GeneralPartner> ;
sbvr:closeMatch
<https://spec.edmcouncil.org/fibo/ontology/BE/Partnerships/Partnerships/GeneralPartnership> ;
sbvr:referenceSupportsMeaning "§ 275.0-2(a)(1)" ;
sbvr:referenceSupportsMeaning "§ 275.0-2(a)" ;
sbvr:isImplicitlyUnderstood "false"^^xsd:boolean ;
sbvr:statement "A non-resident general partner is by definition a general partner who resides in a place not subject to the jurisdiction of the United States." ;
sbvr:designationIsInNamespace cfr-sbvr:CFR_SBVR_275_0_2_NS ;
cfr-sbvr:extractOriginalStatement "A general partner who resides in a place not subject to the jurisdiction of the United States." ;
cfr-sbvr:transformationSemscore 0.8474960390031319 ;
cfr-sbvr:transformationSimilarityScore 0.95 ;
cfr-sbvr:transformationSimilarityScoreConfidence 0.9 ;
cfr-sbvr:transformationAccuracy 0.95 ;
cfr-sbvr:transformationGrammarSyntaxAccuracy 1.0 ;
cfr-sbvr:transformationFinding "The transformed sentence accurately reflects the original definition with the structure provided by the template." ;
cfr-sbvr:transformationFinding "The grammar and syntax of the transformed sentence are correct." ;
cfr-sbvr:transformationFinding "The transformation maintains the meaning and context of the original definition." ;
cfr-sbvr:classificationTemplatesId "T7" ;
cfr-sbvr:classificationType "Definitional" ;
cfr-sbvr:classificationSubtype "Formal intensional definitions" ;
cfr-sbvr:classificationSubtypeConfidence 0.9 ;
cfr-sbvr:classificationSubtypeExplanation "The statement defines 'Non-resident general partner' using a hypernym 'general partner' and a qualifying clause 'who resides in a place'
```

not subject to the jurisdiction of the United States'. This aligns with the template for formal intensional definitions." ;

cfr-sbvr:createDate "2024-12-01T19:21:51Z"^^xsd:dateTime .

Fonte: Elaborado pelo autor.

O Quadro 21 ilustra uma tripla criada para o termo “Non residente general partner”. As triplas com prefixo sbvr seguem a ontologia adaptada para SBVR e as com prefixo cfr-sbvr são customizadas e tem o propósito de auxiliar no entendimento do processo de transformação, são elas:

- cfr-sbvr:extractOriginalStatement: Texto extraído do regulamento, fonte para a transformação;
- cfr-sbvr:transformationSemscore: Pontuação atribuída pelo método SemScore (AYNETDINOV; AKBIK, 2024) para avaliação de similaridade entre a sentença original (cfr-sbvr:extractOriginalStatement) e a transformada (sbvr:statement);
- cfr-sbvr:transformationSimilarityScore: Pontuação atribuída pelo método “LLM as a Judge” (WEI; CHEN; LUO, 2024, DONG; HU; COLLIER, 2024, ZHENG *et al.*, 2023) para avaliação de similaridade entre a sentença original (cfr-sbvr:extractOriginalStatement) e a transformada (sbvr:statement);
- cfr-sbvr:transformationSimilarityScoreConfidence: Pontuação atribuída pelo LLM em relação o quanto confiante foi a atribuição da pontuação “SimilarityScore”;
- cfr-sbvr:transformationAccuracy: Pontuação atribuída pelo LLM em relação o quanto próximo a sentença transformada está do *template* relacionado;
- cfr-sbvr:transformationGrammarSyntaxAccuracy: Pontuação de correção gramatical, indica se foram detectados erros de gramática na sentença transformada;
- cfr-sbvr:transformationFinding: Comentário do modelo sobre o processo de transformação. As justificativas do modelo ajudam a ajustar o processo;
- cfr-sbvr:classificationTemplateId: *Template* usado na transformação;

- cfr-sbvr:classificationType: Classificação no topo da ontologia de Witt (2012). Como já discutido, para termos, nomes e fatos é sempre “Definitional”;
- cfr-sbvr:classificationSubtype: Classificação segundo a ontologia de Witt (2012). Somente subclasses que possuem *templates* são consideradas;
- cfr-sbvr:classificationSubtypeConfidence: Pontuação da confidência do modelo na classificação;
- cfr-sbvr:classificationSubtypeExplanation: O motivo da escolha da subclasse pelo modelo;
- cfr-sbvr:createDate: Data e hora que a tripla foi atualizada no banco de dados.

O predicado “sbvr:exactMatch” está associado a busca semântica pelo termo no KG CFR_SBVR. O predicado “sbvr:closeMatch” aponta para termos que foram identificados na busca semântica no KG FIBO. A associação da designação com a origem do termo na CFR foi através do predicado “sbvr:referenceSupportsMeaning”, essa associação foi feita com literais que representam os parágrafos.

Quadro 22 - Tripla para regras operativas.

```
cfr-sbvr:CommissionOrdersAHearing a sbvr:Rule,
sbvr:BehavioralBusinessRule ;
cfr-sbvr:hasTerm cfr-sbvr:Commission ;
cfr-sbvr:hasTerm cfr-sbvr:Hearing ;
cfr-sbvr:hasTerm cfr-sbvr:PublicInterest ;
cfr-sbvr:hasTerm cfr-sbvr:ProtectionOfInvestors ;
cfr-sbvr:hasVerbSymbol cfr-sbvr:WillOrder ;
cfr-sbvr:hasVerbSymbol cfr-sbvr:Appears ;
cfr-sbvr:hasVerbSymbol cfr-sbvr:IsNecessary ;
cfr-sbvr:hasVerbSymbol cfr-sbvr:IsAppropriate ;
sbvr:isImplicitlyUnderstood "false"^^xsd:boolean ;
sbvr:statement "The Commission may order a hearing on the matter only if it appears that a hearing is necessary or appropriate in the public interest or for the protection of investors." ;
sbvr:designationIsInNamespace cfr-sbvr:CFR_SBVR_275_0_5_NS ;
```

```

cfr-sbvr:extractOriginalStatement "The Commission will order a hearing on the matter, if it appears that a hearing is necessary or appropriate in the public interest or for the protection of investors." ;
cfr-sbvr:transformationSemscore 0.9432317012278398 ;
cfr-sbvr:transformationSimilarityScore 0.95 ;
cfr-sbvr:transformationSimilarityScoreConfidence 0.9 ;
cfr-sbvr:transformationAccuracy 0.95 ;
cfr-sbvr:transformationGrammarSyntaxAccuracy 1.0 ;
cfr-sbvr:classificationType "Activity rules" ;
cfr-sbvr:classificationSubtype "Activity pre-condition rules" ;cfr-sbvr:classificationTypeConfidence 0.9 ;
cfr-sbvr:classificationTypeExplanation "The statement describes an activity (ordering a hearing) that the Commission must perform under certain conditions, which is characteristic of an activity rule." ;
cfr-sbvr:classificationSubtypeConfidence 0.8 ;
cfr-sbvr:classificationSubtypeExplanation "The statement describes a condition under which the Commission will order a hearing. It specifies that a hearing will be ordered if it appears necessary or appropriate in the public interest or for the protection of investors. This aligns with the template for Activity pre-condition rules, which involves an activity that may occur only if a certain condition is met. The 'if' clause in the statement matches the 'if <conditional clause>' part of the template." ;
cfr-sbvr:createDate "2024-12-01T20:14:32Z"^^xsd:dateTime .

```

Fonte: Elaborado pelo autor.

O Quadro 22 ilustra uma tripla criada para uma regra operativa, os predicados são comuns ao das regras de definição, com exceção para classificação de tipos, que segundo a ontologia de Witt (2012) pode ser atividade, dados ou papel e a pontuação da confiança associada a essa classificação.

O algoritmo completo está disponível no repositório de códigos⁵⁹. Após a execução, 16 termos e 2 nomes foram criados, totalizando 247 triplas, esses números variam de execução para execução e são discutidos no capítulo “7 VALIDAÇÃO”.

Quadro 23 - Consulta SPARQL para recuperar referências.

```

PREFIX cfr-sbvr:<http://cfr2sbvr.com/cfr#>
PREFIX sbvr: <https://www.omg.org/spec/SBVR/20190601#>

```

⁵⁹ Arquivo jupyter notebook src/2_b_concept_annotation.ipynb.

```

select ?designation ?reference
FROM NAMED cfr-sbvr:CFR_SBVR
{
graph ?g {
?designation sbvr:referenceSupportsMeaning ?reference .
}
}

```

Fonte: Elaborado pelo autor.

Embora o KG CFR contenha triplas que representam os parágrafos, o parágrafo não pôde ser endereçado diretamente pelo parágrafo por causa dos problemas já discutidos na seção “6.2.1 KG para CFR e FIBO”, uma busca por similaridade identificou o parágrafo com o significante, mas sua identificação (exemplo 275.0-7(b)(2)) estava fragmentada e não correspondia a encontrada no texto original da CFR. O contorno foi associar a literais que podem ser interpretadas por SMEs e usadas em uma busca simples. Todas as referências aos termos podem ser recuperadas com uma consulta SPARQL, como ilustrada no Quadro 23.

Quadro 24 - Pesquisa por elementos sem definição.

```

PREFIX sbvr: <https://www.omg.org/spec/SBVR/20190601#>
SELECT ?s { ?s sbvr:statement "missing" . }

```

Fonte: Elaborado pelo autor.

Algumas definições de termos não estão presentes nos textos das seções e, portanto, não foram identificadas pelo LLM, para indicar uma definição ausente, o predicato “sbvr:statement” foi associado a literal “missing”. Todos os termos sem definição podem ser encontrados pela pesquisa SPARQL do Quadro 24.

Algoritmo “similarity search”

O algoritmo “similarity search”, parte do algoritmo “elements association and creation”, busca nos KG FIBO e CFR_SBVR por termos e nomes, com o objetivo de estabelecer uma ligação entre os elementos e sua definição em outros vocabulários. Caso o elemento não seja localizado, o algoritmo cria triplas com o significante do elemento, tipo, associação com o *namespace* do vocabulário, associação com a origem (seção / parágrafo) e sua definição, mesmo que a definição seja a palavra-chave “missing”. Neste caso, uma query SPARQL

poderá localizar todas as triplas que estão associadas a essa literal e identificar todos os elementos que necessitam de uma definição. Essa abordagem é semelhante ao que Haarst (2013) chama de “colcha de retalhos” (do inglês *pathwork*), na qual o livro de regras SBVR é completado gradualmente, permitindo criar um livro de regras completo (OMG, 2019), no sentido que todos os elementos estão presentes, mas ainda necessitando de uma definição.

Para a busca, o banco de dados AllegroGraph utiliza o algoritmo de busca por similaridade Approximate Nearest Neighbor (ANN)⁶⁰, mais especificamente o Approximate Nearest Neighbors Oh Yeah (Annoy)⁶¹. Esse algoritmo é baseado em projeções aleatórias e árvores, desenvolvido por Erik Bernhardsson em 2015 enquanto trabalhava no Spotify. Ele é otimizado para buscas em conjuntos de dados com até 1000 dimensões densas. O processo de busca de vizinhos mais próximos envolve dividir o conjunto de pontos ao meio repetidamente, até que cada subconjunto contenha k elementos, onde k geralmente é cerca de 100. Esse método equilibra eficiência e precisão em grandes volumes de dados.

Convenientemente o AllegroGraph integra a busca por similaridade as pesquisas SPARQL com o predicado “llm:nearestNeighbor”. As pesquisas seguem a sintaxe: “(?uri ?score ?originalText ?pred ?type) llm:nearestNeighbor (?text ?vectorRepoSpec ?topN ?minScore ?selector)”. No predicado ‘llm:nearestNeighbor’, as variáveis à esquerda representam os resultados da busca: “?uri” é o identificador único do item retornado, “?score” indica a pontuação de similaridade, “?originalText” traz o texto original associado, “?pred” pode fornecer o predicado RDF relacionado, e “?type” indica o tipo de dado do item. À direita, as variáveis de entrada incluem “?text” (o texto de consulta), “?vectorRepoSpec” (especifica o repositório vetorial), “?topN” (quantidade de vizinhos mais próximos), “?minScore” (pontuação mínima) e “?selector” (um filtro opcional).

⁶⁰ <https://thenewstack.io/allegrograph-8-0-incorporates-neuro-symbolic-ai-a-pathway-to-agi/>. Acesso em: 15/mar./2025.

⁶¹ Repositório de código do ANNOY - <https://github.com/spotify/annoy>. Acesso em: 15/mar./2025.

Os termos FIBO e SBVR foram transformados em vetores utilizando o modelo text-embedding-3-small da OpenAI⁶². Foram indexados:

- FiBO - predicado *label* (<http://www.w3.org/2000/01/rdf-schema#label>)
- FIBO - tipos *Class* e *NamedIndividual*
(<http://www.w3.org/2002/07/owl#Class>,
(<http://www.w3.org/2002/07/owl#NamedIndividual>)).
- SBVR - predicado *signifier*
(<https://www.omg.org/spec/SBVR/20190601#signifier>)
- SBVR – tipos *Term*, *Name*
(<https://www.omg.org/spec/SBVR/20190601#Term>,
<https://www.omg.org/spec/SBVR/20190601#Name>)

Quadro 25 - Busca por similaridade no KG FIBO

```
PREFIX fibo: <https://spec.edmcouncil.org/fibo/ontology/master/2024Q2/QuickFIBOProd#>
PREFIX llm: <http://franz.com/ns/allegrograph/8.0.0/llm/>

SELECT ?g ?s ?p ?originalText ?definition ?score ?type
FROM NAMED fibo:FIBO_Graph
WHERE {
  (?uri ?score ?originalText ?p ?type) llm:nearestNeighbor ("asset" "fibo-glossary-3m-vec" 5 0.8)
  .
  graph ?g {
    ?s a ?type .
    ?s ?p ?originalText ;
      skos:definition ?definition .
  }
}
```

Fonte: Elaborado pelo autor.

A busca por similaridade é realizada pela consulta SPARQL ilustrada no Quadro 25. Ela utiliza prefixos para definir os *namespaces* da ontologia FIBO e das funções relacionadas a função no AllegroGraph. A cláusula SELECT especifica

⁶² <https://platform.openai.com/docs/guides/embeddings/embedding-models>. Acesso em: 15/mar./2025.

as variáveis que serão retornadas: o grafo onde o recurso foi encontrado (?g), o sujeito que corresponde ao resultado (?s), a propriedade que relaciona o sujeito ao texto original (?p), o texto original associado ao significante (?originalText), a definição do significante (?definition), a pontuação de similaridade semântica (?score) e o tipo do conceito (?type).

A cláusula “FROM NAMED” indica que os dados devem ser buscados no grafo nomeado “fibo:FIBO_Graph”. Na cláusula “WHERE”, a função “lilm:nearestNeighbor” é utilizada para realizar a busca semântica. Os parâmetros dessa função incluem a palavra procurada “asset”, o banco de dados vetorial “fibo-glossary-3m-vec”, o limite máximo de resultados (5) e a pontuação mínima de similaridade (0.8). As variáveis de saída da função correspondem ao URI do recurso encontrado, a pontuação de similaridade, o texto associado ao recurso que foi comparado semanticamente, a propriedade que relaciona o recurso ao texto original e o tipo do recurso.

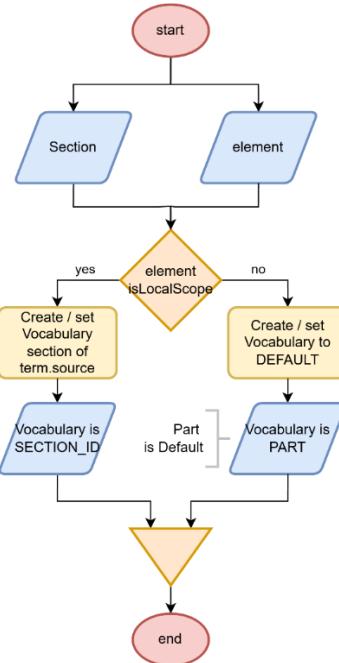
Na cláusula “GRAPH”, são especificadas as condições que devem ser buscadas dentro do grafo “?g”. O sujeito “?s” é do tipo “?type”, possui uma propriedade “?p” com o valor “?originalText” e tem uma definição associada através da propriedade “skos:definition”.

Todos os “labels” foram vetorizados com 1.536 dimensões, se aplica também para os criados pelo algoritmo “Create entity” que é indexado como vetor pelo algoritmo “tokenization”. Os algoritmos “Create as close / exact match”, “Create entity” e “Associate entity to paragraph” são responsáveis por criar os nós no KG SBVR utilizando a ontologia SBVR e as triplas associadas aos parágrafos.

Algoritmo “define vocabulary namespace”

O *namespace* do vocabulário é utilizado para agrupar as triplas das entidades e regras (elementos) em vocabulários evitando eventuais conflitos de definições, considerando que a CFR estabelece que algumas definições são válidas para um escopo específico (exemplo: seção), não podendo ser extrapolado para outras seções do regulamento.

Figura 21 - Algoritmo “define vocabulary namespace”



Fonte: Elaborado pelo autor.

O algoritmo “define vocabulary namespace”, ilustrado na Figura 21, é responsável definir é criar os vocabulários e os *namespaces* para os vocabulários. Esse algoritmo é determinístico, ele recebe uma seção e um indicador se o escopo da definição é local, se sim, cria o vocabulário e o *namespace* e retorna qual o *namespace* que deve ser associado ao elemento usando o predicado “sbvr:designationIsInNamespace”. Embora determinístico, ele depende da identificação do escopo da definição, essa extração é realizada pelo algoritmo “extract / classify elements” e a avaliação da extração da origem foram pontuações de 100% para precisão, acurácia e F1. A implementação dessa função pode ser encontrada no repositório de código no arquivo “code/src/chap_6_semantic_annotation_rules_classification.ipynb”.

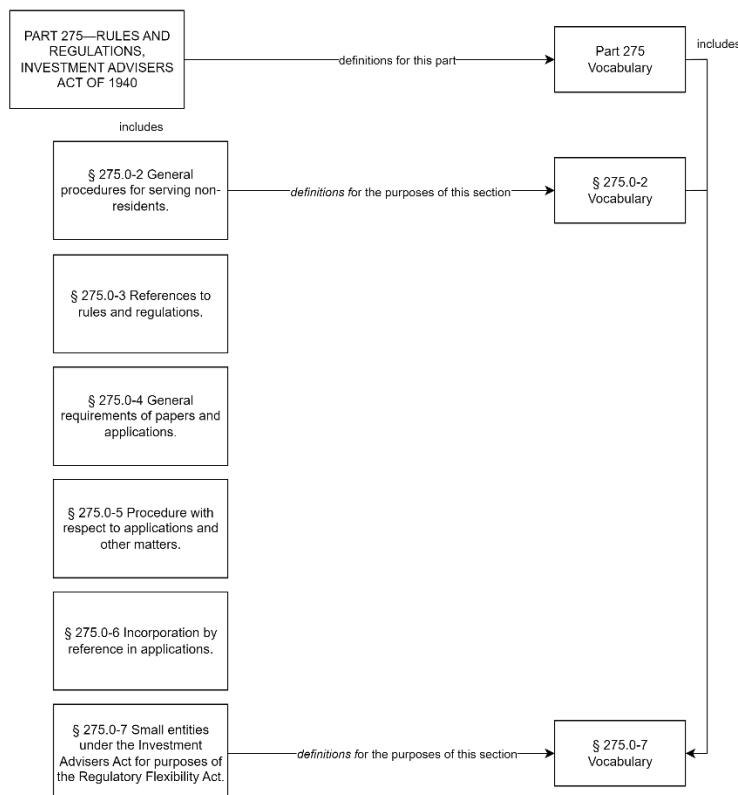
6.2.5 Catálogo de regras

O conteúdo completo de todos os KGs, está armazenado no repositório de códigos na forma de RDF N-Quads⁶³. Esse formato RDF permite armazenar as triplas e o nome do grafo, como quarto elemento, no formato:

⁶³ <https://www.w3.org/TR/n-quads/>. Acesso em: 15/mar./2025.

“<<http://one.example/subject1>> <<http://one.example/predicate1>> <<http://one.example/object1>> <<http://example.org/graph3>> .”. A extração foi realizada com o utilitário “agtool” e o script de extração está no repositório⁶⁴.

Figura 22 - Fragmento da organização dos *namespaces* para os vocabulários da CFR



Fonte: Elaborado pelo autor.

O catálogo de regras criado pelo processo de transformação está associado ao prefixo “cfr-sbvr:<<http://cfr2sbvr.com/cfr#>>”, representado pelo nó “cfr-sbvr:SBVR_CFR_NS” e associado as designações pelo predicado “sbvr:designationIsInNamespace”. O script “create_vocabulary_namespaces.py” é responsável por criar os nós para os vocabulários e *namespaces* padrões para: FIBO_NS – associado a todos os termos importados da ontologia FIBO; CFR_Title_17_Part_275_NS – associado a todas as seções importadas da ontologia FRO; e CFR_SBVR_NS – todos os elementos criados pela transformação CFR2SBVR. Os *namespaces* são associados pelo predicado “sbvr:vocabularyNamespacesDerivedFromVocabulary”.

⁶⁴ <scripts/export-cfr2sbvr-repo-to-nquad.sh>

Quadro 26 - Consulta SPARQL para recuperar vocabulários e *namespaces* dos vocabulários.

| | |
|---|--|
| PREFIX sbvr:
<https://www.omg.org/spec/SBVR/20190601#>
PREFIX skos:
<http://www.w3.org/2004/02/skos/core#>
PREFIX xsd:
<http://www.w3.org/2001/XMLSchema#>
PREFIX cfr-sbvr: <http://cfr2sbvr.com/cfr#>

SELECT ?s ?p ?o
{
?s ?p sbvr:Vocabulary .
}
Output
CFR_SBVR_275_0_7_VOC
CFR_SBVR_275_0_5_VOC
CFR_SBVR_275_0_2_VOC
CFR_SBVR_VOC
FIBO_VOC
CFR_Title_17_Part_275_VOC | PREFIX sbvr:
<https://www.omg.org/spec/SBVR/20190601#>
PREFIX skos:
<http://www.w3.org/2004/02/skos/core#>
PREFIX xsd:
<http://www.w3.org/2001/XMLSchema#>
PREFIX cfr-sbvr: <http://cfr2sbvr.com/cfr#>

SELECT ?s ?p ?o
{
?s ?p sbvr:VocabularyNamespace .
}
Output
CFR_SBVR_275_0_7_NS
CFR_SBVR_275_0_5_NS
CFR_SBVR_275_0_2_NS
CFR_SBVR_NS
FIBO_NS
CFR_Title_17_Part_275_NS |
|---|--|

Fonte: Elaborado pelo autor.

Cada seção pode conter elementos que são definidos para o escopo da seção, a forma para representar essa hierarquia foi criar um vocabulário para a seção. A Figura 22 ilustra o vocabulário CFR_SBVR_VOC (padrão), ele é associado a Part 275 (raíz) e os demais vocabulários são relacionados com ele pelo predicado “sbvr:vocabulary1IncorporatesVocabulary2” e “sbvr:namespace1IncorporatesNamespace2”. O Quadro 26 apresenta as pesquisas SPARQL para recuperar os vocabulários e os *namespaces* dos vocabulários.

Quadro 27 - Consulta SPARQL para recuperar o *namespace* do vocabulário associado a designação.

| |
|---|
| PREFIX sbvr: <https://www.omg.org/spec/SBVR/20190601#>
PREFIX cfr-sbvr: <http://cfr2sbvr.com/cfr#>
SELECT ?s ?o
FROM NAMED cfr-sbvr:CFR_SBVR |
|---|

```
{  
graph ?g {  
?s sbvr:designationIsInNamespace ?o .  
}  
}
```

Fonte: Elaborado pelo autor.

Os elementos são associados aos *namespaces* na sua criação pelo algoritmo de anotação semântica. Os elementos são criados com o prefixo “cfr-sbvr” e associados ao vocabulário pelo predicado “sbvr:designationIsInNamespace”, dessa forma todos os elementos são acessados pelo mesmo prefixo, mas podem ser filtrados por vocabulário utilizando o predicado, como ilustrado no Quadro 27.

7 VALIDAÇÃO

Esta seção estrutura os processos de validação, cada processo do CFR2SBVR foi validado por um processo de validação utilizando um conjunto de técnicas. A seção “7.1 Validação por similaridade” aborda os critérios de similaridade utilizados para avaliar a precisão na transformação de textos regulatórios em SBVR, detalhando como as correspondências entre as traduções automatizadas e manuais foram analisadas. A seção “7.2 Métodos aplicados ao experimento” apresenta os métodos experimentais aplicados, explicando o uso de métricas quantitativas e qualitativas para mensurar a eficácia do modelo. Em “7.3 Resultados”, são descritos os resultados obtidos.

O arcabouço experimental utilizado para validar o CFR2SBVR tem como objetivo principal avaliar a precisão, acurácia, consistência e repetibilidade dos algoritmos em relação aos valores de referência (conjunto de dados ouro), bem como a equivalência semântica da transformação quando comparada com as declarações originais. Secundariamente, o experimento visa avaliar a eficiência, em termos de tempo de processamento e custo monetário por execução. O arcabouço foi organizado da seguinte forma:

1. **Projeto do experimento:** Um conjunto de regulamentações do CFR da seção “5.2.1 Conjunto de dados” foi selecionado como conjunto de teste descritos na seção “5.3 Cenários de validação”. Um grupo de controle, com traduções manuais para SBVR, foi desenvolvido como padrão para comparação. As métricas de avaliação incluem precisão, acurácia, sensibilidade, F1-Score, tempo e custo monetário para o processamento.
2. **Execução:** A ferramenta CFR2SBVR processa os textos selecionados do CFR individualmente para evitar vazamento de dados, mantendo a operação sob condições computacionais consistentes. A saída de um processo é validada e substituída pela referência (padrão de dados ouro) para seguir como entrada para o próximo processo.
3. **Coleta e análise de dados:** A comparação entre as execuções automáticas e manuais, juntamente com os dados de tempo de processamento e quantidade de *tokens*, são coletados para cada cenário de teste.

4. **Interpretação dos resultados e documentação:** Os resultados são discutidos e eventuais ajustes são propostos. As técnicas, testes, resultados e configurações do sistema, são documentadas em relatórios (notebooks).
5. **Refinamento iterativo e discussão:** A partir dos resultados, ajustam-se os algoritmos e a configuração do sistema, realizando ciclos adicionais de testes para validar as modificações. A discussão aborda as limitações e desafios encontrados, propondo melhorias futuras.

Esse arcabouço foi adaptado de Chen *et al.* (2023), que apresentam um processo de avaliação abrangente para validar sua abordagem de modelagem de domínio automatizada usando LLMs, como GPT-3.5 e GPT-4. Inicialmente, preparam um conjunto de dados de dez exemplos diversos, cada uma consistindo em uma descrição textual e um modelo de domínio de referência. Utilizam várias técnicas de engenharia de *prompts* para instruir os LLMs a gerar modelos de domínio. Os modelos gerados são avaliados usando métricas de precisão, sensibilidade e F1-Score, e são submetidos a uma avaliação manual para categorizar cada elemento de modelagem, comparando-os com as soluções de referência. Realizam também uma análise estatística para comparar o desempenho de diferentes técnicas de engenharia de *prompts* e uma análise qualitativa para identificar padrões de falhas. A escolha dos exemplos visa equilibrar diversidade, complexidade e qualidade, proporcionando um ambiente de teste desafiador enquanto mantém a avaliação gerenciável, técnica semelhante foi adotada por Wyner e Peters (2011).

7.1 Validação por similaridade

As métricas tradicionais, como BLEU e ROUGE, frequentemente não se correlacionam bem com a avaliação humana devido à baixa sobreposição lexical com as respostas de referência. Essas métricas não conseguem distinguir adequadamente respostas corretas que utilizam sinônimos ou paráfrases, nem avaliam com precisão respostas para perguntas do tipo “sim/não” e aquelas que envolvem listas de entidades, frequentemente resultando em alta pontuação para respostas incorretas que compartilham muitas palavras com as respostas de referência. Além disso, BLEU e ROUGE não consideram a importância

relativa de termos específicos, levando a uma baixa correlação com o julgamento humano, especialmente em questões que exigem interpretação mais profunda ou precisão factual (YANG *et al.*, 2018).

Avanços recentes em LLMs demonstram um forte alinhamento com o julgamento humano (Gilardi *et al.*, 2023), sugerindo que os rótulos de preferência gerados por LLMs podem ser uma alternativa viável às anotações humanas. Aynetdinov e Akbik (2024) propõem a métrica SemScore para avaliação automatizada de LLMs. O SemScore mede a similaridade semântica entre as respostas geradas pelos modelos e as respostas-alvo, utilizando embeddings de frases e a similaridade de cosseno. Comparado a outras métricas tradicionais e modernas, como BLEU, ROUGE e BERTScore, o SemScore demonstrou uma correlação mais forte com as avaliações humanas, mostrando-se mais eficaz na avaliação da qualidade dos LLMs. Além disso, sua simplicidade conceitual facilita a aplicação prática e escalável, tornando-o uma ferramenta promissora para o cenário atual de rápido desenvolvimento de LLMs.

Wei, Chen e Luo (2024) propõe uma abordagem que envolve a realização de competições entre dois modelos LLM em “rodadas” de avaliação. Cada rodada consiste na apresentação de uma questão retirada de um novo *benchmark* chamado “Real-World Questions” (RWQ), que contém mais de 20 mil perguntas reais obtidas de fontes como Google Trends, Quora e outras plataformas. Esses dois LLMs respondem à mesma pergunta e, em seguida, um juiz (no caso, GPT-4) decide qual modelo apresentou a melhor resposta.

Shankar *et al.* (2024) apresenta a ferramenta EvalGen que auxilia a avaliação de saídas de LLMs ao alinhar os critérios de avaliação com as preferências humanas por meio de uma abordagem mista, combinando automatização e intervenção humana. Ele sugere critérios de avaliação e gera implementações candidatas, que podem ser funções em código ou prompts para LLMs. Enquanto as implementações são geradas, usuários avaliam manualmente um subconjunto das saídas, e este feedback é utilizado para ajustar e selecionar as melhores implementações. O EvalGen então fornece um relatório de alinhamento, mostrando o quanto bem os critérios selecionados se ajustam às preferências humanas, permitindo que o processo seja iterativo e ajustável,

reconhecendo a necessidade de redefinir critérios conforme as saídas dos LLMs são observadas e compreendidas.

He Q. *et al.* (2024) explora como melhorar a capacidade dos modelos de seguir instruções complexas que envolvem múltiplas restrições. Os autores identificam que seguir instruções com diversas exigências é crucial para aumentar a precisão e a confiabilidade desses modelos, mas tal capacidade ainda é pouco explorada. O estudo investiga quais tipos de dados de treinamento são mais eficazes para melhorar a habilidade dos modelos de seguir restrições complexas, mostrando que treinar com dados que contenham múltiplas restrições é mais eficaz do que com dados que abordam apenas restrições individuais.

Os autores propõem métodos para obter e utilizar esses dados de forma eficiente, como o uso de um modelo mais simples (chamado de modelo “estudante”) para gerar saídas iniciais e um modelo mais avançado (o “professor”) para corrigir essas saídas. Além disso, o artigo introduz uma abordagem contrastiva com aprendizado por reforço para otimizar o uso de exemplos positivos e negativos. Os experimentos realizados demonstram que esses métodos melhoram a capacidade dos modelos de seguir instruções em ambientes diversos, mantendo sua eficácia em diferentes domínios e configurações adversárias.

O estudo conclui que a utilização de dados compostos melhora a capacidade de generalização dos modelos, permitindo que eles lidem melhor com tarefas não vistas anteriormente. O trabalho também destaca a importância de dados de alta qualidade e métodos de correção baseados em discriminação, resultando em um desempenho superior em várias métricas de avaliação.

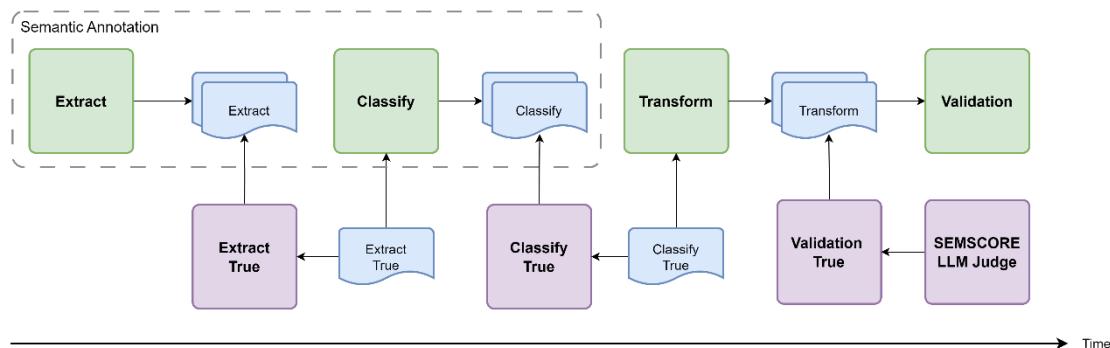
Dong, Hu e Collier (2024) exploram a capacidade dos LLMs de atuar como juízes personalizados, realizando julgamentos baseados nas preferências de usuários específicos através de perfis personalizados. A pesquisa mostrou que o método tradicional de personalização dos LLMs apresenta baixa confiabilidade, com concordância limitada em relação aos julgamentos humanos. Para melhorar essa precisão, os autores introduziram a estimativa de incerteza verbal, com a

qual o modelo calcula sua própria confiança para cada resposta em uma escala de 0 a 100. Esse valor permite identificar e filtrar respostas com baixa certeza, aumentando a concordância em amostras de alta confiança.

A metodologia inclui três configurações para os julgamentos: padrão, com incerteza verbal, e com a opção de empate. Experimentos foram conduzidos em diferentes conjuntos de dados (PRISM, OpinionQA, Empathetic Conversation e Reddit), que variam em complexidade e em detalhes dos perfis de usuários. Os resultados indicam que, com a filtragem por incerteza, os modelos atingem concordância de até 80% em amostras de alta confiança, superando o desempenho humano em alguns casos. A abordagem de incerteza verbal se mostra promissora para personalização em larga escala, ajudando a adaptar os modelos LLM às preferências dos usuários de forma mais confiável e escalável.

7.2 Métodos aplicados ao experimento

Figura 23 - Processo CFR2SBVR



Fonte: Elaborado pelo autor.

Nesta seção são discutidos os métodos selecionados e aplicados ao experimento para a validação. Os processos avaliados são apresentados na Figura 23 e a implementação discutida no capítulo “6.2 Implementação dos principais componentes”. Os documentos na Figura 23 são os *checkpoints*, eles são as saídas dos processos, exceto para os processos de validação, com extensão “True”, esses processos utilizam o *checkpoint* criado manualmente para a validação e constituem o conjunto de dados ouro.

A saída de um processo é normalmente utilizada como entrada para outro, porém, como discutido nesta seção, para minimizar erros, os processos “Classify” e “Transform” receberam o conjunto de dados ouro como entrada. O processo “Transform” foi validado pelo processo “Validation”. Esse processo utiliza o LLM para validar a saída do processo “Transform” ao invés de um conjunto de dados ouro.

Com algumas modificações, o processo CFR2SBVR poderia ser totalmente automático, incluir o *feedback* de SMEs ou combinar ambos. Por exemplo, no processo de extração e validação, bem como no uso de um conjunto de dados ouro, considere o conjunto E de termos, nomes, tipos de fatos e regras operativas que foi extraído: para cada elemento $e \in E$, atribui-se uma pontuação de confiabilidade $c(e) \in [0,1]$ pelo processo de validação, refletindo o grau de confiança no resultado produzido. Ao agrupar os elementos semelhantes por meio de uma função $\phi(e)$, a escolha do(s) elemento(s) com maior confiabilidade em cada grupo g pode ser obtida por meio do operador:

$$\arg \max_{e \in E: \phi(e)=g} c(e)$$

Dessa forma, quando diferentes instâncias do processo (ou *checkpoints*) atribuírem valores distintos de confiabilidade, a função retornará aquele(s) de maior pontuação, podendo indicar mais de um elemento em caso de empate.

A validação também pode ser conduzida pelos SMEs, tornando o processo semiautomático. Outra possibilidade é a combinação das abordagens, em que a validação automática seleciona os melhores resultados, e os SMEs determinam quais seguirão para as próximas etapas, substituindo, assim, o conjunto de dados ouro. Esse método se assemelha com as técnicas de aprendizagem por reforço com *feedback humano* (*Reinforcement Learning with Human Feedback* – RLHF). No entanto, devido ao alto custo e à limitação de dados, esta pesquisa não incluirá o ajuste fino do modelo com feedback, deixando essa possibilidade para investigações futuras, como discutido na seção “8.1 Discussão”.

7.2.1 Avaliação de similaridade de *strings*

Os termos e nomes podem ser identificados pelo seu significante, porém, regras operativas e fatos não tem um significante. Para avaliar a extração e a classificação do previsto com o conjunto de dados-ouro, foi comparado os enunciados usando o método de distância de Levenshtein. Esse método é amplamente usado para a comparação de *strings*, ele não considera a semântica, em vez disso, mede a similaridade ao calcular o número mínimo de operações necessárias para transformar uma *string* em outra (LIND-COMBS *et al.*, 2023; FAROUK, 2019). A comparação é empregada para verificar se um enunciado extraído pelo LLM corresponde ao mesmo enunciado presente no conjunto de dados padrão-ouro. Com isso, pequenas variações são ignoradas e o custo computacional para realizar milhares de comparações permanece baixo.

7.2.2 Avaliação de similaridade semântica

Os requisitos para avaliar a similaridade dos elementos transformados em relação ao texto original precisam considerar a semântica. Nesta pesquisa, essa avaliação será realizada por meio de dois métodos:

1. **SemScore**: este método utiliza um modelo de linguagem grande (LLM) para calcular *embeddings*, apresentando um custo computacional relativamente baixo em comparação com o método juiz.
2. **Avaliação por um LLM juiz**: conforme proposto por Wei, Chen e Luo (2024), Dong, Hu e Collier (2024) e Zheng *et al.* (2023), com critérios adaptados de Shankar *et al.* (2024). Esse método utiliza o LLM para analisar a estrutura e a semântica do enunciado, empregando suas capacidades de compreensão de linguagem natural (NLU). Embora tenha um custo computacional mais elevado, permite uma análise detalhada.

Quadro 28 - Prompt do *LLM as a judge*

| |
|--|
| # Task |
| You're an expert in judging sentence similarity and transformation using a template. |

These criteria should support the evaluation process by verifying classification accuracy, template application, and transformation fidelity.

Check the criteria and evaluate the output:

1. **Similarity Score**

- Given the statement or definition and transformed sentence (transformed), how similar are they from 0 to 1? And how confident are you about your estimation from 0 to 1?

2. **Transformation Accuracy**

- From 0 to 1, how does the transformed sentence (transformed) reflect the original sentence (statement or definition) with the structure and phrasing provided by the template and subtemplates?

3. **Grammar and Syntax Accuracy**

- How is the transformed sentence (transformed) grammatically correct and syntactically accurate from 0 to 1?

Output Format

Record your evaluation in JSON format as follows:

```
```json
{
 "doc_id": "<Document ID>",
 "statement_id": "<Statement ID>",
 "sources": ["<source>"],
 "similarity_score": <Similarity score>,
 "similarity_score_confidence": <Confidence score>,
 "transformation_accuracy": <Transformation score>,
 "grammar_syntax_accuracy": <Grammar score>,
 "findings": [<Things found during the evaluation and worth to be mentioned>,
 "<other things to mention>"],
 "semscore": <original semscore>
}
```

```
```
```
Input example
```

```
```json
{
  "doc_id": <Document ID>,
  "statement_id": <Statement ID>,
  "statement or definition": <original sentence>,
  "sources": [<source>],
  "terms": [
    {"term": <signifier>, "classification": <Proper or Common Noun>},
  ]}
```

```

...  

],  

"verb_symbols": <verbs or phrasal verbs>,  

"element_name": <name of element: Name, Term, Fact, Fact Type, Operative Rule>,  

"transformed": <transformed sentence>,  

"type": <type of element: Definitional, Activity, Party, Data>,  

"subtype": <subtype of element>,  

"templates_ids": ["T8"],  

"semscore": <semscore>  

}  

```
Templates and Subtemplates

{template}

```

Fonte: Elaborado pelo autor.

Devido à ausência de dados reais validados por especialistas, foram empregados métodos automatizados para indicar a similaridade semântica entre a transformação e o texto original. A comparação dos resultados dos dois métodos oferece *insights* sobre a qualidade da transformação e pode alertar os especialistas sobre possíveis discrepâncias. O processo de avaliação segue os seguintes passos:

1. Obter o elemento a ser avaliado do grafo SBVR;
2. Obter o parágrafo associado ao elemento: O parágrafo está relacionado ao elemento transformado no grafo pelo predicado "skos:definition". Esse predicado contém a literal com o texto extraído pelo LLM;
3. Aplicar SemScore;
4. Submeter para o LLM juiz;
5. Comparar os resultados das duas técnicas;
6. Similaridade igual ou superior a 90% foi considerada que o elemento passou nas avaliações.

O Quadro 28 apresenta o prompt usado para instruir o LLM a realizar a avaliação. As pontuações avaliadas são as de *similarity\_score*, de *transformation\_accuracy* e de *grammar\_syntax\_accuracy*. Ele também foi instruído a fornecer uma pontuação de confiança na transformação (*similarity\_score\_confidence*) e

escrever qualquer descoberta sobre a avaliação. O SemScore é obtido obtendo os vetores de cada declaração e calculando a distância do cosseno, embora o algoritmo seja determinístico, a obtenção dos vetores depende de um LLM e podem variar de execução para execução.

### 7.2.3 Avaliação da classificação da taxonomia de regras

Como já discutido nas seções anteriores, a falta de conjunto de dados para a validação dos algoritmos faz necessário empregar diferentes técnicas para alcançar um nível de confiabilidade aceitável. Para a função “*taxonomy classification and templates*”, responsável pela classificação do parágrafo usando a taxonomia definida por Witt (2012, p. 84-99) foi necessário criar um experimento isolado, usando os exemplos do autor como conjunto de dados-ouro e avaliar a eficiência do algoritmo em classificar esses exemplos. Esse experimento pode ser encontrado no repositório de código na pasta “src/code/labs/lab\_6.ipynb”.

Quadro 29 - *Prompt* para classificação do tipo da regra.

You are an expert in SBVR (Semantics of Business Vocabulary and Business Rules). You'll be asked to classify paragraphs into types of rules. You'll be given a definition of types of rules.

#### Definition

1. Definitional: Definitional rules constrain how we define constructs created or used by the organization or the industry within which it operates;
2. Party: Party rules restrict the parties who can perform processes or activities or play roles.
3. Data: Data rules constrain the data included in a transaction (a form or message) or a persistent data set;
4. Activity: Activity rules constrain the operation of one or more business processes or other activities;

Here's an example of the output in JSON:

```
[
 {
 "id": "some id",
 "text": "some text",
 "type": "Definitional"
 },
]
```

```
{
 "id": ...,
 "text": ...,
 "type": ...
},
]
```

Here's the paragraphs you'll need to classify:

Fonte: Elaborado pelo autor.

A função foi realizada em duas partes, na primeira os parágrafos são classificados no topo da ontologia, o *prompt* do Quadro 29 é utilizado para a classificação do tipo.

#### Quadro 30 - Fragmento do *prompt* para a subtipo de regras

You are an expert in SBVR (Semantics of Business Vocabulary and Business Rules). You'll be asked to classify paragraphs into types of rules. You'll be given a definition of the types of rules.

Given the definition of types of rules, classify each rule by following subtypes.

Hints:

- If a section has no templates, look in the next section. E.g. 9.2.1 has no template, so look at 9.2.1.1, it has a template T7;
- If a section has more than one template, check when apply each template. E.g. 9.2.1.2 has two templates, so look at 9.2.1.2.1, it has a template T8;
- The subtype is the section name.

# Definition

#### ## 9.2 Definitional rules

Definitional rules constrain how we define a construct created or used by the organization or the industry within which it operates. Definitional rules can in turn be categorized as:

##### ### 9.2.1 Formal term definitions:

A formal term definition defines a particular business term in a formal manner. They are categorized as:

###### #### 9.2.1.1 Formal intensional definitions

A formal intensional definition defines the subject business term using an intensional definition: one that cites both a hypernym (a term that refers to a superset of the set referred to by the original term) and the characteristics that distinguish members of the set referred to by the original term.

Template: T7

...

Return a JSON object with the following fields.

If the query is invalid, return an empty classification.

Here's an example of output:

```
[
{
 "id": "some id",
 "text": "some text",
 "type": "Definitional"
 "subtype": "Formal intensional definitions",
 "templates": ["T7"],
 "examples": ["R70"]
},
{
 "id": ...
 "text": ...
 "type": ...
 "subtype": ...
 "templates": ...
 "examples": ...
},
]
```

Here's the paragraphs you'll need to classify:

Fonte: Elaborado pelo autor.

A validação da primeira parte foi realizada com dezoito enunciados e os resultados foram pontuações de 100% para precisão, sensibilidade, acurácia e F1. Essa parte é fundamental para a precisão da segunda parte, pois reduz a área de busca para o subtipo. Os dados usados na validação estão no “APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO”.

Quadro 31 - Exemplo da saída do algoritmo “classify”.

|                                                                             |
|-----------------------------------------------------------------------------|
| # Saída da classificação de tipo e entrada para a classificação de subtipo. |
|-----------------------------------------------------------------------------|

```
[
{
 "id": "R348",
 "type": "Data",
 "text": "Each customer complaint must specify a mobile phone number, an e-mail address, or both."
},
{
 "id": "R77",
 "type": "Definitional",
 "text": "The gender of a person is by definition either male or female."
}
]
Classificação do subtipo
[
{
 "id": "R77",
 "text": "The gender of a person is by definition either male or female.",
 "type": "Definitional",
 "subtype": "Categorization scheme enumerations"
},
{
 "id": "R348",
 "text": "Each customer complaint must specify a mobile phone number, an e-mail address, or both.",
 "type": "Data",
 "subtype": "Mandatory group rules"
}
]
```

Fonte: Elaborado pelo autor.

Para a segunda parte, o *prompt* contém toda a definição da taxonomia e quais são os *templates*. A saída da primeira parte é entrada para a segunda. O *prompt* completo está no apêndice e no Quadro 30 são destacados alguns fragmentos. Um exemplo de saída da primeira e segunda partes são apresentados no Quadro 31.

A validação foi realizada com os mesmos enunciados da primeira parte e os resultados foram pontuações de 100% para precisão, 89% para sensibilidade e

acurácia, ficando o F1 com 94%. Os dados usados na validação e a matriz de confusão estão no apêndice.

O *prompt* confundiu a classificação “Formal extensional definitions” como “Categorization scheme enumerations” para o parágrafo “A payment is by definition one of the following: a cash payment, a credit card payment, or an electronic funds transfer payment”. Segundo Witt (2012) a definição de regras extensionais formais é:

Uma definição extensional formal define o termo comercial do assunto usando uma definição extensional: uma que lista um conjunto completo de hipônimos (termos que se referem a subconjuntos do conjunto referido pelo termo original. (Witt, 2012, p. 86).

Para “Enumerações de esquemas de categorização” a definição é:

A enumeração do esquema de categorização define os membros de um esquema de categorização que é mutuamente exclusivo (ou seja, nenhum objeto individual pode ser membro de mais de uma categoria no esquema de categorização) e conjuntamente exaustivo (ou seja, cada membro da classe com a qual o esquema de categorização está associado pertence a uma categoria no esquema de categorização). (Witt, 2012, p. 86).

Pela definição, a classificação está incorreta, porém, a classificação escolhida pelo LLM corresponde a uma lista, com elementos mutuamente exclusivos e exaustiva, sendo a chave para a correta interpretação a identificação de que os elementos da lista não correspondem a hipônimos.

Quadro 32 - Comparativo dos modelos T8 e T10 de escrita de regras

| T8 (previsto)                                                                                                                         | T10 (correto)                                                                                                                                                                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\{A An\} <term\ 1>$<br>$\{of\ \{a an\} <term\ 2>\}$<br>is by definition<br>$[<article> <term\ 3>, or]$<br>$\{of\ that\ <term\ 2>\}.$ | $\{\{A An\} <category\ attribute\ term>\}$<br>The $<category\ attribute\ term>$<br>$\{of\ \{a an\} <entity\ class\ term>\}$<br>is by definition<br>$\{either\ <literal\ 1>\ or\ <literal\ 2>\}$<br>one of the following: $[<literal\ 3>, or\ \}].$ |

Fonte: Elaborado pelo autor.

Porém os *templates* são parecidos e, no Quadro 32, o *template* previsto pelo LLM (T8) e o correto (T10) são comparados e a diferença entre os dois é que

T10 enfatiza a lista de valores únicos que definem o termo, enquanto no modelo T8 isso não é observado.

A segunda confusão é com a classificação “Mandatory option selection rules” e “Mandatory group rules”, o LLM previu que no parágrafo “Each flight booking confirmation must specify exactly one of the following: a postal address, an e-mail address, or a fax number.” A classificação seria a regras de seleção de opções obrigatórias. Ambas as classificações são bem próximas na sua definição, segundo Witt (2012) a regra de grupo obrigatória exige que pelo menos um item de um grupo esteja presente e a regra de seleção de opção obrigatória exige que uma de um conjunto de opções predefinidas seja especificada. Há pouca diferença entre as duas classificações.

Quadro 33 - Comparativo entre modelos de escrita T22 e T23

| T22 (correto)                                                                                                                                                                                                                                                                                                       | T23 (previsto)                                                                                                                                                                                                                                                                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Each &lt;transaction signifier&gt;<br/>must {specify contain}<br/>{(in  for} {each the} &lt;subform term&gt; {(if any)  }<br/>{&lt;qualifying clause&gt;  }   }<br/>{a an} &lt;data item term 1&gt;, {a an} &lt;data item term 2&gt;<br/>, or but not} both<br/>{(if  unless} &lt;conditional clause&gt;  }.</p> | <p>Each &lt;transaction signifier&gt;<br/>must<br/>{(if  unless} &lt;conditional clause&gt;)   }<br/>{specify  contain}<br/>{(in  for} {each the} &lt;subform term&gt; {(if any)  }<br/>{&lt;qualifying clause&gt;  }   }<br/>&lt;cardinality&gt; of the following:<br/>[&lt;data item term&gt;, or].</p> |

Fonte: Elaborado pelo autor.

Os *templates* T22 e T23 (Quadro 33) são utilizados para definir regras obrigatórias envolvendo grupos de dados. O *template* T22 aplica-se quando há exatamente dois itens no grupo e estabelece que deve ser especificado um dos itens, ou ambos, dependendo da regra. Já o *template* T23 é utilizado quando há mais de dois itens no grupo e permite definir quantos itens devem ser especificados, como “exatamente um” ou “ao menos um”, sendo mais flexível e abrangente. A escolha do T23 para a regra, como previsto pelo modelo, não mudaria o entendimento da regra.

A taxonomia de Witt (2012) é bastante abrangente, porém há determinadas definições nas quais a diferença é sutil, se comparada as demais do mesmo tipo, isso pode levar tanto o usuário, como o modelo de LLM a confundi-las. É

interessante para trabalhos futuros acentuar as diferenças de cada uma das categorias de um determinado tipo para torná-las mais fácil de serem identificadas. O resultado dessa validação foi utilizado no projeto (*design*) dos *prompts* do algoritmo “*taxonomy classification and templates*” apresentados no Quadro 16 e Quadro 17.

#### 7.2.4 Conjunto de dados

Para o processo CFR2SBVR foram selecionados uma amostra da CFR, as seções § 275.0-2, § 275.0-5 e § 275.0-7. Essas seções foram selecionadas pelo seu tamanho e conteúdo. Elas representam um misto de fatos, regras operativas e termos e nomes com e sem definições. Representam também definições locais, que propicia a criação de múltiplos vocabulários e a quantidade de tokens<sup>65</sup>.

Tabela 1 - Conjunto de dados CFR

| Seção     | Parágrafos | Palavras | Média palavras / parágrafos | Declarações (Elementos) |
|-----------|------------|----------|-----------------------------|-------------------------|
| § 275.0-2 | 14         | 362      | 26                          | 7                       |
| § 275.0-5 | 10         | 260      | 26                          | 5                       |
| § 275.0-7 | 19         | 513      | 27                          | 10                      |
| Total     | 43         | 1.135    | 26                          | 22                      |

Fonte: Elaborado pelo autor.

As seções analisadas apresentam, em média, 14 parágrafos, 7 declarações e cerca de 26 palavras por parágrafo, conforme mostrado na Tabela 1. As declarações correspondem aos trechos identificados como regras operativas ou tipos de fatos. A proporção de aproximadamente dois parágrafos por declaração indica a presença de uma quantidade relativamente alta de conteúdo irrelevante para a extração de regras. No entanto, outras partes do texto, além das declarações, também foram utilizadas na definição de termos e nomes. A escolha desse conjunto de seções permitiu explorar a CFR com baixo custo computacional e financeiro por execução. Além disso, contribuiu para a construção do conjunto de dados ouro necessário à validação, uma vez que a

---

<sup>65</sup> Tiktoken foi a biblioteca usada na contagem. Ela é uma biblioteca de tokenização de código-livre rápida e eficiente desenvolvida pela OpenAI - <https://github.com/openai/tiktoken>.

extração manual, classificação e transformação dos elementos é um processo demorado e oneroso. Os algoritmos foram projetados para operar com qualquer número de seções e, para trabalhos futuros, outras partes da CFR poderão ser processadas.

O conjunto de dados são divididos em “previsão” e “verdade”. Em aprendizado de máquina, os conjuntos de dados referem-se, respectivamente, às previsões feitas pelo modelo e aos valores reais observados. O conjunto “previsão” contém as saídas estimadas pelo modelo treinado quando aplicado a um conjunto de entrada, enquanto o conjunto “verdade” representa os valores verdadeiros associados a essas entradas, provenientes de dados rotulados manualmente pelo autor desta pesquisa. A comparação entre esses conjuntos é discutida no capítulo “7 VALIDAÇÃO”.

Tabela 2 - Conjunto de dados ouro para validação (P1)

| Id        | Elementos | Tipo de fatos | Regras operativas | Símbolos verbais | Nomes | Termos |
|-----------|-----------|---------------|-------------------|------------------|-------|--------|
| § 275.0-2 | 7         | 5             | 2                 | 21               | 3     | 27     |
| § 275.0-5 | 5         | 1             | 4                 | 12               | 3     | 20     |
| § 275.0-7 | 10        | 10            | 0                 | 28               | 2     | 30     |
| Total     | 22        | 16            | 6                 | 61               | 8     | 77     |

Fonte: Elaborado pelo autor.

A Tabela 2 resume os dados ouro de avaliação para a primeira parte da avaliação do algoritmo de extração. Os elementos são as declarações extraídas do texto das seções e classificados como tipo de fatos (*Fact Types*) ou regras operativas (*Operative Rules*). De cada um dos elementos, foi extraído os símbolos verbais, nomes e termos.

Tabela 3 - Termos e nomes com definição.

| Id        | Termos totais | Termos com definição | Nomes totais | Nomes com definição |
|-----------|---------------|----------------------|--------------|---------------------|
| § 275.0-2 | 27            | 4                    | 3            | 2                   |
| § 275.0-5 | 20            | 16                   | 3            | 3                   |
| § 275.0-7 | 30            | 8                    | 2            | 0                   |
|           | 77            | 28                   | 8            | 5                   |

Fonte: Elaborado pelo autor.

Para a segunda parte (P2), obtenção das definições, os termos duplicados, que são referenciados em mais de um parágrafo na mesma seção com o mesmo

significado, são eliminados, a Tabela 3 resume os termos e nomes por documento.

Na etapa de classificação, foram criados conjuntos de dados ouro, para avaliar as regras operativas em relação ao topo da taxonomia de Witt (2012) – Atividade, Dados e Papeis, e para os demais elementos e níveis. Com exceção das regras operativas, as avaliações foram realizadas em uma única etapa, visto que o topo da taxonomia para esses elementos é a regra de definição. Nessa avaliação, apenas os elementos com declarações são avaliados, uma vez que o LLM deriva suas previsões das declarações. Com isso o número de termos foi reduzido para 28 e nomes para cinco. As regras operativas e os tipos de fato são, por definição, as declarações, já os termos e nomes são compostos de duas partes, o significante e o significado, ou definição. Dado o escopo reduzido do CFR2SBVR, termos e nomes, referenciados nas seções avaliadas, podem ser definidos em seções fora do escopo, ficando sem definição. Isso justifica a redução de termos para as próximas etapas.

Termos e nomes sem definição foram mantidos na inserção para o KG por ilustrarem o processo incremental da construção das regras de negócio chamado de “colcha de retalhos” por Haarst (2013). O algoritmo de transformação não dispõe de um conjunto de dados verdade e sua avaliação é realizada pelos métodos SemScore e *LLM as a judge*.

Os conjuntos de dados de previsão são análogos em estrutura, porém obtidos pela execução do algoritmo. Cada algoritmo foi executado dez vezes, o que pareceu suficiente visto que a variação entre as execuções não foram significativas, como apresentado na próxima seção.

Todos os conjuntos de dados verdade e previsão estão no repositório de código do autor na pasta “code/data”.

### 7.3 Resultados

Os principais algoritmos foram validados individualmente e são discutidos nessa seção. A principal cadeia de dependência de processos é ilustrada na Figura 12. Como cada processo depende dos dados do processo subjacente, há um

acúmulo de erros que se propaga pelos processos. Os processos probabilísticos de extração, classificação (anotação semântica) e transformação são os mais impactados. Para minimizar esse acúmulo, cada processo foi medido isoladamente e os dados para o processo seguinte reiniciados com os dados do conjunto de dados ouro. Isso permitiu, além de controlar os erros, simular a intervenção humana no processo, no qual os SMEs podem verificar as regras candidatas a cada etapa, como discutido na seção “5.2.4 Regras candidatas”. Uma lista completa das declarações e elementos extraídos, classificados e transformados está no “APÊNDICE D – RESULTADO DA TRANSFORMAÇÃO”.

O algoritmo “transform2sbvr”, como discutido anteriormente, não foi avaliado por ser o orquestrador e, para o experimento, a orquestração foi realizada manualmente. A seção “7.3.4 Uso do LLM” discute a quantidade de tokens, o tempo de execução dos processos e os custos associados.

### 7.3.1 create kg

A criação do KG é um processo determinístico e não foi aplicado método de validação. Foi realizado uma inspeção manual para determinar se o resultado estava coerente com os dados de entrada. Nesta validação foram identificados os problemas, já discutido com o CFR-FRO. Foram carregados 2.252 termos e 3.531 nomes para FIBO, totalizando 5.783. Para CFR-FRO foram carregados 1.960 parágrafos, mas como discutido, esse número não corresponde ao número de parágrafos da CFR 275<sup>66</sup> que é de aproximadamente 836.

### 7.3.2 semantic annotation

Nesta seção são avaliados os algoritmos de extração e classificação. Foram extraídos o total de 215 elementos.

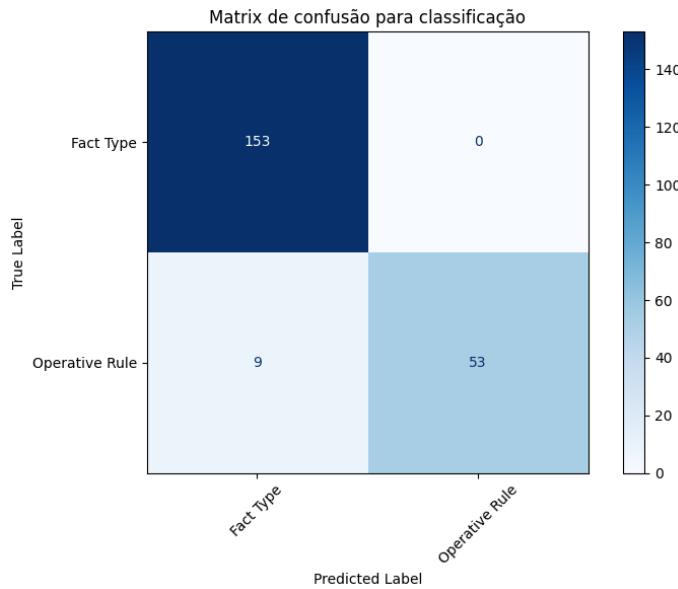
Tabela 4 - Resumo métricas do processo de extração.

| Elementos                              | Cont. | Precisão | Sensibilidade | Acurácia | F1   |
|----------------------------------------|-------|----------|---------------|----------|------|
| Extração de tipos (Fatos e Regras op.) | 215   | 0,96     | 0,96          | 0,96     | 0,96 |
| Identificação da origem                | 215   | 0,99     | 0,96          | 0,96     | 0,97 |
| Extração de termos e verbos            | 215   | 0,72     | 1,00          | 0,62     | 0,83 |

<sup>66</sup> Elementos <p> em <https://www.ecfr.gov/api/versioner/v1/full/2024-11-27/title-17.xml?part=275>. Acesso em: 15/mar./2025.

Fonte: Elaborado pelo autor.

Figura 24 - Extração e classificação das regras operativas e tipos de fatos

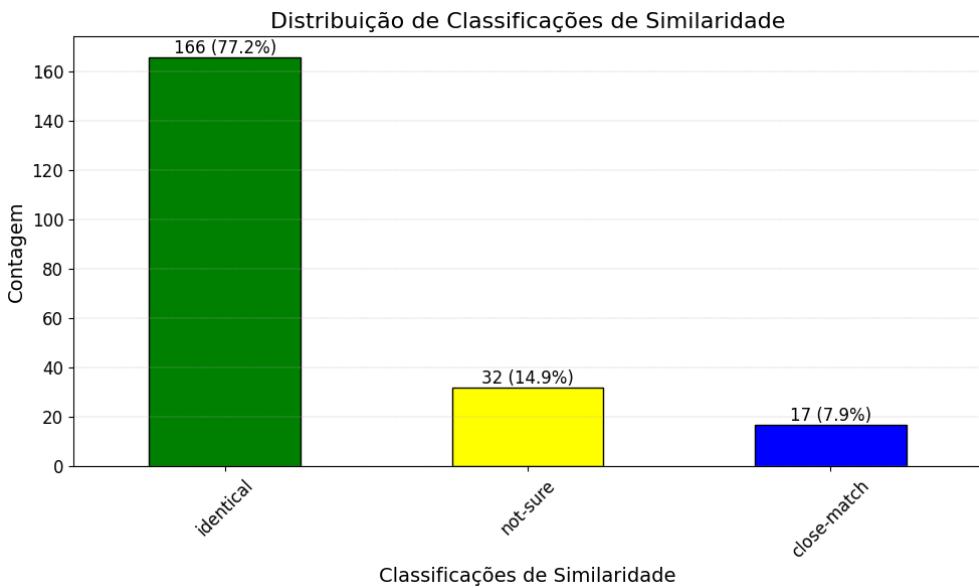


Fonte: Elaborado pelo autor.

### *Extração das regras operativas e tipos de fatos*

A anotação semântica foi avaliada em duas partes: a extração (P1) e a classificação (P2). Para a extração os dados de entrada do algoritmo são as três seções da CFR, discutidas na seção “7.2.4 Conjunto de dados”. Destas seções, foram avaliadas as declarações das regras operativas e tipos de fatos. No total foram extraídas 215 declarações classificadas em regras operativas ou tipos de fatos. O resultado da extração é apresentado na Figura 24 e os dados na Tabela 4. Os resultados demonstraram uma acurácia geral de 0,96 no conjunto de elementos avaliados. Para a classe “Fact Type”, os valores de precisão (0,95), sensibilidade (1) e F1 (0,97) indicam eficiência na identificação desta classe, sem a ocorrência de falsos negativos. Em contrapartida, para a classe “Operative Rule”, a precisão alcançou 1,00, enquanto o recall foi de 0,85, resultando em um F1 de 0,92, devido a presença de alguns falsos negativos para esta classe. As médias macro (0,94) e ponderada (0,96) dos F1-scores evidenciam um desempenho equilibrado, embora ligeiramente favorecendo a classe majoritária em termos de suporte.

Figura 25 - Similaridade entre as declarações extraídas e do conjunto de dados ouro.



Fonte: Elaborado pelo autor.

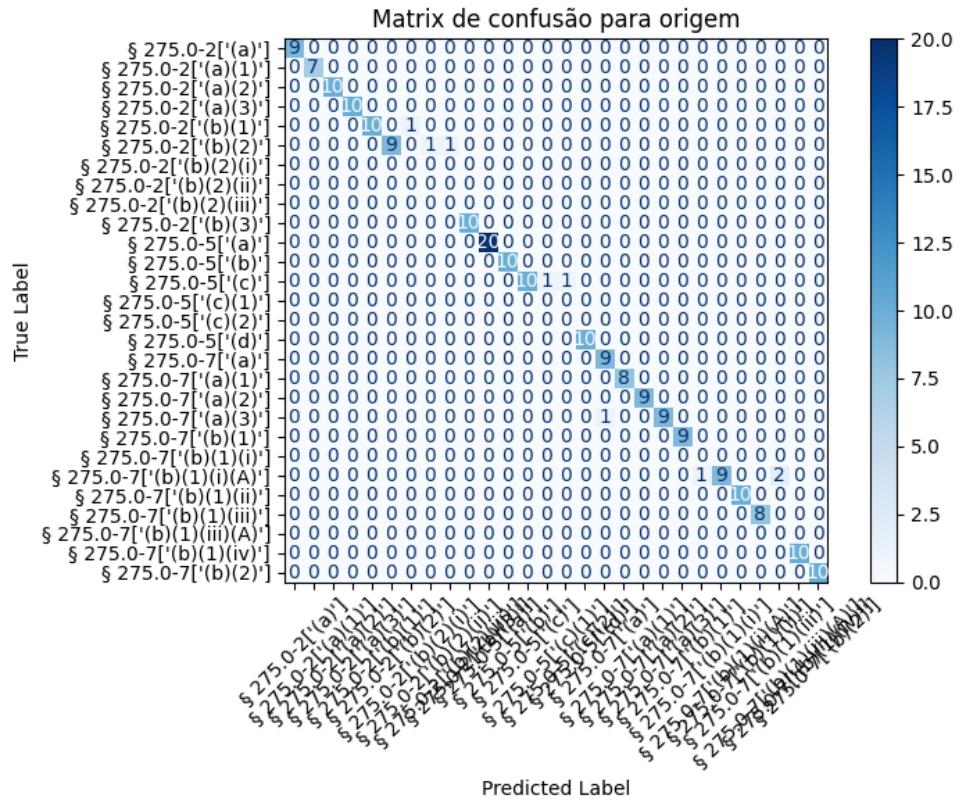
Esse resultado resume o processo de extração, todas as declarações foram extraídas e classificadas. Adicionalmente, foi realizada uma verificação de similaridade entre as declarações. O resultado é apresentado na Figura 25. As sentenças são classificadas de acordo com o percentual de similaridade: “close-match” é acima de 0,90 de similaridade e “not-sure” abaixo desse valor. A comparação foi realizada utilizando a distância de Levenshtein para computar a similaridade. A classe “not-sure” está com uma média de 0,5 de similaridade. Avaliando as 17 declarações, a baixa similaridade se deu pela violação do LLM para a instrução de extração de declarações atômicas. Em todos os casos, a declaração prevista é mais longa que a verdadeira e a verdadeira estava contida na prevista. A comparação de todas as declarações está no notebook de avaliação da extração<sup>67</sup>.

Nas execuções do processo de extração, 215 declarações (fatos e regras operativas) foram obtidas, dessas declarações foram extraídos 966 termos, 124 nomes e 607 símbolos verbais.

---

<sup>67</sup> [code/src/chap\\_7\\_validation\\_elements\\_extraction.ipynb](#)

Figura 26 - Matriz de confusão da identificação da origem dos elementos.



Fonte: Elaborado pelo autor.

Foram avaliados também a origem (identificador do parágrafo) das declarações. Os termos, nomes e símbolos verbais herdaram a origem de suas declarações e esse mecanismo é a principal forma de rastreamento dos elementos transformados as suas origens. A precisão foi de 0,996, sensibilidade de 0,96 e F1 de 0,97, indicando um desempenho consistente. O principal motivo da queda de precisão foi a unificação de sentenças. Por exemplo, os parágrafos (b)(2)(i), (b)(2)(ii) e (b)(2)(iii) da seção § 275.0-2 foram reescritos para tornar a declaração atômica. A decisão no conjunto de dados verdade foi de considerar o parágrafo (b)(2) como origem. Pela matriz de confusão, apresentado na Figura 26, é evidenciado que a maioria das execuções extraiu declarações atômicas, porém algumas execuções não, mas identificaram a origem correta.

## *Extração dos símbolos verbais, termos e nomes*

Quando comparados com o conjunto de dados ouro, a precisão da extração para símbolos verbais foi de 0,8 e a sensibilidade de 1,00, com F1 de 0,89. A métrica

foi afetada pelos falsos positivos (108), extraídos principalmente das sentenças mais longas, caracterizando a propagação do erro.

A avaliação dos termos e nomes foi realizada de forma quantitativa e qualitativa, avaliando a similaridade das suas definições. Quantitativamente, dos 966 termos extraídos, 934 estão relacionados ao conjunto de dados ouro e 32 termos “extras” foram extraídos. Os termos extras são, na sua maioria, termos compostos de dois ou mais termos que deveriam ser separados pela instrução dada ao LLM (Quadro 14), por exemplo, o termo extra “Process, pleadings, or other papers” foi extraído nas demais execuções como “Process”, “Pleadings” e “Papers”. Isso fez com que a acurácia (0,61) e a precisão (0,72) fossem baixas, mas a sensibilidade é 1,0, ficando o F1 em 0,84. Os nomes tiveram resultado semelhante, mas como as amostras são menores, os falsos positivos tiveram uma influência ainda maior nas pontuações, ficando com precisão de 0,63, acurácia de 0,42, mas com sensibilidade de 1,0 e F1 de 0,77.

Outro atributo dos termos e nomes é se a definição é local (*isLocalScope*), usada para definir o vocabulário. O resultado, para os termos e nomes com definições foi precisão de 0.71, sensibilidade de 0.99 e com F1 de 0.83. Novamente, os falsos positivos contribuíram para reduzir a precisão. Para os termos e nomes sem definição, o valor deste atributo é sempre falso.

A característica probabilística do modelo pode ser observada na classificação de falsos positivos, a cada execução um conjunto aparentemente aleatório é introduzido, afetando as pontuações finais. A forma de lidar com esse problema é aumentando o número de amostras e selecionado apenas as com maior representatividade ou submeter os resultados a uma avaliação dos SMEs.

A segunda avaliação foi a de similaridade, realizada comparando as definições obtidas para os termos (P2) com as do conjunto de dados ouro. Os termos e nomes sem definição também foram comparados para verificar se o modelo “alucinou”, criando definições que contrariam a instrução de extrair apenas textos presentes na seção. Apenas os termos e nomes foram avaliados. O resultado foi a totalidade dos termos avaliados obterem pontuação 1,00 de similaridade, indicando que todas as declarações extraídas estão presentes no texto original.

A partir da extração, a classificação de cada elemento segundo a taxonomia de Witt (2012) é avaliada independentemente.

Tabela 5 - Conjunto de dados ouro para a classificação (P1)

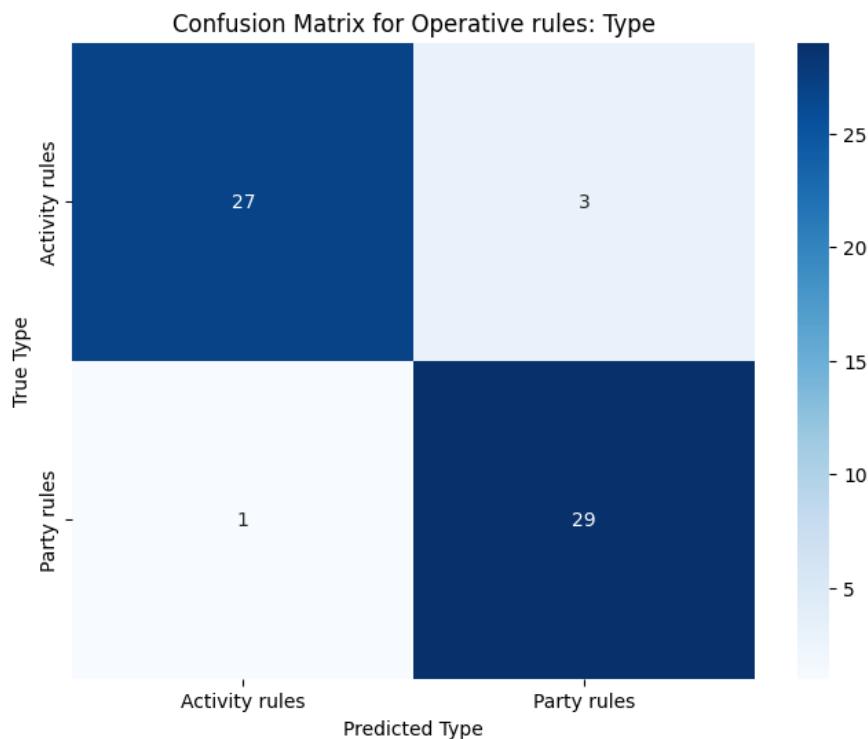
| Id        | Tipo de fatos | Regras operativas | Nomes | Termos |
|-----------|---------------|-------------------|-------|--------|
| § 275.0-2 | 5             | 2                 | 2     | 4      |
| § 275.0-5 | 1             | 4                 | 3     | 16     |
| § 275.0-7 | 10            | 0                 | 0     | 8      |
| Total     | 16            | 6                 | 5     | 28     |

Fonte: Elaborado pelo autor.

#### *Classificação de regras operativas, fatos, termos e nomes*

A execução foi realizada utilizando os elementos do conjunto de dados ouro para não propagar os erros como discutido anteriormente, o Tabela 5 resume os elementos que foram utilizados na classificação. Após a execução foram obtidas 60 regras classificadas.

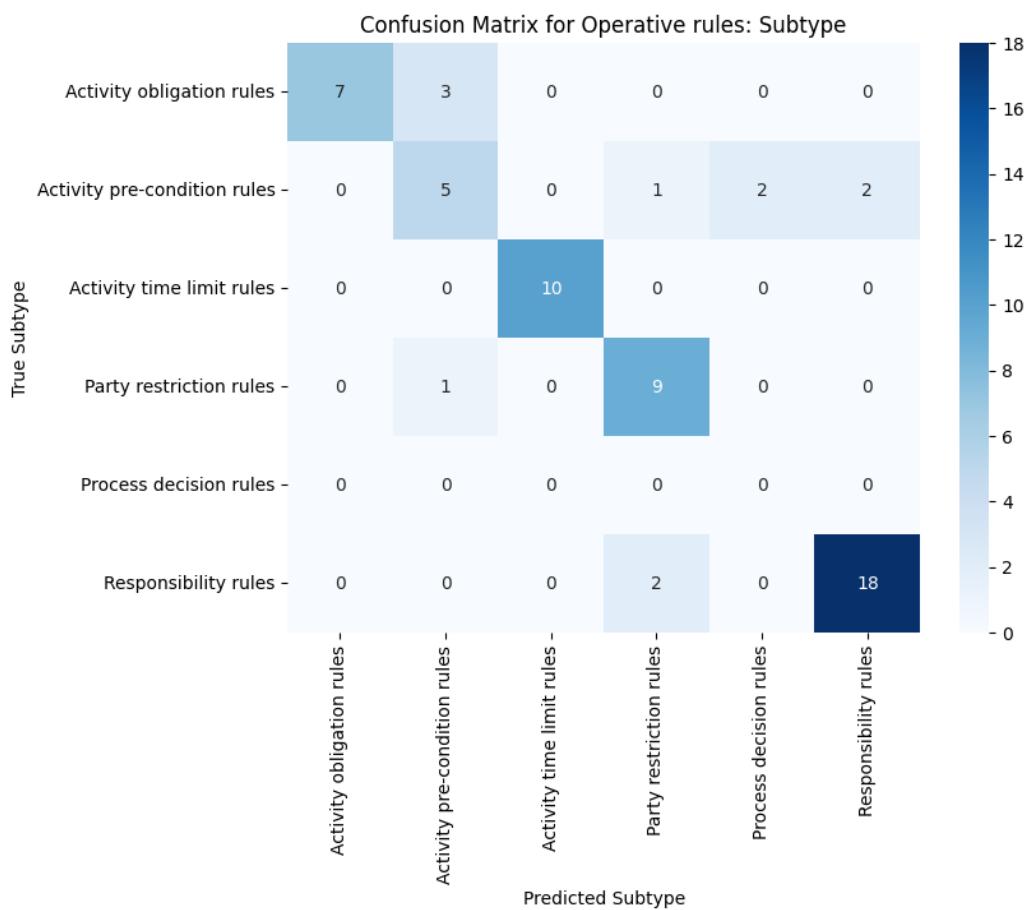
Figura 27 - Classificação das regras operativas (P1)



Fonte: Elaborado pelo autor.

A classificação das regras operativas foi dividida em duas partes, sendo a primeira classificando as regras operativas identificadas na extração no topo da taxonomia e a segunda nos demais níveis. O resultado é apresentado na Figura 27: a precisão foi de 0,94, sensibilidade 0,93 e F1 de 0,93. Duas regras foram classificadas de forma equivocada, o parágrafo § 275.0-5 - ['(c)'], três vezes e o § 275.0-5 - ['(a)'], sem um motivo aparente.

Figura 28 – Classificação (P2) das regras operativas nos subníveis da taxonomia.

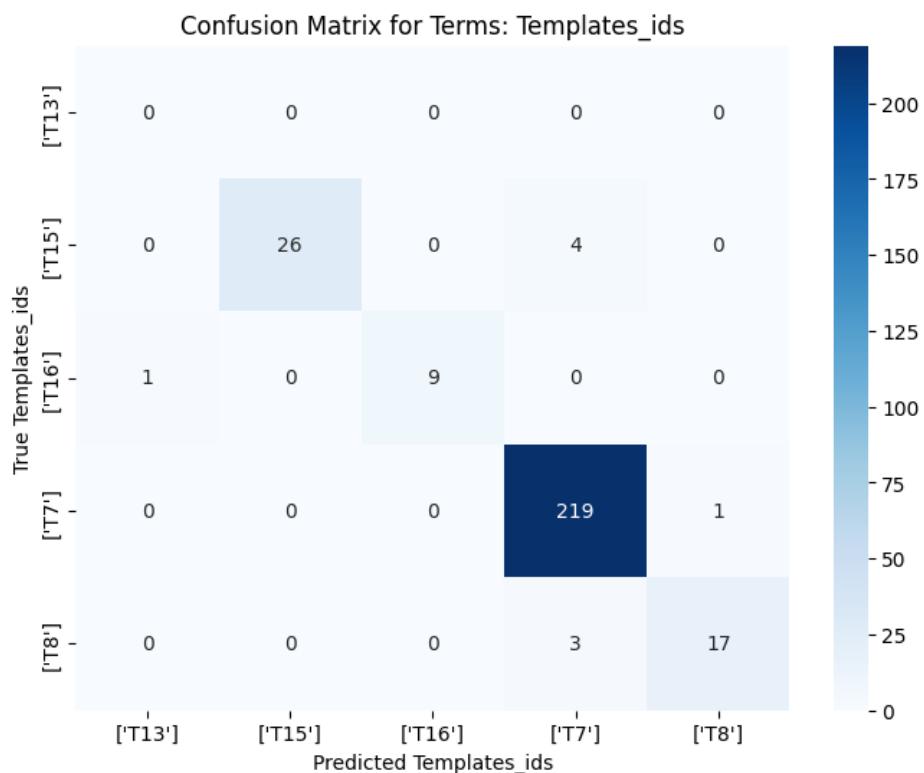


Fonte: Elaborado pelo autor.

A segunda parte classificou todos os elementos nos demais níveis da taxonomia. Para termos, nomes e tipos de fato, a classificação de topo é regras de definição. Somente as regras operativas necessitaram de dois níveis de classificação. Os resultados foram precisão de 0,85, sensibilidade de 0,83 e F1 de 0,83. O principal motivo que reduziu a precisão foi identificar o subtipo de regra de atividade - regra de pré-condição de atividade, que proíbe um processo de negócios ou

outra atividade, a menos que alguma outra atividade ou evento tenha ocorrido anteriormente ou exista alguma condição de pré-requisito (Witt, 2012 p. 96).

Figura 29 - *Templates* para as classificações de termos.



Fonte: Elaborado pelo autor.

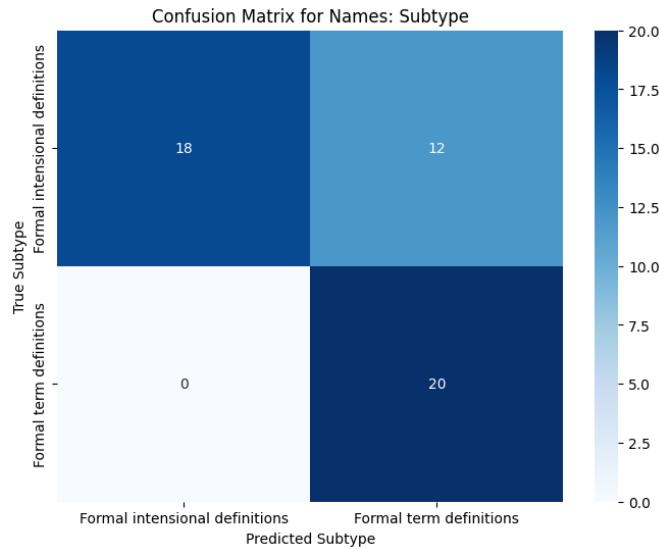
Na Figura 28, com cinco classificações erradas das onze, a declaração “The Commission will order a hearing on the matter, if it appears that a hearing is necessary or appropriate in the public interest or for the protection of investors, upon the request of any interested person or upon its own motion.” foi interpretada como restrição de papéis, decisão de processo e responsabilidade de papéis, sendo a classificação correta a de pré-condição para iniciar uma atividade. Outros três erros envolvem o entendimento do LLM sobre as regras de pré-condição de atividade. Como discutido na seção “7.2.3 Avaliação da classificação da taxonomia de regras”, a confusão do modelo é exemplo da sutileza das classes definidas por Witt (2012).

Figura 30 - Classificação (P2) para termos.



Outro exemplo é ilustrado na Figura 30. A precisão para a classificação de termos foi de 0,97, com sensibilidade de 0,90 e F1 de 0,94, porém, quando analisado pela dimensão dos *templates*, o resultado alcançou precisão, sensibilidade e F1 de 0,97 (Figura 29), porque algumas classes compartilham o mesmo *template*. A classe que teve maior influência sobre a classificação dos termos foi a definição formal. Essa é uma superclasse da classe definição formal intensional – que define citando tanto um hiperônimo (um termo que significa um superconjunto do conjunto significado pelo termo original) quanto as características que distinguem os membros do conjunto significado pelo termo original (Witt, 2012). Sem um motivo aparente, quatro termos extraídos da seção § 275.0-5: “Facts” (a); “Reasons” (a); “Hearing” (a), (b), (c) e “Nature of interest” (a) foram classificados com a superclasse, contabilizando 18 dos 27 erros de classificação.

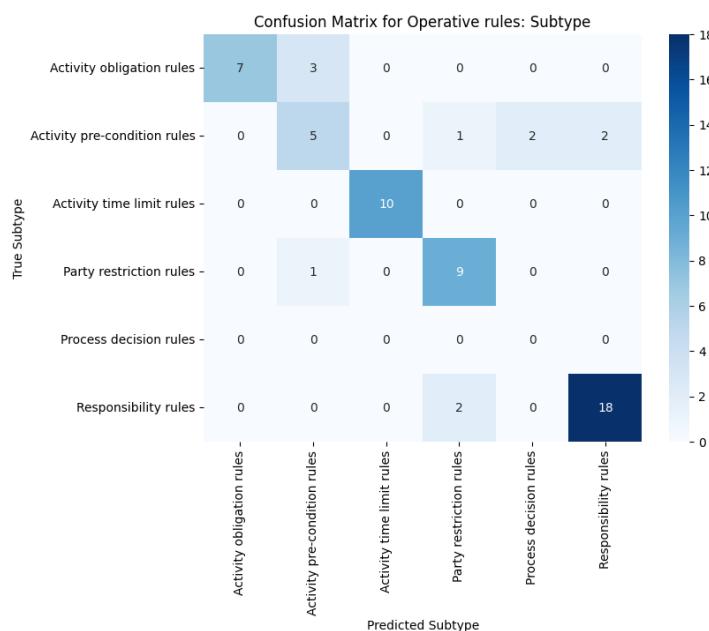
Figura 31 - Classificação (P2) para nomes.



Fonte: Elaborado pelo autor.

A classificação dos nomes, apresentados na Figura 31, teve um problema semelhante de classificação, mas como a quantidade de amostras é menor, a influência foi mais acentuada nas pontuações de precisão (0,85), sensibilidade (0,76) e F1 (0,76), mas com impacto baixo para a transformação, visto que todos utilizam o mesmo *template* (T7).

Figura 32 - Classificação (P2) para tipo de fatos



Fonte: Elaborado pelo autor.

Na Figura 32, são apresentados os resultados dos tipos de fatos que obtiveram pontuações de 0,85 para precisão, 0,82 para sensibilidade e 0,83 para F1. Duas declarações foram as principais responsáveis pelas pontuações, ambas classificadas no conjunto de dados verdade como regras de pré-condição para atividade. A mesma classe que foi confundida nas regras operativas. De fato, são os tipos de fato que dão suporte as regras operativas nos parágrafos § 275.0-5 (a) e (c).

### *Resumo*

Tabela 6 - Resumo das métricas de classificação.

| Elementos                 | Cont. | Precisão | Sensibilidade | Acurácia | F1   |
|---------------------------|-------|----------|---------------|----------|------|
| Tipos de fatos subtipo    | 160   | 0,97     | 0,90          | 0,90     | 0,93 |
| Nomes subtipo             | 50    | 0,85     | 0,76          | 0,76     | 0,76 |
| Regras operativas tipo    | 60    | 0,94     | 0,93          | 0,93     | 0,93 |
| Regras operativas subtipo | 60    | 0,86     | 0,83          | 0,83     | 0,84 |
| Termos subtipo            | 280   | 0,97     | 0,90          | 0,90     | 0,94 |

Fonte: Elaborado pelo autor.

Em geral, todos os elementos classificados obtiveram uma precisão acima de 0,85, como apresentado na Tabela 6, com variações mínimas entre as execuções, porém cabe ressaltar que esse resultado foi obtido ajustando a entrada dos processos subsequentes com os dados do conjunto de dados ouro, simulando a intervenção dos SMEs a cada etapa do processo. Se não considerarmos esses ajustes, o acúmulo de erros, tomando a pontuação F1 como referência, seria de:

$$P(\text{Sucesso}) = \bar{P}_e \cdot \bar{P}_c \cdot \bar{P}_t = 0,73$$

Onde,  $\bar{P}_e = 0,90$  é a média do F1 da extração (Tabela 4),  $P_c = 0,88$  é a média do F1 da classificação (Tabela 6) e  $\bar{P}_t = 0,92$  é a média da pontuação de similaridade do *LLM as a judge* para a transformação. O uso da avaliação como parte do processo, pode auxiliar na escolha dos melhores resultados de cada execução (métricas próximas de 1,00) como saída para o próximo processo. Veja os notebooks `chap_7_validation_*_cumulative.ipynb` para os resultados com erro acumulativo.

### 7.3.3 Transformação

Para a etapa de transformação não foi criado um conjunto de dados ouro, a avaliação foi realizada por dois algoritmos: SemScore (AYNETDINOV; AKBIK, 2024) e “LLM as a Judge” (WEI; CHEN; LUO, 2024), (DONG; HU; COLLIER, 2024), (ZHENG *et al.*, 2023)., como discutido na seção “7.2.2 Avaliação de similaridade semântica”. O resultado de cada algoritmo e a correlação entre eles é discutido nas seções seguintes.

As principais métricas utilizadas são o SemScore, que mede a similaridade semântica entre as respostas geradas pelos modelos e as respostas-alvo. Essa métrica utiliza *embeddings* de frases e a similaridade de cosseno, para a comparação do conteúdo semântico das sentenças. Também foi empregado o *Similarity Score*, calculado a partir de um LLM utilizando o método *LLM as a judge*. Outras duas métricas foram calculadas, a *Transformation Accuracy* mede a fidelidade da sentença transformada em relação à estrutura da sentença original em uma escala de 0 a 1, enquanto a *Grammar and Syntax Accuracy* avalia a correção gramatical e sintática das sentenças transformadas, também em uma escala de 0 a 1. Essas métricas adicionais permitem que a análise dos resultados vá além da similaridade semântica, incluindo a precisão estrutural e gramatical das transformações, proporcionando uma avaliação mais completa.

O conjunto de dados avaliado é o resultado da transformação e é constituído de 60 regras operativas, 160 tipos de fatos, 280 termos e 50 nomes. Da mesma forma que as etapas anteriores, a etapa de transformação foi realizada utilizando os dados do conjunto de dados ouro de classificação.

#### SemScore

Tabela 7 - Estatística das métricas combinadas.

| Elementos         | Cont. | SemScore |      |      |      |      | LLM as a judge (Similarity Score) |      |      |      |      |
|-------------------|-------|----------|------|------|------|------|-----------------------------------|------|------|------|------|
|                   |       | Média    | DP   | Min  | 75%  | Max  | Média                             | DP   | Min  | 75%  | Max  |
| Tipos de fatos    | 160   | 0,87     | 0,05 | 0,71 | 0,92 | 0,97 | 0,92                              | 0,06 | 0,70 | 0,95 | 0,95 |
| Nomes             | 50    | 0,88     | 0,03 | 0,81 | 0,90 | 0,91 | 0,95                              | 0,01 | 0,90 | 0,95 | 0,95 |
| Regras operativas | 60    | 0,91     | 0,02 | 0,87 | 0,92 | 0,93 | 0,90                              | 0,05 | 0,75 | 0,95 | 0,95 |
| Termos            | 280   | 0,85     | 0,08 | 0,50 | 0,90 | 0,96 | 0,92                              | 0,05 | 0,60 | 0,95 | 1,00 |

Fonte: Elaborado pelo autor.

A métrica do SemScore é o resultado do cálculo da distância do cosseno para os vetores densos da declaração original e da transformada. O cálculo dos vetores foi realizado utilizando o modelo “text-embedding-3-large” da OpenAI conforme discutido na seção “6.1 Ferramentas e infraestrutura”. Essa é uma operação custosa tanto monetariamente, quanto em tempo. O resultado é resumido na Tabela 7, onde a tradução dos elementos obteve um desempenho homogêneo, as regras operativas tiveram um desempenho ligeiramente melhor do que os outros elementos.

Os vetores são fortemente influenciados pelas palavras. O principal caso em que o SemScore ficou abaixo de 0,8 foi para o termo “Protection of investors”. Todas as dez instâncias desse termo obtiveram uma pontuação abaixo de 0,59, embora a pontuação de *Similarity Score* tenha variado de 0,85 a 0,90. A declaração original “A consideration for ordering a hearing if it appears necessary or appropriate.” foi classificada como uma regra de definição do tipo definição intensional e a declaração transforma foi “A consideration for ordering a hearing is by definition a consideration that appears necessary or appropriate.”. O resultado da transformação preservou o significado do termo, porém modificou a ordem das palavras e adicionou outras.

Na métrica SemScore, o vetor resultante ficou distante da declaração original, embora o significado não. Há poucas instâncias com essa discrepância, o menor SemScore é 0,50 com um *Similarity Score* de 0,85. Um exemplo inverso, que ilustra a característica probabilística do LLM, foi o termo “Order of the Commission”, cuja declaração original é “An order issued by the Commission under the Act.” e a transformação é “An order is by definition the order of the Commission that is issued under the Act.”. Para o SemScore, a pontuação foi acima de 0,81 em todas as instâncias desse temos, porém, para o *Similarity Score*, em duas instâncias, de dez, a avaliação foi de 0,6, com a justificativa de que a frase transformada apresenta erros de gramática, omissão de termos essenciais (“issued by” e “issued by the Commission”) e falha em refletir com precisão o significado original. Ainda segundo o LLM, a omissão do “issued by” leva a entender que a ordem é a comissão, na qual claramente não houve omissão e o significado não leva a entender isso.

Tabela 8 - Pontuações de qualidade da transformação

| Elementos     | Cont. | Transformation Accuracy |      |      |      |      | Grammar Syntax Accuracy |      |      |      |      |
|---------------|-------|-------------------------|------|------|------|------|-------------------------|------|------|------|------|
|               |       | Média                   | DP   | Min  | 75%  | Max  | Média                   | DP   | Min  | 75%  | Max  |
| Tipo de fatos | 160   | 0,87                    | 0,09 | 0,50 | 0,90 | 0,95 | 0,93                    | 0,09 | 0,50 | 0,95 | 1,00 |
| Nomes         | 50    | 0,92                    | 0,03 | 0,85 | 0,95 | 0,95 | 0,97                    | 0,02 | 0,95 | 1,00 | 1,00 |
| Regras op.    | 60    | 0,84                    | 0,07 | 0,70 | 0,90 | 0,90 | 0,93                    | 0,05 | 0,70 | 0,95 | 0,95 |
| Termos        | 280   | 0,88                    | 0,08 | 0,50 | 0,95 | 1,00 | 0,95                    | 0,06 | 0,40 | 1,00 | 1,00 |

Fonte: Elaborado pelo autor.

Adicional a avaliação de similaridade, como proposto por Shankar *et al.* (2024), foram adicionadas duas métricas, a acurácia na tradução (*Transformation Accuracy*) e a acurácia na gramática e sintaxe (*Grammar Syntax Accuracy*), apresentados na Tabela 8.

### *LLM as a Judge*

Tabela 9 - Pontuações de confiança no julgamento do LLM

| Elementos         | Cont. | Média | DP   | Min  | 25%  | 50%  | 75%  | Max  |
|-------------------|-------|-------|------|------|------|------|------|------|
| Tipo de fatos     | 160   | 0,89  | 0,03 | 0,70 | 0,90 | 0,90 | 0,90 | 0,90 |
| Nomes             | 50    | 0,90  | 0,01 | 0,90 | 0,90 | 0,90 | 0,90 | 0,95 |
| Regras operativas | 60    | 0,88  | 0,03 | 0,80 | 0,85 | 0,90 | 0,90 | 0,90 |
| Termos            | 280   | 0,89  | 0,03 | 0,80 | 0,85 | 0,90 | 0,90 | 1,00 |

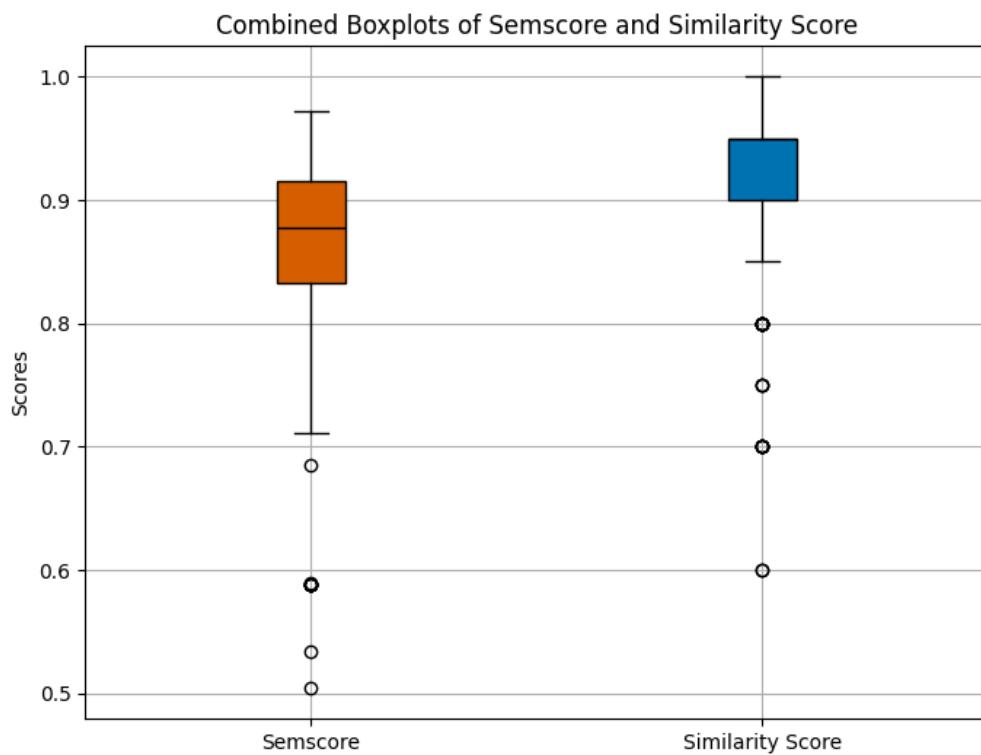
Fonte: Elaborado pelo autor.

Os resultados da avaliação do *LLM as a judge*, métrica *Similarity Score*, são resumidos na Tabela 7 e a pontuação da confiança no julgamento na Tabela 9, essas pontuações indicam a avaliação de similaridade e o nível de confiança, do LLM, nesta avaliação. No geral, o LLM forneceu, com alta pontuação de confiança apresentado na Tabela 9, pontuações de similaridade ligeiramente mais altas do que as do SemScore, embora, concordem na maioria das instâncias, as avaliações de maior interesse são quando discordam, como será discutido na seção “7.3.3 Transformação - Relação entre as métricas”.

Em relação a qualidade da transformação, o LLM foi instruído a verificar a correspondência da declaração traduzida em relação ao *template* utilizado. Na maioria dos casos, a pontuação *Transformation Accuracy* foi acima de 0,9, indicando que a tradução foi realizada seguindo corretamente o *template*.

A avaliação da métrica *Grammar Syntax Accuracy* auxiliou no entendimento se o uso do *template* resultou em uma declaração gramaticalmente correta, visto que o usuário da declaração são SMEs. Vale destacar que as piores pontuações para essa avaliação foram atribuídas ao termo “Order of the Commission”, já discutido, embora não foram identificados problemas aparentes.

Figura 33 - Distribuição para todos os elementos das pontuações de similaridade semântica

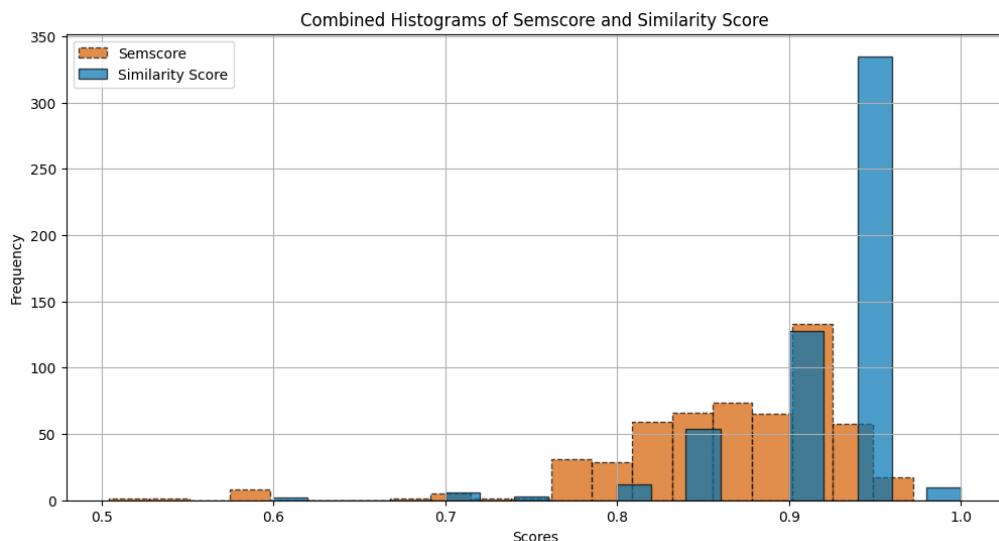


Fonte: Elaborado pelo autor.

Outro exemplo, explicado pela característica probabilística do LLM, foi o termo “Trust”. Sua declaração original é “A person is presumed to control a trust if the person is a trustee or managing agent of the trust.” e a versão transformada é “A person is by definition a controller of a trust that is presumed to control (the trust) if the person is a trustee or managing agent of the trust.”. Em duas instâncias o “the trust” foi omitido. Em uma única instância, a pontuação *transformation\_accuracy* atribuído foi de 0,6 e de 0,5 para *grammar\_syntax\_accuracy*. A justificativa do LLM para as pontuações baixas foi a interpretação de que a definição de pessoa é “presumida com controladora”.

Em todos os casos, o *LLM as a judge* e o SemScore atribuíram uma nota acima de 0,8 para a similaridade e as notas de qualidade foram acima de 0,8.

Figura 34 - Histograma de distribuição do *Similarity Score* e *SemScore*

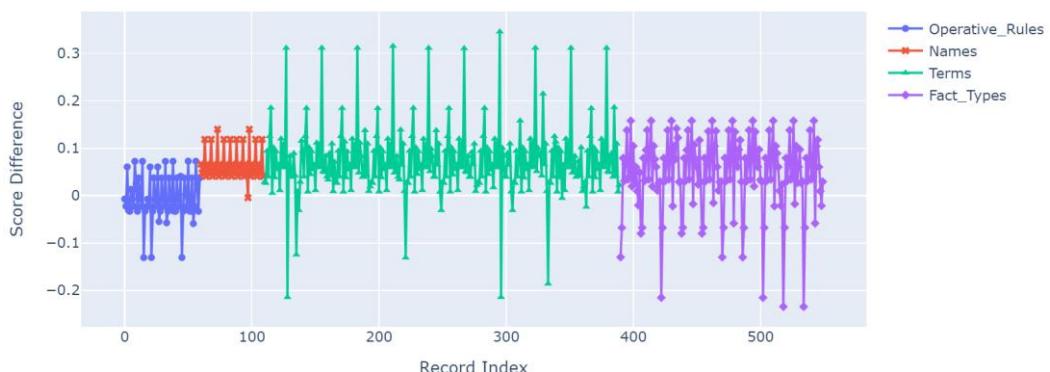


Fonte: Elaborado pelo autor.

A Figura 34 compara as distribuições de SemScore e *LLM as a judge* (*Similarity Score*). O *Similarity Score* apresenta maior concentração de valores próximos a 1,0, com menor dispersão e uma mediana mais alta. Já o SemScore exibe maior variabilidade, com a mediana ligeiramente abaixo de 0,9 e valores discrepantes mais baixos, chegando próximo a 0,5, como apresentado na Figura 33, confirmando o viés do SemScore de atribuir valores mais baixos do que o LLM.

### Relação entre as métricas

Figura 35 - Diferença entre *Similarity Score* (*LLM as a judge*) e *SemScore* por tipo de elemento.



Fonte: Elaborado pelo autor.

A Figura 35 ilustra a variação das diferenças do *Similarity Score* e *SemScore*. Essa análise por tipo de elemento revela padrões distintos, sendo o valor zero quando as métricas concordam, é um (positivo ou negativo), quando discordam, valores positivos indicam que a métrica de *Similarity Score* é maior do que a *SemScore*, o que é verdade para 75% das instâncias. “Operative\_Rules” apresentam a menor variação nas diferenças, agrupando-se de forma consistente em torno de zero, o que indica um maior nível de concordância entre as pontuações. “Names” mostram um intervalo moderado de diferenças, com valores ocasionalmente mais elevados, sugerindo uma variabilidade ligeiramente maior. No entanto, a avaliação combinada de similaridade foi a mais elevada. “Terms” demonstram a maior amplitude, com valores extremos positivos e negativos, evidenciando discordâncias significativas. “Fact\_Types”, embora tenha mais variações que “Operative\_Rules”, apresentam uma faixa de diferenças moderadamente consistente, indicando um nível de concordância entre as pontuações maior do que em “Terms”. Essas diferenças sugerem que a natureza do conteúdo em cada tipo de elemento influencia o alinhamento entre *Similarity Score* e *SemScore*, em certas categorias, como “Terms”, exibindo uma maior discrepância nas pontuações.

A variação em torno de 0,3 sugere que as métricas não discordam fortemente, embora, com uma tolerância de 0,1, as métricas concordam apenas 6% das instâncias, sendo que *Similarity Score* tende a indicar, com maior frequência, uma similaridade maior do que *SemScore*.

A análise das métricas de correlação — Pearson, Spearman e Kendall — forneceu insights complementares sobre a relação entre as variáveis investigadas. O coeficiente de correlação de Pearson ( $r = 0,2573, p < 10^{-9}$ ) indicou uma relação linear positiva fraca a moderada entre as variáveis, sugerindo uma associação estatisticamente significativa, mas não fortemente linear. O coeficiente de correlação de Spearman ( $\rho = 0,3294, p < 10^{-14}$ ) refletiu uma relação monotônica positiva moderada, capturando uma associação mais forte do que a de Pearson, consistente com sua capacidade de detectar padrões monotônicos lineares e não lineares. Da mesma forma, o coeficiente de

correlação de Kendall ( $\tau = 0,2617, p < 10^{-14}$ ) confirmou uma relação monotônica estatisticamente significativa, mas ligeiramente mais conservadora, destacando a proporção de pares de classificação concordantes e discordantes. A discrepância entre essas métricas apontou para a possível presença de dinâmicas não lineares, já que Spearman e Kendall são mais robustos a *outliers* e não linearidades em comparação com Pearson. Coletivamente, esses resultados indicaram uma associação moderada e estatisticamente significativa entre as variáveis, com a relação sendo mais bem capturada pelas métricas baseadas em *ranking* devido à sua sensibilidade a padrões monotônicos.

Os exemplos apresentados, ilustram que o *Similarity Score* baseado em LLM captura melhor a discrepancia de significado, evidenciando as limitações do SemScore em isolar falhas contextuais e de interpretação. Isso sugere a necessidade de complementar o SemScore com avaliações mais contextualizadas, como o *Similarity Score* com suas observações explicativas, para assegurar que as transformações mantêm a fidelidade semântica e estrutural das sentenças originais.

Os resultados sugerem que o uso isolado de *Similarity Score* pode ser insuficiente para capturar nuances semânticas, e que a combinação com métricas como o SemScore é necessária para uma avaliação mais completa. O contrário também se mostrou verdadeiro, no qual o SemScore falhou em capturar um aspecto mais contextual. Em estudos futuros, outras métricas semânticas podem ser combinadas para que aspectos de similaridade estrutural e semântica sejam mais bem avaliados. Todas os valores computados para as métricas podem ser encontrados no repositório de código do autor no arquivo “code/src/chap\_7\_validation.ipynb”.

#### 7.3.4 Uso do LLM

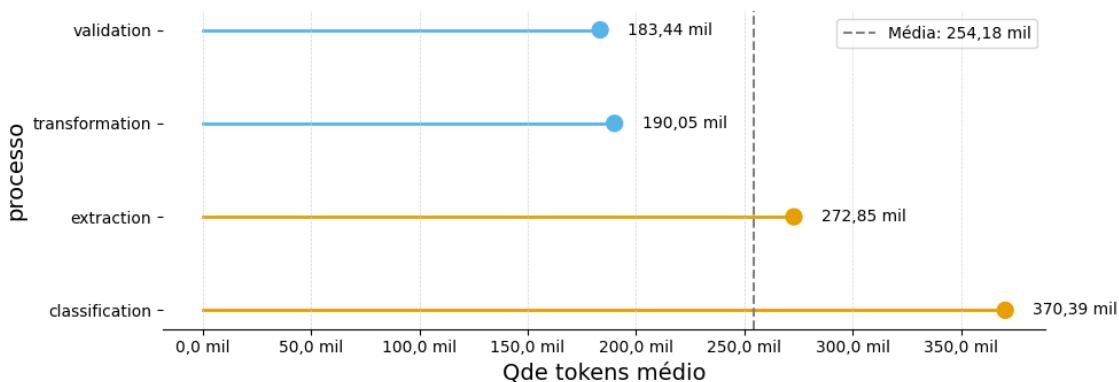
Esta seção examina os dados coletados durante o processo de transformação das regras em NL para SBVR, esses dados são obtidos a cada execução da API que retorna, juntamente com a resposta, o objeto “chat completion”<sup>68</sup>. A média

---

<sup>68</sup> <https://platform.openai.com/docs/api-reference/chat/object>. Acesso em: 15/mar./2025.

de tokens computados durante uma execução e de aproximadamente 1,02 milhões de tokens. Foram exploradas as métricas de uso de tokens, tempo de execução e custo para oferecer insights sobre os aspectos operacionais e financeiros do uso de LLMs nesse cenário.

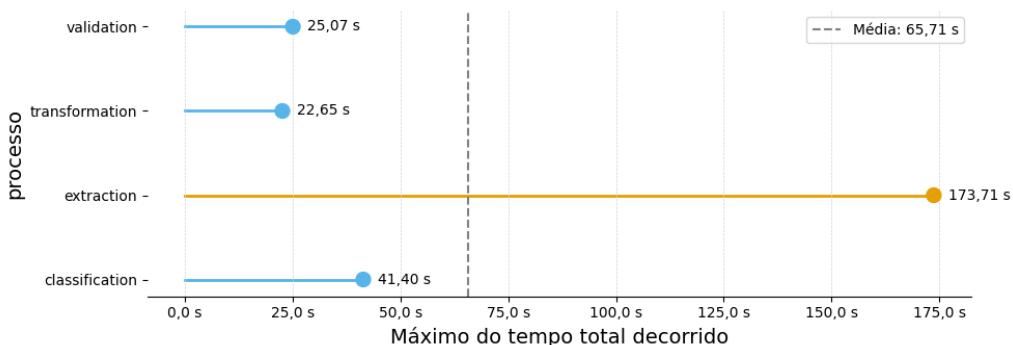
Figura 36 – Uso dos *tokens* por processo.



Fonte: Elaborado pelo autor.

A distribuição média de *tokens* por execução destaca um uso médio de aproximadamente 254,18 mil *tokens* por processo. O processo de extração extraiu 168 elementos, sendo 55 elementos com definição e 113 sem definição. O processo de classificação tem o maior uso de tokens, como apresentado na Figura 36. Esse processo classifica apenas os termos com definição dos elementos extraídos, sendo eles: 6 regras operativas, 28 termos, 16 tipos de fatos e 5 para nomes (ver Tabela 2 e Tabela 3). A entrada do processo de transformação são os elementos classificados e o de Validação os elementos transformados.

Figura 37 - Tempo total por processo



Fonte: Elaborado pelo autor.

O tempo médio decorrido para as execuções dos processos (Figura 37) foi de 262,83 segundos (4,38 minutos). O processo de extração tem a maior duração. Inicialmente, esse processo extraiu todos os 168 elementos (P1) e, em uma segunda execução, extraiu a definição dos 85 termos e nomes (P2). Todos os processos foram executados 10 vezes e a duração média de um processo é de 63,71 segundos.

Do ponto de vista da complexidade computacional, o mecanismo de autoatenção em transformadores — principal componente das arquiteturas de LLMs — apresenta complexidade quadrática em relação ao comprimento da sequência de entrada, denotado por  $n$ . Para uma entrada de tamanho  $n$  e dimensão vetorial  $d$ , a complexidade é  $O(n^2 \cdot d)$ . Isso se deve ao cálculo da matriz de similaridade, baseado em produtos internos, e à aplicação da operação softmax, que exigem operações entre todos os pares de tokens, o que leva ao crescimento quadrático conforme  $n$  aumenta (KELES et al., 2022). Por outro lado, no contexto do CFR2SBVR, a chamada à API externa apresenta complexidade  $O(n)$ . Como detalhado na seção “6.2 Implementação dos principais componentes”, os algoritmos utilizados possuem, em geral, um único laço que itera sobre  $n$  seções, declarações ou termos. Assim, tanto a complexidade de cada etapa quanto a complexidade total permanecem em  $O(n)$ .

O custo monetário para executar todo o processo (extração, classificação e transformação) foi de aproximadamente USD 25.50, calculado com base em um custo de USD 2.50 por milhão de *tokens*<sup>69</sup>. Os dados destacam a importância de otimizar tanto a estrutura dos *prompts* quanto das respostas para alcançar operações mais econômicas sem comprometer a qualidade do resultado. O custo calculado foi para a execução de todos os *prompts* para as três seções da CFR. Porém, durante a fase de desenvolvimento, cada *prompt* foi executado no mínimo 100 vezes, resultando em um custo cerca de 100 vezes maior do que o calculado para uma execução.

Essa análise fornece uma estimativa do custo do uso da ferramenta CFR2SBVR para suporte aos SMEs. O principal objetivo é avaliar a viabilidade do uso pela

---

<sup>69</sup> <https://openai.com/api/pricing/>. Acesso em: 15/mar./2025.

perspectiva econômica e tempo de execução, sendo o seu uso viável para aprimorar a eficiência e a escalabilidade a um custo recorrente moderado. Para mais métricas e seus respectivos dados, veja os *jupyter notebooks* “chap\_7\_validation\_\*” no repositório do autor.

## 8 CONSIDERAÇÕES FINAIS

### 8.1 Discussão

A transformação de regulamentações financeiras em regras SBVR é uma tarefa complexa que se beneficia da combinação de técnicas avançadas de NLP e refinamento conduzido por especialistas. Os desafios centrais – desde a interpretação de jargão jurídico e sentenças longas até o tratamento de referências cruzadas e ambiguidades – exigem uma metodologia rigorosa. Modelos de linguagem desempenham um papel importante na automação de parte dessa transformação, fornecendo um entendimento avançado da linguagem para extrair e estruturar as regras. No entanto, suas limitações, como possíveis imprecisões e falta de explicabilidade, exigem que sejam utilizados dentro de um fluxo de trabalho controlado e auditável. Garantir precisão e conformidade requer validação e auditoria em todas as etapas e o envolvimento de especialistas.

A análise realizada na seção “6.2 Implementação dos principais componentes” apresentou como os LLMs têm dificuldades em lidar com instruções complexas, especialmente quando a informação relevante está no meio do contexto, um fenômeno descrito por Liu et al. (2023) como *Lost in the Middle*. Estudos como os de He Q. (2024) e Sun et al. (2024) mostram que quanto maior a complexidade das instruções, maior a probabilidade de erro, especialmente quando há múltiplas restrições e diretrizes. A opacidade dos LLMs, frequentemente descritos como “caixa-preta”, é um obstáculo adicional, dificultando a explicação das decisões tomadas e limitando seu uso em áreas críticas como finanças e saúde (BHATTACHARJEE et al., 2024; GAT et al., 2023).

Outro fator relevante é que a inconsistência dos LLMs em classificar sentenças complexas reflete a sensibilidade à estrutura do *prompt*. Jiang et al. (2020) destacam que pequenas variações na geração de *tokens* podem levar a interpretações divergentes, enquanto Wei et al. (2022) observam que o raciocínio encadeado (CoT) pode ser impreciso em modelos menores. No experimento, a classificação de declarações regulatórias evidenciou essa fragilidade: sem um

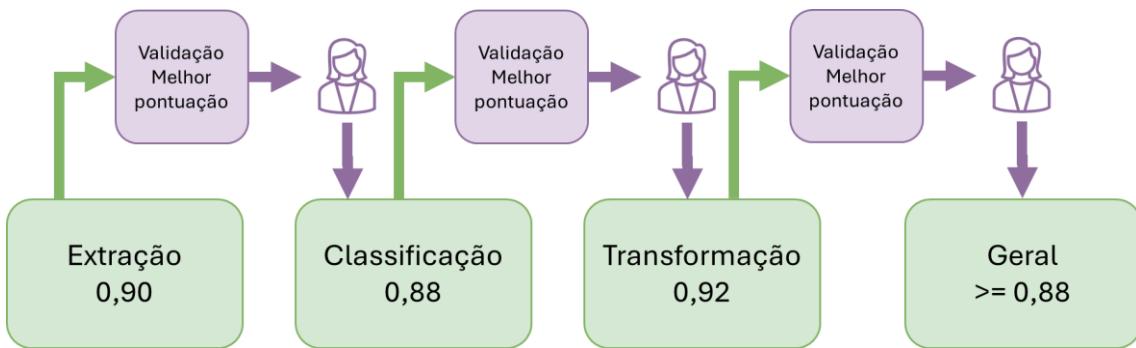
contexto explícito no *prompt* e a redução de complexidade, o modelo interpretava as regras de diferentes formas e por vezes ignorava certos elementos ou alucinava uma definição. Ao reduzir a complexidade e adicionar contexto sobre a aplicação regulatória, a confiança do modelo na classificação aumentou, demonstrando a importância da engenharia de *prompts* e a sensibilidade dos modelos às instruções, como discutido na seção “6.2.3 Anotações semânticas - Discussão”.

Inicialmente projetado para operar de forma totalmente automatizada, o CFR2SBVR foi adaptado metodologicamente com foco na confiabilidade. A decisão de avaliar os processos de forma isolada permitiu demonstrar a independência e a utilidade de cada etapa para os especialistas no assunto (*Subject Matter Experts – SMEs*). A extração automática de elementos do código de regulamentos federais (*Code of Federal Regulations - CFR*) viabilizou a identificação de trechos com diferentes funções, como definições de fatos e regras operacionais. Esse processo contribui para ampliar a interpretação do regulamento e direciona a atenção dos SMEs para os conteúdos mais relevantes. Adicionalmente, a identificação de termos, nomes e suas definições apoia a construção de um vocabulário institucional, essencial para o alinhamento com o domínio, seja em contextos técnicos, como o desenvolvimento de software, ou em contextos organizacionais, como a elaboração de políticas e normas.

A classificação de elementos conforme uma taxonomia alinha a intenção de cada elemento com um significado mais amplo, permitindo entender a intenção e agrupar elementos pelo seu propósito. Funciona, assim, como uma ferramenta de aprendizado, concentrando conhecimento sobre o que a instituição reconhece de importante para aquela classe, como definições, responsabilidades, padrões etc., e conecta esse conhecimento a outros, através de grafos de conhecimento, como proposto nesta pesquisa. Combinado com o processo de transformação, dá-se um passo adiante ao formalizar a forma como as declarações são representadas, permitindo reduzir a carga cognitiva na interpretação pelos SMEs e abrindo a possibilidade de incluir padrões de escrita,

como os discutidos na seção “3.2.2 Linguagem Controlada”, que possam ser interpretados por outros algoritmos, incluindo os de IA.

Figura 38 - CFR2SBVR semi-automático.



Fonte: Elaborado pelo autor.

A automação baseada em modelos probabilísticos, como LLMs, traz desafios inerentes, tais com variância, alucinação e vieses. Esta pesquisa tomou o rumo de preservar a característica automática dos processos, porém dando ênfase na maximização do valor para o SME em cada processo. O ajuste da entrada de cada processo com valores alinhados com a preferência dos SMEs, representado pelos autores, demonstrou que, no geral, o uso do CFR2SBVR como uma ferramenta para os SMEs, como ilustrado na Figura 38, contribui para mitigar os riscos e melhorar o resultado geral, abrindo possibilidades para especializar o modelo em cada processo, incorporando o processo de validação e o conjunto de dados ouro, ou dados sintéticos, no seu ajuste fino.

O ajuste fino é uma forma reduzida de treinamento para alinhar as respostas dos LLMs com a preferência e *feedbacks* humanos. Ele é menor porque utiliza um modelo pré-treinado, mas ainda enfrenta problemas de escassez de dados e os relacionados a participação humana, tais como custos, tempo, vieses, entre outros. Esse é um campo que tem atraído grande interesse para pesquisas com o advento dos LLMs, pois é, atualmente, a principal forma de especializar um LLM para tarefas específicas a um custo relativamente moderado. Praticamente todos os modelos de LLM modernos permitem algum tipo de ajuste fino e existem diversas técnicas para o treinamento, sendo a mais comum a *Reinforcement Learning with Human Feedback* (RLHF).

A técnica *direct-Reinforcement Learning from AI Feedback* (d-RЛАIF), proposto por Lee et al. (2024), é uma técnica de aprendizado por reforço que utiliza *feedback* gerado por um LLM em vez de anotações humanas. Ela elimina a necessidade de treinar um modelo de recompensa, utilizando diretamente um LLM pronto para atribuir pontuações de recompensa durante o aprendizado por reforço. Em vez de depender de um modelo de recompensa pré-treinado, o LLM, em tempo real, avalia respostas geradas com base em critérios como coerência e utilidade. O LLM atribui uma nota (por exemplo, de 1 a 10), que é normalizada em um sinal de recompensa usado para atualizar o modelo de política por meio do aprendizado por reforço.

Os autores demonstraram que o desempenho do *Reinforcement Learning from AI Feedback* (RLAIF) foi igual ou superior ao RLHF (STIENNON, NISSAN et al., 2020) em diversas tarefas, como resumo de textos e geração de diálogos. Em testes com avaliadores humanos, o d-RЛАIF superou tanto o RLHF quanto o RLAIF tradicional, principalmente por evitar problemas de obsolescência do modelo de recompensa e permitir que o *feedback* da IA fosse gerado dinamicamente durante o treinamento.

Ainda nessa linha, Rafailov et al. (2024) introduzem o DPO, que elimina a necessidade de um modelo de recompensa explícito e do RLHF. O problema é tratado como uma tarefa de classificação sobre os dados de preferência humana, utilizando uma função de perda baseada em entropia cruzada binária. Isso resulta em um treinamento mais estável, eficiente e computacionalmente leve. Os experimentos mostram que o DPO alcança ou supera o desempenho do RLHF tradicional em tarefas como análise de sentimento, sumarização e diálogo.

Ethayarajh et al. (2024) apresentam uma nova abordagem para o alinhamento de LLMs baseada na teoria da perspectiva de Kahneman e Tversky. Os autores argumentam que métodos amplamente utilizados, como RLHF e DPO, já incorporam indiretamente vieses cognitivos humanos, como a aversão à perda. Para formalizar essa observação, eles introduzem a classe das *human-aware losses* (HALOs), que considera como os humanos percebem o valor das respostas geradas pelos modelos. Como contribuição principal, os autores propõem o *Kahneman-Tversky Optimization* (KTO), uma nova função de perda

que maximiza diretamente a utilidade das saídas do modelo, em vez de apenas otimizar a probabilidade de preferências observadas. Diferentemente de abordagens baseadas em pares de preferência, o KTO utiliza um sinal binário (desejável / indesejável), reduzindo significativamente os custos de coleta de dados. Os experimentos realizados mostram que o KTO apresenta desempenho igual ou superior ao DPO em modelos variando de 1B a 30B parâmetros, além de oferecer maior estabilidade no treinamento e lidar melhor com conjuntos de dados desbalanceados. A proposta dos autores abre possibilidades para a pesquisa sobre funções de perda inspiradas na economia comportamental e destaca a importância de considerar como os humanos avaliam a qualidade das respostas geradas por modelos de linguagem.

Esta pesquisa não empregou a técnica de ajuste fino, porém os critérios usados para validação das declarações pelo *LLM as a judge* e SemScore constituem uma base para o treinamento de modelos de recompensas e/ou funções de perda, simplificando e melhorando a precisão do CFR2SBVR, com o treinamento de modelos otimizados para cada um dos seus processos, e potencial para aprimorá-lo para a adaptação à diferentes domínios, ficando como trabalhos futuros.

Um dos motivadores desta pesquisa foram os processos de governança, risco e conformidade (*Governance, Risk and Compliance* - GRC) com ênfase no mercado financeiro e a sua falta de suporte para processos automatizados. A formalização de políticas e regulamentos reconhecidamente traria benefícios aos seus processos, porém, como discutido na revisão sistemática da literatura, poucos trabalhos combinam o formalismo do SBVR com a demanda do GRC. Mesmo produtos comerciais não incorporam o seu uso e pouco exploram o potencial de combiná-lo com técnicas avançadas de NLP, como demonstrado nesta pesquisa. Ao combinar o formalismo do SBVR com as técnicas de AI discutidas e demonstradas nesta pesquisa, abre-se um novo conjunto de possibilidade de automação para os processos de GRC.

A integração de técnicas de NLP avançadas para a formalização de regulamentos em SBVR constitui uma das principais contribuições técnicas desta pesquisa. O estudo amplia a compreensão de como métodos

automatizados podem ser aplicados em domínios especializados, como o regulatório, fornecendo uma base teórica para estudos futuros. Do ponto de vista prático, o estudo comprovou a viabilidade do conceito e mostrou que o método proposto é aplicável em instituições financeiras e órgãos reguladores, como ferramenta para acelerar a interpretação de normas. Observa-se, no entanto, que a baixa velocidade das mudanças na CFR e o seu alto valor, contribuem para a adoção de práticas tradicionais na indústria financeira, como a contratação de consultorias especializadas. Mesmo assim, o método e as ferramentas discutidas nesta pesquisa podem ser aplicados como ferramenta para essas consultorias e em outros domínios, como o de contratos, que são mais dinâmicos e com valores relativamente menores.

Explorado por Joshi e Saha (2020), a técnica de adaptação de domínio pode ser aplicada ao CFR2SBVR para adaptá-lo, por exemplo, para os contratos de Kowalsi e Datoo (2022) e Joshi et al. (2021), discutidos na seção “3.3.2 Outras fontes”. A implementação do CFR2SBVR não treinou ou ajustou (fine-tunning) os algoritmos probabilísticos para o domínio financeiro. Ela contou apenas com o conhecimento de NL do modelo pré-treinado, implicando uma solução genérica que pode ser adaptada para outro domínio com relativa facilidade, ficando em aberto como lidar com os dados de entrada, tais como as seções da CFR e parágrafos de um contrato. Os demais elementos da solução, como metodologia, processos, algoritmos, taxonomia e ontologia SBVR podem ser aplicados para outros propósitos.

No contexto da CFR, as posições jurídicas de Hohfeld (1917), discutidas na seção “3.4 Posições jurídicas e modalidades deôntica e aléticas”, não desempenharam nenhum papel relevante para a transformação dos regulamentos avaliados nesta pesquisa, porém, em outros domínios, como os contratos, teria um papel relevante. Uma forma de incorporá-las seria tratá-las como um problema de classificação e associá-las a taxonomia das regras, como a de Witt (2012), e ao processo de classificação.

Quanto à validade interna deste estudo, que trata da relação causal entre as intervenções realizadas e os resultados obtidos, a principal ameaça é a dependência do conjunto de dados ouro utilizado na validação. A presença de

possíveis lacunas ou vieses na representação dos regulamentos poderia comprometer a avaliação do modelo. Para mitigar essa ameaça, o conjunto de dados foi revisado iterativamente e os resultados foram comparados com os resultados gerados pelo modelo.

No caso da validade externa, a limitação do escopo dos regulamentos analisados representou um desafio significativo. O estudo concentrou-se em três seções da CFR. Como mitigação, foi adotada a especificação SBVR, assegurando um potencial transferência para outros domínios, descrita por Joshi et al. (2021).

Quanto à validade de construto, uma ameaça é a possibilidade de que métricas quantitativas, como a similaridade semântica, não capturassem integralmente a semântica dos regulamentos transformados. Para enfrentar esse problema, a validação do modelo combinou métricas quantitativas e análises qualitativas automatizadas, como a avaliação crítica da semântica das regras transformadas, em que o modelo de linguagem foi instruído a justificar suas avaliações, posteriormente analisadas pelo autor.

A confiabilidade, definida por Wieringa (2014) como a capacidade de reproduzir os resultados sob condições semelhantes, foi desafiada pela natureza estocástica do modelo empregado. Para mitigar essa ameaça, foram realizadas múltiplas execuções para o mesmo conjunto de dados, permitindo avaliar a estabilidade dos resultados. Além disso, o código-fonte e os dados utilizados na pesquisa foram disponibilizados publicamente, para assegurar a replicabilidade e a possibilidade de verificação independente dos achados.

Para posicionar os resultados desta pesquisa em relação aos trabalhos da seção “4.2 Abordagens de transformação de NLP para SBVR”, foram selecionados os trabalhos que apresentaram resultados, mesmo que parciais, embora nenhum deles disponibilize todos os elementos necessários para replicá-los, tais como conjuntos de dados e algoritmos. Esse cenário motivou a criação de um conjunto de dados exclusivo para esta pesquisa, que pode ser utilizado em futuras comparações. Apesar das limitações, os fragmentos apresentados nos trabalhos selecionados permitiram construir um quadro comparativo preliminar e embasar as conclusões desta pesquisa com os dos autores discutidos a seguir.

Abi-Lahoud et al. (2013) desenvolveram uma prova de conceito para a interpretação de regulamentos utilizando SBVR, aplicada às regulações de *Anti-Money Laundering* (AML) dos EUA. Identificaram e formalizaram 300 conceitos de substantivos e cerca de 200 regras comportamentais a partir de seções específicas do registro federal dos EUA (*Federal Register - FR*). A abordagem combinou a competência de especialistas de domínio com ferramentas como Confluence e Designs for Management™ para validação e exportação dos resultados. Apesar de não apresentarem métricas quantitativas, destacaram a consistência e completude das formalizações. As limitações incluíram dependência de interpretações manuais e escopo restrito, enquanto os trabalhos futuros sugeriram automatizar transformações em ontologias OWL e modelos de dados.

Chittimalli et al. (2020) propuseram uma abordagem não supervisionada para extrair vocabulários e regras SBVR de documentos empresariais. Testada em dois conjuntos de dados, EU-Rent-A-Car (64 regras) e KYC (630 sentenças, das quais 185 irrelevantes), a metodologia apresentou ganhos significativos, como aumentos de precisão entre 7-20% e reduções de tempo de processamento em 25%. A precisão foi de 97,83% para entidades e 83,69% para regras no KYC. Contudo, a abordagem não cobre regras implícitas, correferências e idiomas além do inglês, limitando sua generalização.

Haj et al. (2021) apresentaram uma abordagem automática para transformar regras textuais em modelos SBVR completos, avaliando 152 sentenças divididas em preferências de clientes, reservas de voos e empréstimos. Obtiveram uma F1 média de 87%, com precisão entre 89-96% para conceitos nominais e 100% para negações. Comparado a trabalhos anteriores, a abordagem mostrou-se mais completa, incorporando definições e relações implícitas. Contudo, a dependência de sentenças bem-formadas e a falta de validações em outros idiomas foram destacadas como limitações. Trabalhos futuros sugerem suporte a formulações gramaticais mais complexas e testes com dados maiores.

Joshi e Saha (2020) desenvolveram um framework semântico para representar o conhecimento do CFR, aplicado ao Título 48. A abordagem extraiu 9.084 expressões deônticas, organizando-as em um grafo de conhecimento que incluiu

permissões, obrigações, proibições e dispensas. Apesar da ausência de métricas quantitativas, a eficácia foi validada qualitativamente. As limitações incluíram dependência de textos estruturados e ausência de priorização de regras. O desenvolvimento futuro abrange suporte a todos os títulos do CFR e hierarquias entre regras.

Roychoudhury et al. (2017) propuseram um framework para transformar textos regulatórios em modelos SBVR utilizando *Structured English* (SE) como intermediário. Validada em subconjuntos de MiFID II e KYC, a abordagem reduziu o tempo de modelagem de 71 para 9 horas em 187 regras do KYC. Embora eficiente em termos de tempo e usabilidade, carece de métricas quantitativas detalhadas e validações em outros domínios regulatórios. Melhorias futuras incluem suporte a modalidades mais complexas e integração com sistemas de conformidade existentes.

Esta pesquisa destaca-se pela inovação no uso de LLMs, que representa uma evolução em relação a abordagens anteriores, como as de Roychoudhury et al. (2017) e Haj et al. (2021), que usaram linguagens intermediárias e técnicas manuais para criar modelos SBVR. Esta abordagem apresenta uma avaliação abrangente, combinando métricas quantitativas (precisão, F1-score, sensibilidade) e qualitativas (preservação semântica com SemScore e *LLM as a judge*), o que o torna a avaliação metodologicamente mais completa do que alguns trabalhos que focalizam apenas a avaliação qualitativa, como o de Joshi e Saha (2020). Os resultados quantitativos indicam alta precisão, com acurácia média de 95,81% na transformação para SBVR, equiparando aos resultados de trabalhos como os de Chittimalli et al. (2020), que apresentaram precisão entre 89% e 97%.

Quadro 34 - Resultado dos autores.

| <b>Autores</b>            | <b>Precisão<br/>(regras def.)</b> | <b>Precisão<br/>(regras op.)</b> | <b>Conj. dados</b>                           | <b>Qde Extraídas /<br/>Avaliadas</b>                    | <b>Tempo de<br/>Execução</b>              |
|---------------------------|-----------------------------------|----------------------------------|----------------------------------------------|---------------------------------------------------------|-------------------------------------------|
| Esta pesquisa             | 85%-95%                           | 90%-95%                          | Título 17, Cap. II, Parte 275. CFR           | 22 sentenças, 85 termos / nomes, 16 fatos, 6 regras op. | 4,38 min (262,83s)                        |
| Chittimalli et al. (2020) | 89%-97%                           | 84%                              | Conjuntos de teste simulados (KYC e EU-Rent) | KYC: 548 entidades, 350 fatos, 638 regras. EU-Rent: 277 | EURent: 3 min (181s), KYC: 10 min (597s). |

| Autores                       | Precisão<br>(regras def.) | Precisão<br>(regras op.) | Conj. dados                   | Qde Extraídas /<br>Avaliadas                                                                            | Tempo de<br>Execução                          |
|-------------------------------|---------------------------|--------------------------|-------------------------------|---------------------------------------------------------------------------------------------------------|-----------------------------------------------|
|                               |                           |                          |                               | entidades, 177<br>fatos e 64 regras                                                                     |                                               |
| Roychoudhury<br>et al. (2017) | Não<br>informado          | Não<br>informado         | MiFID II e KYC                | 187 regras (KYC)                                                                                        | 9h (KYC com<br>SE)                            |
| Abi-Lahoud et<br>al. (2013)   | Não<br>informado          | Não<br>informado         | 76 FR 45403 -<br>18 páginas   | 300 substantivos<br>200 regras<br>comportamentais                                                       | Não<br>especificado<br>(totalmente<br>manual) |
| Haj et al.<br>(2021)          | 89%-96%                   | 67%-95%                  | Extraído de<br>Livros 1, 2, 3 | 152 sentenças<br>(preferências,<br>reservas e<br>emprestimos)                                           | Não<br>especificado<br>(automatizado)         |
| Joshi e Saha<br>(2020)        | Não<br>informado          | Não<br>informado         | Título 48 CFR                 | 9.084 expressões<br>deônica (710<br>permissões, 698<br>obrigações, 479<br>proibições, 149<br>dispensas) | Não informado                                 |

Fonte: Elaborado pelo autor.

O Quadro 34 resume os principais resultados discutidos nesta seção. São destacadas as métricas de precisão, quantidades avaliadas e tempos de execução dos trabalhos revisados no capítulo “4 DISCUSSÃO DAS SOLUÇÕES”, para uma visão comparativa de cada abordagem.

Os resultados apresentados nesta seção demonstram que o CFR2SBVR alcançou precisão e tempo de execução similares aos dos outros trabalhos. Por não ser baseado em regras, ele é mais facilmente adaptável a mudanças; em muitos casos basta modificar as instruções dos *prompts*, porém, dadas as características probabilística e de caixa-preta dos LLMs, os resultados não são sempre consistentes e difíceis de justificar. Múltiplas execuções podem mitigar esse problema, escolhendo a resposta com maior frequência e melhor avaliação. Contudo, a completa automação, excluindo o ser humano no processo (*human-in-the-loop*), não foi um objetivo perseguido, devido aos riscos já discutidos e o entendimento de que modelos de IA, por mais sofisticados que sejam, ainda precisam de supervisão humana, seja no seu treinamento, para alinhar as preferências, ou na inferência, na qual seus resultados podem divergir da preferência devido à sua natureza probabilística, vieses e ou falhas. A supervisão humana pode ser entendida como um esforço conjunto, em que o julgamento humano e do modelo são combinados para obter um melhor resultado.

## 8.2 Conclusão

A pergunta central desta pesquisa foi “como a aplicação de técnicas de NLP e ontologias, como a FIBO, podem aprimorar a transformação automatizada de regulamentos financeiros em regras e vocabulário estruturados, conforme a especificação SBVR”. O experimento CFR2SBVR, discutida no capítulo “6 FERRAMENTAS DE SUPORTE” demonstrou como aplicar essas técnicas e mostrou-se viável técnica e economicamente, com pontuações de precisão, acurácia, sensibilidade e F1 score acima de 0,85 em média, comparáveis aos dos autores estudados e custo por execução de aproximadamente USD 25.00, com tempo médio de aproximadamente 4 minutos.

A pesquisa revelou que a associação dos elementos com as ontologias FRO e FIBO tiveram um papel secundário na mecânica da transformação, apresentada na seção “6.2.4 Transformação - Algoritmo “transformação para SBVR”, no entanto, ela cumpriu a tarefa de melhorar o entendimento das regras, preservando suas relações e a desambiguação para os SMEs, relacionando os termos a vocabulários do mercado e identificando lacunas nas definições a serem preenchidas por outros vocabulários ou pelos SMEs.

No caso da ontologia FRO, sua exclusão do processo foi devido aos erros na importação e atualização da CFR (Quadro 11). A solução foi utilizar o modelo para interpretar as partes do texto regulatório, com parágrafos e referências a outras seções, o que foi realizado com sucesso às custas do aumento da complexidade dos *prompts*, apresentado no Quadro 12. Para a ontologia FIBO, foi identificado um desalinhamento entre os termos e nomes extraídos pelo CFR2SBVR e os termos na FIBO. Foram encontrados 62% de termos similares (*closeMatch*) e 14% de correspondências exatas (*exactMatch*), e para o restante dos 22% dos termos nenhuma similaridade foi encontrada. A expectativa era que a busca na FIBO (Figura 20) tivesse uma taxa maior de correspondências exatas, dado a sua origem nos textos regulatórios do setor financeiro.

No entanto, a baixa taxa de localização dos termos e nomes na FIBO teve baixo impacto no resultado das transformações, devido ao fato de o modelo não precisar conhecer o significado do termo para “entender” o seu papel na

declaração e ser capaz de classificá-los ou mapeá-los para os *templates* de transformação. No entanto, para o SME, o relacionamento com a FIBO ou outros vocabulários pode auxiliar na definição de termos, fornecendo um ponto de partida para alinhar vocabulários. Nesse sentido, essa baixa taxa de sucesso traz prejuízos para o processo.

O CFR2SBVR como ferramenta de suporte para os SMEs cobre todos os processos, de extração à validação, guiando o modelo de linguagem com um processo rígido, auditável e estável, sem prejudicar a capacidade de generalização, uma vez que as tarefas de identificação, classificação e geração de texto estruturado em SBVR são facilmente adaptáveis com engenharia de *prompt*, como realizado nesta pesquisa, ou dando um passo a diante com a especialização de modelos utilizando técnicas de ajuste fino.

Como discutido nas seções “7.2.4 Conjunto de dados” e no capítulo “7 VALIDAÇÃO”, foi necessário a criação de um conjunto de dados para a validação. No entanto, esse conjunto de dados é limitado.

Em relação as pontuações dos resultados, eles foram obtidos ajustando a entrada dos processos para o conjunto de dados padrão ouro com o intuito de reduzir o acúmulo de erros e simular a intervenção humana / IA no processo. Sem essa intervenção, a precisão geral poderia ser significativamente mais baixa, porém, ao incorporar a avaliação como parte do processo e a validação dos SMEs, o potencial de alcançar pontuações mais elevadas é factível, como discutido na seção “7.3.2 semantic annotation”.

Em setores que impactam diretamente a vida de seus usuários, como o financeiro, a obtenção de boa pontuação de desempenho não é suficiente, uma vez que o impacto negativo de um único erro pode ser significativo tanto em termos sociais, econômicos e de credibilidade para os usuários do sistema financeiro. Embora o CFR2SBVR tenha sido projetado como um processo completamente automatizado, seu uso como ferramenta de apoio para os SMEs, pode não apenas mitigar esse risco, mas também aumentar a confiabilidade dos processos de GRC. Além disso, os resultados intermediários obtidos nesta pesquisa — incluindo a decomposição do texto regulatório em elementos, sua

classificação conforme uma taxonomia e a associação entre os elementos e com o texto original — podem auxiliar os SMEs na compreensão da composição, do propósito e da intenção dos textos normativos e seus relacionamentos.

Além do desenvolvimento do processo e das ferramentas, esta pesquisa contribui com uma análise da situação atual do emprego do SBVR, que embora presente nas discussões acadêmicas, respeitando as limitações desta pesquisa, é inexpressiva na indústria. Demonstrou-se o potencial do uso do SBVR e a sua viabilidade, contribuindo para aproximar os trabalhos acadêmicos aos da indústria. A pesquisa também disponibiliza um conjunto de dados e um processo de avaliação que podem ser utilizados para comparações com estudos futuros. Para garantir a reproduzibilidade, todos os dados e códigos necessários para a replicação dos experimentos foram disponibilizados.

### 8.3 Trabalhos futuros

Com base nos resultados obtidos, há diversas direções possíveis para a continuidade desta pesquisa.

Na seção “3.1 Abordagens formais de verificação de conformidade”, foi apresentado um arcabouço conceitual em que esta pesquisa se concentrou nas atividades de identificação e extração por serem pré-requisitos. Destacou-se ainda a integração de abordagens formais de verificação de conformidade em ferramentas comerciais. Um estudo não exaustivo das soluções de mercado mostrou que elas empregam IA para resolver problemas mais simples do que os propostos pela academia. Para trabalhos futuros, com os resultados desta pesquisa, expandir o desenvolvimento das demais atividades do arcabouço da Figura 5 e conduzir estudos estruturados com fornecedores de soluções de GRC para alinhar os desafios enfrentados pela academia com soluções aplicadas na indústria.

O SBVR possui um metamodelo ontológico que define conceitos, relações e regras de negócio de forma estruturada, baseado na lógica modal e de predicados e na ISO 1087. Ele organiza vocabulários e regras de negócio de maneira formal, garantindo precisão semântica, mas não é nativamente

representado em formatos como OWL ou RDF. No entanto, pode ser mapeado para ontologias formais para interoperabilidade com sistemas semânticos, como discutido na seção “5.2.8 Ontologia SBVR”. Sua integração com a ontologia MVF ou grafos conceituais, como exposto nesta pesquisa na seção “3.2.5 Modelo de conteúdo”.

Como discutido nesta seção, não existe um padrão de jure para a classificação das regras de negócio, sendo a sua escolha dependente do propósito. Nesta pesquisa, o objetivo é identificar padrões de regras e associá-los a *templates* de escrita a fim de obter um resultado consistente na sua formulação, em que a taxonomia de Witt (2012), ao combinar a classificação e um *template* de escrita, atende a esse propósito.

“3.2.5 Modelo de conteúdo”, permitiria representar termos em múltiplos idiomas e terminologias, facilitando a inferência, a comunicação e o compartilhamento de conhecimento entre diferentes domínios e sistemas.

Ainda sobre SBVR, a exploração da sinonímia e dos conceitos verbais pode ser aprimorada com o uso de LLMs. O uso de grafos para relacionar essas informações com termos e declarações poderá revelar novas relações e significados que vão além da extração e transformação. Essa abordagem pode contribuir para uma representação mais rica e estruturada do conhecimento, aprimorando a interpretação das regras de negócio.

No contexto da transformação de regras financeiras (CFR), as posições jurídicas formuladas por Hohfeld (1917), conforme discutido na “3.4 Posições jurídicas e modalidades deôntica e aléticas”, não exerceram um papel significativo na conversão do conjunto de regulamentos analisado nesta pesquisa. Entretanto, em investigações futuras, sua incorporação poderia ser viabilizada por meio da integração com taxonomias de regras, como a proposta por Witt (2012), possibilitando uma interpretação mais detalhada da intenção do texto normativo.

No que se refere à modelagem de dados, Abi-Lahoud (2013) destaca a relevância do desenvolvimento de ontologias e modelos baseados em vocabulários SBVR, especialmente no setor financeiro, com o objetivo de

aprimorar a automação e reduzir a dependência de especialistas humanos. A construção de uma ontologia específica para SBVR permitiria a exploração das capacidades inferenciais de grafos de conhecimento, possibilitando a descoberta de novas conexões e aprofundando a compreensão das relações semânticas entre os dados regulatórios.

Sistemas multiagentes e novos modelos, treinados para raciocínios complexos, como OpenAI o3, DeepSeek R1, Anthropic Claude 3.5 Opus e Llama 3, ampliam as possibilidades de melhoria na compreensão das nuances de NL.

A especialização de modelos para cada processo descrito na seção “6.2 Implementação dos principais componentes”. Também simplificam e melhoram a precisão geral da solução ao incorporar, no treinamento do modelo, os critérios de validação. Essa especialização pode ser obtida por ajuste fino dos modelos usando técnicas como RLAIF e RFT que pretende mitigar a escassez de dados para treinamento, bem como técnicas mais recentes como GRPO, DPO e KTO, que aceleram o processo de treinamento e reduzem os custos ao eliminarem a necessidade do modelo de recompensa, mantendo o alinhamento com a preferências dos SMEs.

A incorporação da etapa de validação no *pipeline* do processo de transformação CFR2SBVR pode ajudar a lidar com a natureza probabilística do modelo, permitindo a estratificação dos resultados com maior pontuação antes de submetê-los aos SMEs para *feedback*. Isso contribui para um uso mais eficiente do tempo, além de melhorar a precisão e a padronização dos resultados. O LLM, além de justificar suas avaliações, pode atuar em um ciclo de *feedback* com os SMEs, sugerindo melhorias nas declarações para torná-las mais comprehensíveis, alinhadas aos vocabulários do mercado e corporativo e semanticamente equivalentes. Para viabilizar essa interação, é essencial o desenvolvimento de interfaces gráficas que permitam aos SMEs interagir com o LLM em todas as etapas do processo e registrar seus *feedbacks*.

Por fim, investigar a sinergia entre abordagens voltadas ao alinhamento entre conceitos de negócio, como SBVR e *Domain-Driven Design* de Eric Evans<sup>70</sup> e as formas como podem contribuir para a solução de desafios na integração de sistemas. Dentre esses desafios, destacam-se a heterogeneidade semântica, que se refere às diferenças no significado e na interpretação dos dados entre sistemas, resultando em inconsistências, e a heterogeneidade estrutural, caracterizada por discrepâncias na organização dos dados, falta de padronização e baixa qualidade das informações (SCHMIDT; OTTO; ÖSTERLE, 2010). Esse desalinhamento semântico aumenta a complexidade e os custos da integração, demandando soluções sofisticadas e ferramentas avançadas de governança e integração de dados. Nesse contexto, abordagens como o CFR2SBVR oferecem um caminho para o alinhamento semântico, conforme discutido por Sunkle *et al.* (2015a), e contribuem para uma governança mais eficiente.

Ao investir em automação e ferramentas, como CFR2SBVR, as instituições financeiras podem alcançar um repositório digital de regras regulatórias, facilitando sua atualização, interpretação e operacionalização. Isso não apenas contribui para a conformidade com regulamentos atuais, mas também proporciona maior agilidade para se adaptar a novas legislações. No longo prazo, a sinergia entre NLP baseada em LLMs e a estrutura formal do SBVR, através das técnicas e métodos elaborados na academia, representa um caminho promissor para gerenciar a crescente complexidade regulatória de maneira sistemática e transparente. Esse processo exige rigor e a combinação adequada de tecnologia e competência humana, mas os benefícios resultam em um programa de conformidade robusto e inteligível, transformando mandatos legais em regras de negócios acionáveis com segurança e eficiência.

---

<sup>70</sup> <https://www.domainlanguage.com/>. Acesso em: 15/mar./2025.

## REFERÊNCIAS

360 MARKET UPDATES. **CAGR of 15.42%** - Governance, Risk Management and Compliance (GRC) Market 2023 Is Set to Fly Excessive Growth in Years to Come. 2023. Disponível em: <https://www.linkedin.com/pulse/cagr-1542-governance-risk-management-compliance-grc-market/>. Acesso em: 10 mar. 2024.

ABDULLAH, Hanifa. Analyzing the technological challenges of Governance, Risk and Compliance (GRC). In: 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT). IEEE, p. 274-282, 2019

ABI-LAHOUD, Elie *et al.* Interpreting Regulations with SBVR, in. Collaboration Technologies and Systems (CTS), IEEE, p. 439-445, Maio 2013.

ABIDIN, Nor Najihah Zainal; A. MANAF, Nurulhuda. “OR” of Rule-Based Specification for Service Choreography. In: **International Conference on Services Computing**. Cham: Springer Nature Switzerland, 2023. p. 3-15

ABU-SALIH, Bilal. Domain-specific knowledge graphs: A survey. **Journal of Network and Computer Applications**, [S.I.], v. 185, p. 103076, jul. 2021.

AGNOLONI, T.; FRANCESCONI, E.. Modelling semantic profiles in legislative documents for enhanced norm accessibility. **Proceedings Of The 13Th International Conference On Artificial Intelligence And Law**, [S.I.], p. 111-115, 6 jun. 2011. ACM.

AGRAWAL, Garima; KUMARAGE, Tharindu; ALGHAMDI, Zeyad; LIU, Huan. Can Knowledge Graphs Reduce Hallucinations in LLMs?: a survey. **Proceedings Of The 2024 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies** (Volume 1, [S.I.], v. 1, p. 3947-3960, jun. 2024. Association for Computational Linguistics.

AIT-MLOUK, Addi; JIANG, Lili. KBot: a knowledge graph based chatbot for natural language understanding over linked data. **ieee Access**, [S.I.], v. 8, p. 149220-149230, ago. 2020. Institute of Electrical and Electronics Engineers (IEEE).

AKEHURST, D. H.; BORDBAR, B.; EVANS, M. J.; HOWELLS, W. G. J.; McDONALD-MAIER, K. D.. SiTra: simple transformations in java. **Lecture Notes In Computer Science**, [S.I.], v. 4199, p. 351-364, 2006. Springer Berlin Heidelberg.

ALEXANDER, Christopher. **The timeless way of building**. New York: Oxford university press, 1979.

AMARAL, Glenda; SALES, Tiago Prince; GUIZZARDI, Giancarlo. Towards an Ontology Network in Finance and Economics. *In: Enterprise Engineering Working Conference*. Cham: Springer International Publishing, 2021. p. 42-57

ANBIMA. **Mercado de capitais mantém trajetória de recuperação e chega a R\$ 229 bilhões em emissões no ano.** 2023. Disponível em: [https://www.anbima.com.br/pt\\_br/informar/relatorios/mercado-de-capitais/boletim-de-mercado-de-capitais/mercado-de-capitais-mantem-trajetoria-de-recuperacao-e-chega-a-r-229-bilhoes-em-emissoes-no-ano.htm](https://www.anbima.com.br/pt_br/informar/relatorios/mercado-de-capitais/boletim-de-mercado-de-capitais/mercado-de-capitais-mantem-trajetoria-de-recuperacao-e-chega-a-r-229-bilhoes-em-emissoes-no-ano.htm). Acesso em: 22 out. 2023.

ANH, Dang Hoang *et al.* The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey. *In: 2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, p. 1-7, 2023.

ASHFAQ, Fariha; BAJWA, Imran Sarwar. Natural language ambiguity resolution by intelligent semantic annotation of software requirements. **Automated Software Engineering**, [S.I.], v. 28, n. 2, p. 13-13, 21 jul. 2021. Springer Science and Business Media LLC.

ASHOK, Dhananjay; LIPTON, Zachary C. PromptNER: Prompting for Named Entity Recognition. **[S.I.: s.n.]**, v. 1, n. 1, p. 1-11, maio 2023.

ATTARD, Judie; ORLANDI, Fabrizio; AUER, Sören. Data Driven Governments: creating value through open government data. **Lecture Notes In Computer Science**, [S.I.], p. 84-110, set. 2016. Springer Berlin Heidelberg.

AYNETDINOV, Ansar; AKBIK, Alan. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. **[S.I.: s.n.]**, v. 1, n. 1, p. 1-11, p. 1-9, fev. 2024.

**B3. Mercado de renda variável movimenta R\$ 7,46 trilhões na B3 em 2022.** 2023a. Disponível em: [https://www.b3.com.br/pt\\_br/noticias/mercado-de-renda-variavel-movimenta-r-7-46-trilhoes-na-b3-em-2022.htm](https://www.b3.com.br/pt_br/noticias/mercado-de-renda-variavel-movimenta-r-7-46-trilhoes-na-b3-em-2022.htm). Acesso em: 21 out. 2023.

**B3. Ofertas públicas.** 2023b. Disponível em: [https://www.b3.com.br/pt\\_br/produtos-e-servicos/solucoes-para-emissores/ofertas-publicas/estatisticas/](https://www.b3.com.br/pt_br/produtos-e-servicos/solucoes-para-emissores/ofertas-publicas/estatisticas/). Acesso em: 21 out. 2023.

**B3. Valor de mercado das empresas listadas.** 2023c. Disponível em: [https://www.b3.com.br/pt\\_br/market-data-e-indices/254elatório-de-dados/market-data/consultas/mercado-a-vista/valor-de-mercado-das-empresas-listadas/bolsa-de-valores-diario/](https://www.b3.com.br/pt_br/market-data-e-indices/254elatório-de-dados/market-data/consultas/mercado-a-vista/valor-de-mercado-das-empresas-listadas/bolsa-de-valores-diario/). Acesso em: 22 out. 2023.

BAJWA, Imran Sarwar; LEE, Mark G.; BORDBAR, Behzad. SBVR business rules generation from natural language specification. *In: 2011 AAAI Spring Symposium Series*. 2011.

BAJWA, Imran Sarwar; SHAHZADA, Muhammad Anwar. Automated generation of OCL constraints: NL based approach vs pattern based approach. **Mehran University Research Journal of Engineering & Technology**, [S.I.], v. 36, n. 2, p. 243-254, abr. 2017.

BENNETT, Michael G.. Standards for Knowledge Graphs in the Financial Sector. **Lecture Notes In Networks And Systems**, [S.I.], p. 717-729, 24 out. 2021.

BERNOTAITYTE, Gintare *et al.* Developing SBVR Vocabularies and Business Rules from OWL2 Ontologies. **Communications In Computer And Information Science**, [S.I.], p. 134-145, 2013. Springer Berlin Heidelberg.

BHAROSA, Nitesh *et al.* Towards a lean-government using new IT-architectures for compliance monitoring. **Proceedings Of The 5Th International Conference On Theory And Practice Of Electronic Governance**, [S.I.], p. 147-156, 26 set. 2011. ACM.

BHATTACHARJEE, Amrita *et al.* Towards IIm-guided causal explainability for black-box text classifiers. *In: AAAI 2024 Workshop on Responsible Language Models, Vancouver, BC, Canada*, p. 1-6, jan. 2024.

BHATTACHARYYA, Abhidip *et al.* An Approach to Mine Business Rule Intents from Domain-specific Documents. **Proceedings Of The 10Th Innovations In Software Engineering Conference**, [S.I.], p. 96-106, 5 fev. 2017. ACM.

BIOCHINI, Jorge *et al.* Systematic review in software engineering. **System engineering and computer science department COPPE/UFRJ, Technical Report ES**, Rio de Janeiro, v. 679, n. 05, p. 45, maio 2005.

BOELLA, Guido *et al.* Managing legal interpretation in regulatory compliance. **Proceedings Of The Fourteenth International Conference On Artificial Intelligence And Law**, [S.I.], p. 23-32, 10 jun. 2013. ACM.

BOUZIDI, Khalil Riad *et al.* An Ontological Approach for Modeling Technical Standards for Compliance Checking. **Lecture Notes In Computer Science**, [S.I.], p. 244-249, 2011. Springer Berlin Heidelberg.

CABOT, Jordi *et al.* From UML/OCL to SBVR specifications: a challenging transformation. **Information Systems**, [S.I.], v. 35, n. 4, p. 417-440, jun. 2010. Elsevier BV.

CÁMARA, Javier *et al.* On the assessment of generative AI in modeling tasks: an experience report with chatgpt and uml. **Software And Systems Modeling**, [S.I.], v. 22, n. 3, p. 781-793, 22 maio 2023. Springer Science and Business Media LLC.

CASELLAS, Núria *et al.* From Legal Information to Open Legal Data: a case study in u.s. federal legal information. **Ssrn Electronic Journal**, [S.I.], p. 1-5, set. 2011. Elsevier BV.

CASELLAS, Nuria *et al.* Linked legal data. **Proceedings Of The 13Th Annual International Conference On Digital Government Research**, [S.I.], p. 280-281, 4 jun. 2012. ACM.

CHANG, Yupeng *et al.* A Survey on Evaluation of Large Language Models. **Acm Transactions On Intelligent Systems And Technology**, [S.I.], v. 15, n. 3, p. 1-45, 29 mar. 2024. Association for Computing Machinery (ACM).

CHEN, Kua *et al.* Automated Domain Modeling with Large Language Models: A Comparative Study. *In:* 2023 ACM/IEEE 26th International Conference on Model Driven Engineering Languages and Systems (MODELS). **IEEE**, p. 162-172, 2023.

CHITTIMALLI, Pavan Kumar *et al.* BuRRiT0: a framework to extract, specify, verify and analyze business rules. *In:* 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE). **IEEE**, p. 1190-1193, 2019a.

CHITTIMALLI, Pavan Kumar; BHATTACHARYYA, Abhidip. Sbvr-based business rule creation for legacy programs using variable provenance. *In:* **Proceedings of the 12th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference)**. 2019b. p. 1-11.

CHITTIMALLI, Pavan Kumar *et al.* An Approach to Mine SBVR Vocabularies and Rules from Business Documents. **Proceedings Of The 13Th Innovations In Software Engineering Conference On Formerly Known As India Software Engineering Conference**, [S.I.], p. 1-11, 27 fev. 2020. ACM.

CIO.gov, U.S. Chief Information Offices Council. **Open Government Data Act (2018)**. 2019. Disponível em: <https://www.cio.gov/handbook/it-laws/ogda/>. Acesso em: 14 nov. 2023.

COMBEMALE, Benoit *et al.* ChatGPT in software modeling. **Software And Systems Modeling**, [S.I.], v. 22, n. 3, p. 777-779, 11 maio 2023. Springer Science and Business Media LLC.

CONGRESS, U. S. **Dodd-Frank Wall Street reform and consumer protection act**. 2010. Disponível em: <https://www.congress.gov/111/plaws/publ203/PLAW-111publ203.pdf>. Acesso em: 16 mar. 2025.

CFR. **How is the eCFR updated?** 2024a. Disponível em: <https://www.ecfr.gov/reader-aids/understanding-the-ecfr/how-is-the-ecfr-updated>. Acesso em: 03 abr. 2024.

CFR. **Code of Federal Regulations (CFR), 1996 to Present**. 2024b. Disponível em: <https://www.govinfo.gov/help/cfr>. Acesso em: 03 abr. 2024

DANENAS, Paulius *et al.* Natural language processing-enhanced extraction of SBVR business vocabularies and business rules from UML use case

diagrams. **Data & Knowledge Engineering**, [S.I.], v. 128, p. 101822, jul. 2020. Elsevier BV.

DE OLIVEIRA RODRIGUES, Cleyton Mario et al. Legal ontologies over time: A systematic mapping study. **Expert Systems with Applications**, v. 130, p. 12-30, 2019.

DONG, Yijiang River; HU, Tiancheng; COLLIER, Nigel. Can LLM be a Personalized Judge? **Findings Of The Association For Computational Linguistics: EMNLP 2024**, [S.I.], p. 10126-10141, nov. 2024. Association for Computational Linguistics.

DRAGONI, Mauro et al. Combining Natural Language Processing Approaches for Rule Extraction from Legal Documents. **Lecture Notes In Computer Science**, [S.I.], p. 287-300, 2018. Springer International Publishing.

EDMC (org.). **The Financial Industry Business Ontology**: the global standard ontology for efficient and unambiguous financial services. 2024. Disponível em: <https://spec.edmcouncil.org/fibo/>. Acesso em: 03 abr. 2024.

ETHAYARAJH, Kawin et al. Kto: Model alignment as prospect theoretic optimization. **arXiv preprint arXiv:2402.01306**, 2024.

FAROUK, Mamdouh et al. Measuring Sentences Similarity: a survey. **Indian Journal Of Science And Technology**, [S.I.], v. 12, n. 25, p. 1-11, 1 jul. 2019. Indian Society for Education and Environment.

FENSEL, Dieter et al. Introduction: what is a knowledge graph?. **Knowledge Graphs**, [S.I.], p. 1-10, 2020. Springer International Publishing.

FERREIRA, Fernando; LI, Jennie; NOSSIG, Rebecca. **Raio-XP**: retrospectiva 2022. Retrospectiva 2022. 2023. Disponível em: <https://conteudos.xpi.com.br/acoes/relatorios/raio-xp-retrospectiva-2022/>. Acesso em: 21 out. 2023.

FILTZ, Erwin. Building and Processing a Knowledge-Graph for Legal Data. **Lecture Notes In Computer Science**, [S.I.], p. 184-194, maio 2017. Springer International Publishing.

FINRA. **Rule Filings**. 2023. Disponível em: <https://www.finra.org/rules-guidance/rule-filings>. Acesso em: 22 out. 2023.

FORD, Reginald et al. Automating Financial Regulatory Compliance Using Ontology+Rules and Sunflower. **Proceedings Of The 12Th International Conference On Semantic Systems**, [S.I.], p. 113-120, 12 set. 2016. ACM.

FRANCESCONI, Enrico; GOVERNATORI, Guido. Patterns for legal compliance checking in a decidable framework of linked open data. **Artificial Intelligence And Law**, [S.I.], v. 31, n. 3, p. 445-464, 4 jul. 2022. Springer Science and Business Media LLC.

FRANZ INC. **AllegroGraph**. Disponível em:  
<https://allegrograph.com/products/allegrograph/>. Acesso em: 21 set. 2024.

GAMMA, Erich. **Design patterns: elements of reusable object-oriented software**. Pearson Education India, 1995.

GAT, Yair *et al.* Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals. [S.I.: s.n.], p. 1-34, nov. 2023.

GAVA, Vagner Luiz. **Metodologia de pesquisa científica**: Aula 01. São Paulo: Instituto de Pesquisas Tecnológicas do Estado de São Paulo, 2022.

GOEBEL, Randy *et al.* Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. **Proceedings Of The Nineteenth International Conference On Artificial Intelligence And Law**, [S.I.], p. 472-480, 19 jun. 2023. ACM.

GOOGLE. **Google Colaboratory**. Disponível em:  
<https://colab.research.google.com/>. Acesso em: 21 set. 2024.

GOVERNATORI, Guido; ROTOLI, Antonino. Deontic Ambiguities in Legal Reasoning. **Proceedings Of The Nineteenth International Conference On Artificial Intelligence And Law**, [S.I.], p. 91-100, 19 jun. 2023. ACM.

GOU, Zhibin *et al.* Critic: Large language models can self-correct with tool-interactive critiquing. [S.I.: s.n.], p. 1-78, fev. 2023.

GREENHALGH, Trisha; PEACOCK, Richard. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. **Bmj**, [S.I.], v. 331, n. 7524, p. 1064-1065, 17 out. 2005. BMJ.

GRUZITIS, Normunds; BARZDINS, Guntis. Polysemy in Controlled Natural Language Texts. **Lecture Notes In Computer Science**, [S.I.], p. 102-120, 2010. Springer Berlin Heidelberg.

HAARST, R. V. **SBVR Made Easy**: Business Vocabulary and Rules as a Critical Asset. Amsterdam: Wardy Poelstra, 2013.

HAJ, Abdellatif *et al.* Automated Checking of Conformance to SBVR Structured English Notation. *In: International Conference on Advanced Intelligent Systems for Sustainable Development*. Cham: Springer International Publishing, 2018. p.697-706

HAJ, Abdellatif *et al.* Automatic Extraction of SBVR Based Business Vocabulary from Natural Language Business Rules. **Lecture Notes In Networks And Systems**, [S.I.], p. 170-182, 2019. Springer International Publishing.

HAJ, Abdellatif *et al.* The Semantic of Business Vocabulary and Business Rules: an automatic generation from textual statements. **Ieee Access**, [S.I.], v. 9, p. 56506-56522, 2021a. Institute of Electrical and Electronics Engineers (IEEE).

HAJ, Abdellatif *et al.* Automated generation of terminological dictionary from textual business rules. **Journal Of Software: Evolution and Process**, [S.I.], v. 33, n. 5, p. 2339-2339, 19 mar. 2021b. Wiley.

HAJ, Abdellatif *et al.* Automated Identification of Semantic Similarity between Concepts of Textual Business Rules. **International Journal Of Intelligent Engineering And Systems**, [S.I.], v. 14, n. 1, p. 147-156, 28 fev. 2021c.

HE, Junqing *et al.* Never Lost in the Middle: mastering long-context question answering with position-agnostic decompositional training. [S.I.: s.n.], p. 13628-13642, 2023.

HE, Qianyu *et al.* From Complex to Simple: enhancing multi-constraint complex instruction following ability of large language models. [S.I.: s.n.], p. 1-18, jun. 2024.

HE, Xiaoxin *et al.* G-Retriever: retrieval-augmented generation for textual graph understanding and question answering. [S.I.: s.n.], p. 132876-132907, 2024.

HINKELMANN, Knut *et al.* Ontology-based metamodeling. **Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change**. Cidade: Editora, 2018. p. 177-194

HOEKSTRA, Rinke *et al.* The LKIF Core Ontology of Basic Legal Concepts. **LOAIT**, [S.I.], v. 321, p. 43-64, jun. 2007.

HOGAN, Aidan *et al.* Knowledge Graphs. **Acm Computing Surveys**, [S.I.], v. 54, n. 4, p. 1-37, 2 jul. 2021. Association for Computing Machinery (ACM).

HOHFELD, Wesley Newcomb. Fundamental Legal Conceptions as Applied in Judicial Reasoning. **The Yale Law Journal**, [S.I.], v. 26, n. 8, p. 710, jun. 1917.

HOUGH, Alexander R.; GLUCK, Kevin A. The understanding problem in cognitive science. **Advances in Cognitive Systems**, [S.I.], v. 8, p. 13-32, dez. 2019.

HUANG, Hui *et al.* An Empirical Study of LLM-as-a-Judge for LLM Evaluation: fine-tuned judge model is not a general substitute for gpt-4. [S.I.: s.n.], p. 1-19, nov. 2024.

ICAIL 23. **Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law**. Braga - Portugal: Association for Computing Machinery, 2023.

INFOMONEY. **NYSE**: o que é e como funciona a Bolsa de Nova York. 2022. Disponível em: <https://www.infomoney.com.br/guias/o-que-e-nyse/>. Acesso em: 22 out. 2023.

INVESTOR.GOV. **The Laws That Govern the Securities Industry**. 2023. Disponível em: <https://www.investor.gov/introduction-investing/investing-basics/role-sec/laws-govern-securities-industry>. Acesso em: 22 out. 2023.

JACKSON, Philip. Understanding understanding and ambiguity in natural language. **Procedia Computer Science**, [S.I.], v. 169, p. 209-225, 2020. Elsevier BV.

JAYZED DATA MODELS INC. **Semantic compliance in finance**: The Financial Regulation Ontologies (FRO) are an open source design published in Ontology Web Language (OWL). 2021. Disponível em: <https://finregont.com/>. Acesso em: 18 nov. 2023.

JANSSEN, Kathleen. The influence of the PSI directive on open government data: an overview of recent developments. **Government Information Quarterly**, [S.I.], v. 28, n. 4, p. 446-456, out. 2011. Elsevier BV.

JIANG, Zhengbao *et al.* How Can We Know What Language Models Know? **Transactions Of The Association For Computational Linguistics**, [S.I.], v. 8, p. 423-438, dez. 2020. MIT Press - Journals.

JING, Kun; XU, Jungang. A Survey on Neural Network Language Models. **[S.N.]**, [S.I.], p. 1-7, jun. 2019.

JOSHI, Karuna Pande; SAHA, Srishty. A semantically rich framework for knowledge representation of code of federal regulations. **Digital Government: Research and Practice**, v. 1, n. 3, p. 1-17, 2020.

JOSHI, Vivek; ANISH, Preethu Rose; GHAISAS, Smita. Domain adaptation for an automated classification of deontic modalities in software engineering contracts. In: **Proceedings of the 29<sup>th</sup> ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering**. 2021. p. 1275-1280.

JUNIOR, Ademar Crotti *et al.* Using mapping languages for building legal knowledge graphs from xml files. **arXiv preprint arXiv:1911.07673**, 2019.

KELES, Feyza Duman *et al.* On The Computational Complexity of Self-Attention. **International Conference On Algorithmic Learning Theory**, [S.I.], p. 597-619, 2022.

KHURANA, Diksha *et al.* Natural language processing: state of the art, current trends and challenges. **Multimedia Tools And Applications**, [S.I.], v. 82, n. 3, p. 3713-3744, 14 jul. 2022. Springer Science and Business Media LLC.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, [S.I.], v. 33, n. 2004, p. 1-26, jul. 2004.

KOWALSKI, Robert; DATOO, Akber. Logical English meets legal English for swaps and derivatives. **Artificial Intelligence And Law**, [S.I.], v. 30, n. 2, p. 163-197, 12 ago. 2021. Springer Science and Business Media LLC.

KULKARNI, Vinay *et al.* Toward automated regulatory compliance. **Csi Transactions On Ict**, [S.I.], v. 9, n. 2, p. 95-104, jun. 2021. Springer Science and Business Media LLC.

KULKARNI, Vinay. Regulatory Compliance at Optimal Cost with Minimum Exposure to Risk. *In: The AI-Enabled Enterprise*. Cham: Springer International Publishing, 2023. p. 35-55

LEE, Harrison *et al.* Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. **arXiv preprint arXiv:2309.00267**, 2023.

LEE, Seungwoo *et al.* Multi-faceted navigation of legal documents. *In: 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*. **IEEE**, p. 537-540, 2011.

LEWIS, Patrick *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, NY, v. 33, p. 9459-9474, dec. 2020.

LIND-COMBS, Holly C. *et al.* Comparing Levenshtein distance and dynamic time warping in predicting listeners' judgments of accent distance. **Speech Communication**, [S.I.], v. 155, p. 102987, nov. 2023. Elsevier BV.

LIU, Nelson F. *et al.* Lost in the Middle: how language models use long contexts. **Transactions Of The Association For Computational Linguistics**, [S.I.], v. 12, p. 157-173, 2023. ArXiv.

LIU, Zihan *et al.* NER-BERT: a pre-trained model for low-resource entity tagging. **[S.n.]**, Hong Kong, p. 1-15, dez. 2021. ArXiv.

LONGO, Luca *et al.* Explainable Artificial Intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. **Information Fusion**, [S.I.], v. 106, p. 102301, jun. 2024. Elsevier BV.

LOUIS, Antoine *et al.* Finding the Law: enhancing statutory article retrieval via graph neural networks. **Proceedings Of The 17Th Conference Of The European Chapter Of The Association For Computational Linguistics**, [S.I.], p. 2753-2768, 2023. Association for Computational Linguistics.

MAO, Xuting *et al.* Financial fraud detection using the related-party transaction knowledge graph. **Procedia Computer Science**, [S.I.], v. 199, p. 733-740, 2022. Elsevier BV.

MARTÍNEZ-FERNÁNDEZ, José L. *et al.* A Preliminary Approach to the Automatic Extraction of Business Rules from Unrestricted Text in the Banking

Industry. **Lecture Notes In Computer Science**, Londres, p. 299-310, 2008. Springer Berlin Heidelberg.

MCCABE, T.J.. A Complexity Measure. **Ieee Transactions On Software Engineering**, [S.I.], v. -2, n. 4, p. 308-320, dez. 1976. Institute of Electrical and Electronics Engineers (IEEE).

MEYER, Lars-Peter *et al.* LLM-assisted Knowledge Graph Engineering: experiments with chatgpt. **Informatik Aktuell**, [S.I.], p. 103-115, 2024. Springer Fachmedien Wiesbaden.

MIN, Bonan *et al.* Recent Advances in Natural Language Processing via Large Pre-trained Language Models: a survey. **Acm Computing Surveys**, [S.I.], v. 56, n. 2, p. 1-40, 14 set. 2023. Association for Computing Machinery (ACM).

MITRA, Sayandeept *et al.* Matgap: A systematic approach to perform match and gap analysis among sbvr-based domain specific business rules. *In: 2018 25th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, p. 551-560, 2018.

NAKAGAWA, Shinichi *et al.* Meta-evaluation of meta-analysis: ten appraisal questions for biologists. **Bmc Biology**, [S.I.], v. 15, n. 1, p. 1-14, 3 mar. 2017. Springer Science and Business Media LLC.

NASDAQ. **Stock Screener**. 2023. Disponível em:  
<https://www.nasdaq.com/market-activity/stocks/screener>. Acesso em: 21 out. 2023.

NEILL, James O' *et al.* Classifying sentential modality in legal language. **Proceedings Of The 16Th Edition Of The International Conference On Artificial Intelligence And Law**, [S.I.], p. 159-168, 12 jun. 2017. ACM.

NGUYEN, Ha-Thanh *et al.* Transformer-Based Approaches for Legal Text Processing. **The Review Of Socionetwork Strategies**, [S.I.], v. 16, n. 1, p. 135-155, 25 jan. 2022. Springer Science and Business Media LLC.  
<http://dx.doi.org/10.1007/s12626-022-00102-2>.

OCEG. **GRC Capability Model 3.5 (Red Book) FREE VERSION**. 2023. Disponível em: <https://www.oceg.org/grc-capability-model-red-book/>. Acesso em: 22 out. 2023.

OMAR, Reham *et al.* ChatGPT versus Traditional Question Answering for Knowledge Graphs: current status and future directions towards knowledge graph chatbots. **[S.I.: s.n.]**, p. 1-9, fev. 2023. ArXiv.

OMG - Object Management Group. **Semantics of Business Vocabulary and Business Rules Version 1.5**. 2019. Disponível em:  
<https://www.omg.org/spec/SBVR/1.5/PDF>. Acesso em: 22 out. 2023.

OPENAI. **GPT-4 (versão gpt-4o-2024-08-06)**. Disponível em:  
<https://chat.openai.com/>. Acesso em: 21 set. 2024.

OPSAHL-ONG, Krista *et al.* Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. [S.I.: s.n.], p. 1-31, out. 2024. ArXiv.

OXLEY, Michael G. **Sarbanes-Oxley Act of 2002**. 2002. Disponível em:  
<https://www.congress.gov/bill/107th-congress/house-bill/3763>. Acesso em: 03 abr. 2024.

PAN, Shirui *et al.* Unifying Large Language Models and Knowledge Graphs: a roadmap. **IEEE Transactions On Knowledge And Data Engineering**, [S.I.], v. 36, n. 7, p. 3580-3599, jul. 2024. Institute of Electrical and Electronics Engineers (IEEE).

PARVIZIMOSAED, Alireza *et al.* Model-checking legal contracts with SymboleoPC. **Proceedings Of The 25Th International Conference On Model Driven Engineering Languages And Systems**, [S.I.], p. 278-288, 23 out. 2022. ACM.

PETTICREW, Mark; ROBERTS, Helen. **Systematic reviews in the social sciences**: A practical guide. Oxford: John Wiley & Sons, 2008.

POLONIO, Fabio Augusto. **Método para obtenção de SBVR com definições bilíngues**. Orientador: Paulo Sérgio Muniz. 2018. 98 f. Dissertação (Mestrado em Ciência da Computação) - Curso de Mestrado em Engenharia de Computação, Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, São Paulo, 2018.

RABELO, Juliano *et al.* A Summary of the COLIEE 2019 Competition. **Lecture Notes In Computer Science**, [S.I.], p. 34-49, 2020. Springer International Publishing.

RACZ, Nicolas; WEIPPL, Edgar; SEUFERT, Andreas. A frame of reference for research of integrated governance, risk and compliance (GRC). *In:* Communications and Multimedia Security: 11<sup>th</sup> IFIP TC 6/TC, 11., 2010, Linz – Austria. **Anais...** Berlin, Heidelberg: CMS, 2010. p.106-117.

RAFAILOV, Rafael *et al.* Direct preference optimization: Your language model is secretly a reward model. **Advances in Neural Information Processing Systems**, v. 36, p. 53728-53741, 2023.

RAMZAN, Shabana *et al.* A model transformation from NL to SBVR. *In:* Ninth International Conference on Digital Information Management (ICDIM 2014). **IEEE**, p. 220-225, 2014.

RODRIGUES, Cleyton Mário de Oliveira *et al.* Legal ontologies over time: a systematic mapping study. **Expert Systems With Applications**, [S.I.], v. 130, p. 12-30, set. 2019. Elsevier BV.

ROYCHOUDHURY, Suman *et al.* From natural language to SBVR model authoring using structured english for compliance checking. In: 2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC). IEEE, p. 73-78, 2017.

SADIQ, Shazia; GOVERNATORI, Guido; NAMIRI, Kioumars. Modeling control objectives for business process compliance. In: Business Process Management: 5th International Conference, 5., 2007, Berlin, Heidelberg. **Anais...** Brisbane, Australia: BPM 2007. p. 149-164

SANSONE, Carlo; SPERLÍ, Giancarlo. Legal Information Retrieval systems: state-of-the-art and open issues. **Information Systems**, [S.I.], v. 106, p. 101967, maio 2022. Elsevier BV.

SCHMIDT, Alexander; OTTO, Boris; ÖSTERLE, Hubert. Integrating information systems: case studies on current challenges. **Electronic Markets**, [S.I.], v. 20, n. 2, p. 161-174, 18 maio 2010. Springer Science and Business Media LLC.

**SEC. The Laws That Govern the Securities Industry.** Disponível em: <https://www.investor.gov/introduction-investing/investing-basics/role-sec/laws-govern-securities-industry>. Acesso em: 22 out. 2023.

SELWAY, Matt *et al.* Formalising natural language specifications using a cognitive linguistic/configuration based approach. **Information Systems**, [S.I.], v. 54, p. 191-208, dez. 2015. Elsevier BV.

SEQUEDA, Juan *et al.* A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. **Proceedings Of The 7Th Joint Workshop On Graph Data Management Experiences & Systems (Grades) And Network Data Analytics (Nda)**, [S.I.], p. 1-12, 9 jun. 2024. ACM.

SHARIFI, Sepehr *et al.* Symboleo: towards a specification language for legal contracts. **2020 Ieee 28Th International Requirements Engineering Conference (Re)**, [S.I.], p. 364-369, ago. 2020. IEEE.

SHANKAR, Shreya *et al.* Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. **Proceedings Of The 37Th Annual Acm Symposium On User Interface Software And Technology**, [S.I.], p. 1-14, 11 out. 2024. ACM.

**SIFMA. Capital Markets Fact Book, 2023.** 2023. Disponível em: <https://www.sifma.org/resources/research/fact-book/>. Acesso em: 22 out. 2023.

SKERSYS, Tomas *et al.* Extracting SBVR business vocabularies and business rules from UML use case diagrams. **Journal Of Systems And Software**, [S.I.], v. 141, p. 111-130, jul. 2018. Elsevier BV.

SKERSYS, Tomas et al. Extracting SBVR Business Vocabularies from UML Use Case Models Using M2M Transformations Based on Drag-and-Drop Actions. **Applied Sciences**, [S.I.], v. 11, n. 14, p. 6464, 13 jul. 2021. MDPI AG.

SKERSYS, Tomas et al. Transforming BPMN Processes to SBVR Process Rules with Deontic Modalities. **Applied Sciences**, [S.I.], v. 12, n. 18, p. 8976, 7 set. 2022. MDPI AG.

SOWA, John F. Conceptual graphs. **Foundations of artificial intelligence**, v. 3, p. 213-237, 2008.

STATISTA. **Comparison of the number of listed companies on the New York Stock Exchange (NYSE) and Nasdaq from 2018 to 1<sup>st</sup> quarter 2023, by domicile.** 2023. Disponível em: <https://www.statista.com/statistics/1277216/nyse-nasdaq-comparison-number-listed-companies/>. Acesso em: 22 out. 2023.

STRĄK, Tomasz; TUSZYŃSKI, Michał. Nlp based retrieval of semantically similar private tax rulings. **Procedia Computer Science**, v. 207, p. 2853-2864, 2022.

SUN, Haoran et al. Conifer: improving complex constrained instruction-following ability of large language models. **[S.N.]**, [S.I.], p. 1-16, abr. 2024. ArXiv.

SUN, Tian-Xiang et al. Paradigm Shift in Natural Language Processing. **Machine Intelligence Research**, [S.I.], v. 19, n. 3, p. 169-183, 28 maio 2022. Springer Science and Business Media LLC.

SUNKLE, Sagar et al. Toward Better Mapping between Regulations and Operations of Enterprises Using Vocabularies and Semantic Similarity. **Complex Systems Informatics And Modeling Quarterly**, [S.I.], n. 5, p. 39-60, 29 dez. 2015a.

SUNKLE, Sagar et al. Explanation of Proofs of Regulatory (Non-)Compliance Using Semantic Vocabularies. **Lecture Notes In Computer Science**, [S.I.], p. 388-403, 2015b. Springer International Publishing.

THOMSON REUTERS. **2023 Cost of Compliance**: regulatory burden poses operational challenges for compliance. 2023. Disponível em: <https://legal.thomsonreuters.com/en/insights/reports/cost-of-compliance-2023>. Acesso em: 22 out. 2023.

TONMOY, S. M Towhidul Islam et al. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. **[S.N.]**, [S.I.], p. 1-20, jan. 2024. ArXiv.

WANG, Jiayin et al. A User-Centric Multi-Intent Benchmark for Evaluating Large Language Models. **[S.N.]**, [S.I.], p. 1-24, set. 2024. ArXiv.

WATAOKA, Koki; TAKAHASHI, Tsubasa; RI, Ryokan. Self-Preference Bias in LLM-as-a-Judge. **[S.N.]**, [S.I.], p. 1-11, out. 2024. ArXiv.

WEI, Fangyun; CHEN, Xi; LUO, Lin. Rethinking generative large language model evaluation for semantic comprehension. **arXiv preprint arXiv:2403.07872**, 2024.

WEI, Jason *et al.* Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, [S.I.], v. 35, p. 24824-24837, jan. 2022.

WIERINGA, Roel. **Design science methodology for information systems and software engineering**. Springer-Verlag Berlin Heidelberg, 2014.

WIMMER, Manuel; BURGUEÑO, Loli. Testing M2T/T2M Transformations. **Lecture Notes In Computer Science**, [S.I.], p. 203-219, 2013. Springer Berlin Heidelberg.

WITT, Graham. **Writing effective business rules**. Elsevier, 2012.

WYNER, Adam; PETERS, Wim. On rule extraction from regulations. *In: Legal knowledge and information systems. IOS Press*, [S.I.], v. 235, p. 113-122, 2011.

WOHLIN, Claes. Guidelines for snowballing in systematic literature studies and a replication in software engineering. **Proceedings Of The 18Th International Conference On Evaluation And Assessment In Software Engineering**, [S.I.], p. 1-10, 13 maio 2014. ACM.

UFSCAR. **StArt**. Disponível em: <https://www.lapes.ufscar.br/resources/tools-1/start-1>. Acesso em: 03 abr. 2024.

VON HALLE, Barbara. **Business rules applied**: building better systems using the business rules approach. [S.I.]: Wiley Publishing, oct. 2001.

YANG, An *et al.* Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task. **[S.N.]**, [S.I.], p. 1-7, jun. 2018. ArXiv.

YU, Fangyi *et al.* Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. **Findings Of The Association For Computational Linguistics**: ACL 2023, [S.I.], p. 13582-13596, 2023. Association for Computational Linguistics.

ZARATIANA, Urchade *et al.* GLiNER: generalist model for named entity recognition using bidirectional transformer. **[S.N.]**, [S.I.], p. 1-11, nov. 2023. ArXiv.

ZHENG, Lianmin *et al.* Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *In: Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023. p. 46595-46623.

ZHONG, Haoxi *et al.* How Does NLP Benefit Legal System: a summary of legal artificial intelligence. **[S.N.]**, [S.I.], p. 1-13, maio 2020. ArXiv.

## APÊNDICE A – RESULTADOS DA RSL

A fase exploratória os estudos base de apoio à revisão sistemática foram identificados por meio de consulta bibliográfica de artigos publicados em conferências, periódicos e dissertações de mestrado (Biblioteca IPT<sup>71</sup>), obtidas por consultas na base de dados eletrônica Google Scholar<sup>72</sup>

Quadro 35 - Artigos base para a revisão sistemática

| # | Título                                                                                                                                                                                                                                                                                         | Autor(es)                                  | Base          |
|---|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|---------------|
| 1 | ABI-LAHOUD, Elie <i>et al.</i> Interpreting Regulations with SBVR, in: <b>Collaboration Technologies and Systems (CTS)</b> , IEEE, 23-27 Maio 2013, pp. 439-445.                                                                                                                               | Abi-Lahoud <i>et al.</i> (2013)            | IEEE          |
| 2 | BAJWA, Imran Sarwar; SHAHZADA, Muhammad Anwar. Automated generation of OCL constraints: NL based approach vs pattern based approach. <b>Mehran University Research Journal of Engineering &amp; Technology</b> , v. 36, n. 2, p. 243-254, 2017.                                                | Bajwa <i>et al.</i> (2017)                 | SnowBalling   |
| 3 | BERNOTAITYTE, Gintare <i>et al.</i> Developing SBVR vocabularies and business rules from OWL2 ontologies. In: Information and Software Technologies: 19th International Conference, ICIST 2013, Kaunas, Lithuania, October 2013. Proceedings 19. Springer Berlin Heidelberg, 2013. p. 134-145. | Bernotaityte <i>et al.</i> (2013)          | Springer      |
| 4 | CHITTIMALLI, Pavan Kumar <i>et al.</i> An approach to mine SBVR vocabularies and rules from business documents. In: <b>Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference</b> . 2020. p. 1-11.                  | Chittimalli <i>et al.</i> (2020)           | ACM           |
| 5 | DANENAS, Paulius; SKERSYS, Tomas; BUTLERIS, Rimantas. Natural language processing-enhanced extraction of SBVR business vocabularies and business rules from UML use case diagrams. <b>Data &amp; Knowledge Engineering</b> , v. 128, p. 101822, 2020.                                          | Danenas <i>et al.</i> (2020)               | Springer      |
| 6 | DE OLIVEIRA RODRIGUES, Cleyton Mario <i>et al.</i> Legal ontologies over time: A systematic mapping study. <b>Expert Systems with Applications</b> , v. 130, p. 12-30, 2019.                                                                                                                   | De Oliveira Rodrigues <i>et al.</i> (2019) | ScienceDirect |
| 7 | HAJ, Abdellatif <i>et al.</i> The semantic of business vocabulary and business rules: an automatic generation from textual statements. <b>IEEE Access</b> , v. 9, p. 56506-56522, 2021.                                                                                                        | HAJ, Abdellatif <i>et al.</i> (2021)       | IEEE          |
| 8 | JOSHI, Karuna Pande; SAHA, Srishty. A. Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations. <b>Digital Government: Research and Practice</b> , v. 1, n. 3, p. 1-17, 2020.                                                                                  | Joshi e Saha (2020)                        | ACM           |
| 9 | ROYCHOUDHURY, Suman <i>et al.</i> From natural language to SBVR model authoring using structured                                                                                                                                                                                               | Roychoudhury <i>et al.</i> (2017)          | IEEE          |

<sup>71</sup> <http://aleph.ipt.br/F>

<sup>72</sup> <http://scholar.google.com/>

|           |                                                                                                                                                                                                          |                              |             |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|-------------|
|           | english for compliance checking. In: <b>2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC)</b> . IEEE, 2017. p. 73-78.                                               |                              |             |
| <b>10</b> | SELWAY, Matt <i>et al.</i> Formalising natural language specifications using a cognitive linguistic/configuration based approach. <b>Information Systems</b> , v. 54, p. 191-208, 2015.                  | Selway <i>et al.</i> (2015)  | ACM         |
| <b>11</b> | SHARIFI, Sepehr <i>et al.</i> Symboleo: Towards a specification language for legal contracts. In: <b>2020 IEEE 28th international requirements engineering conference (RE)</b> . IEEE, 2020. p. 364-369. | Sharifi <i>et al.</i> (2020) | IEEE        |
| <b>12</b> | SKERSYS, Tomas <i>et al.</i> Transforming BPMN Processes to SBVR Process Rules with Deontic Modalities. <b>Applied Sciences</b> , v. 12, n. 18, p. 8976, 2022.                                           | Skersys <i>et al.</i> (2022) | SnowBalling |

Fonte: Elaborado pelo autor.

O Quadro 35 apresenta os artigos que foram julgados relevantes e representativos. Esses artigos foram obtidos de com a string de busca: “sbvr” OR “Semantics of Business Vocabulary and Rules”.

#### **Critérios de inclusão e exclusão dos trabalhos:**

Os critérios de inclusão e exclusão, limitam a seleção dos trabalhos conforme sua relevância para o objetivo da pesquisa. Para a avaliação qualitativa foram definidos os seguintes critérios:

##### *Critérios de inclusão (atender pelo menos um):*

- (I) (1) O estudo trata da transformação de regras
- (I) (2) O estudo relata o uso de técnicas de NLP
- (I) (3) O estudo relata dificuldades na condução da transformação
- (I) (4) O estudo apresenta alguma melhoria no processo de transformação ou GRC
- (I) (5) O estudo é referente ao domínio financeiro
- (I) (6) O estudo é referente a GRC

##### *Critérios de exclusão (atender um):*

- (E) (a) O estudo não trata da transformação de regras ou GRC
- (E) (b) O estudo não cita uso de técnicas de NLP

(E) (c) Estudo completo não está disponível para acesso por meio digital.

### **Extração dos dados dos artigos**

Para cada artigo aceito foi preenchido o formulário de extração de dados. Os dados extraídos foram:

- Para todos os artigos:
  - Metodologia de Pesquisa: Estudo de Caso, Experimento, Revisão Bibliográfica, Survey, Pesquisa-Ação, DSR, Outros)
  - Qual o objetivo do estudo?
  - Em que população o estudo foi aplicado?
  - Qual a motivação?
  - Quais as contribuições?
  - Quais as limitações do trabalho?
  - Quais os resultados?
  - Como os resultados foram validados?
  - Quais são os trabalhos futuros?
  - Referências de interesse (snowballing)
- Para artigos da *string 1* (Transformação para SBVR):
  - Alguma linguagem controlada foi utilizada? Qual? Como?
  - Como os textos de origem são processados? (P1)
  - Grafos de conhecimento foram utilizados? Como? (P1, P4, P5)
  - Qual o processo para identificar/extrair as regras? (P2, P3)
  - Ontologias foram utilizadas? Quais? E como? (P4, P5)
  - Qual o processo para identificar/criar o vocabulário? (P4, P5)
  - Como a desambiguação de termos é resolvida? (P5)
  - Qual o processo para converter as regras em linguagem controlada? (P7)
- Para artigos da *string 2* (GRC):
  - Como o processo de GRC é realizado?
  - Uso de algum tipo de IA? Com qual objetivo?
  - Quais os desafios para os processos de GRC?

### **Resultados**

O Quadro 36 apresenta todos os artigos retornados por base de dados. Foram encontrados um total de 211 artigos, onde: 30 na ACM, 25 no IEEE, 62 no ScienceDirect, 48 na Springer e 46 na WoS.

Quadro 36 - Artigos retornados por base de dados

| ID   | Title                                                                                                                        | Authors                                                                                                                                              | Ano  | A1 | A2 | BD |
|------|------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 7519 | Automated Natural Language Processing-Based Supplier Discovery for Financial Services                                        | Papa, Mauro and Chatzigiannakis, Ioannis and Anagnostopoulos, Aris                                                                                   | 2024 | N  | N  | 0  |
| 7520 | Ten Ways of Leveraging Ontologies for Natural Language Processing and its Enterprise Applications                            | Erekhinskaya, Tatiana and Strebkov, Dmitriy and Patel, Sujal and Balakrishna, Mithun and Tatu, Marta and Moldovan, Dan                               | 2020 | N  | N  | 0  |
| 7521 | The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain                                       | Mansar, Youness and Kang, Juyeon and El Maarouf, Ismail                                                                                              | 2021 | N  | N  | 0  |
| 7522 | An ontological artifact for classifying social media: Text mining analysis for financial data                                | Alzamil, Zamil and Appelbaum, Deniz and Nehmer, Robert                                                                                               | 2020 | N  | N  | 0  |
| 7523 | Financial markets sentiment analysis: developing a specialized Lexicon                                                       | Yekrangi, Mehdi and Abdolvand, Neda                                                                                                                  | 2021 | N  | N  | 0  |
| 7524 | Toward Robust Food Ontology Mapping                                                                                          | Stojanov, Riste and Kocev, Ilija and Gramatikov, Sasho and Popovski, Gorjan and Seljak, Barbara Korousic and Eftimov, Tome                           | 2020 | N  | N  | 0  |
| 7525 | Extraction of Formal Manufacturing Rules from Unstructured English Text                                                      | Kang, SungKu and Patil, Lalit and Rangarajan, Arvind and Moitra, Abha and Jia, Tao and Robinson, Dean and Ameri, Farhad and Dutta, Debasish          | 2021 | N  | N  | 0  |
| 7526 | Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development  | Tang, Xianming and Feng, Zhiqiang and Xiao, Yitian and Wang, Ming and Ye, Tianrui and Zhou, Yujie and Meng, Jin and Zhang, Baosen and Zhang, Dongwei | 2023 | N  | N  | 0  |
| 7527 | Automated feedback generation for formal manufacturing rule extraction                                                       | Kang, SungKu and Path, Lalit and Rangarajan, Arvind and Moitra, Abha and Jia, Tao and Robinson, Dean and Dutta, Debasish                             | 2019 | N  | N  | 0  |
| 7528 | Multi-ontology embeddings approach on human-aligned multi-ontologies representation for gene-disease associations prediction | Wang, Yihao and Wegner, Philipp and Domingo-Fernandez, Daniel and Kodamullil, Alpha Tom                                                              | 2023 | N  | N  | 0  |
| 7529 | Towards a Knowledge Base of Financial Relations: Overview and Project Description                                            | Jabbari, Ali and Sauvage, Olivier and Cabioch, Nicolas                                                                                               | 2019 | N  | N  | 0  |

| ID   | Title                                                                                                                           | Authors                                                                                                                                                                     | Ano  | A1 | A2 | BD |
|------|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 7530 | Ontology-based Interpretable Machine Learning for Textual Data                                                                  | Lai, Phung and Phan, NhatHai and Hu, Han and Badeti, Anuja and Newman,David and Dou, Dejing                                                                                 | 2020 | N  | N  | 0  |
| 7531 | Uncovering Social Media Users' Emotions Towards Companies Using Semantic Web Technologies                                       | Cotfas, Livu-Adrian and Delcea, Camelia and Nica, Ionut                                                                                                                     | 2020 | N  | N  | 0  |
| 7532 | Ontology-Based Ambiguity Resolution of Manufacturing Text for Formal Rule Extraction                                            | Kang, SungKu and Patil, Lalit and Rangarajan, Arvind and Moitra, Abha and Robinson, Dean and Jia, Tao and Dutta, Debasish                                                   | 2019 | N  | N  | 0  |
| 7533 | Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data                               | Ciroku, Fiorela and De Giorgis, Stefano and Gangemi, Aldo and Martinez-Pandiani, Delfina S. and Presutti, Valentina                                                         | 2024 | N  | N  | 0  |
| 7534 | Transforming text into knowledge graph: Extracting and structuring information from spatial development plans                   | Kaczmarek, Iwona                                                                                                                                                            | 2023 | N  | N  | 0  |
| 7535 | A Semantic Annotation Pipeline towards the Generation of Knowledge Graphs in Tribology                                          | Kuegler, Patricia and Marian, Max and Dorsch, Rene and Schleich, Benjamin and Wartzack, Sandro                                                                              | 2022 | N  | N  | 0  |
| 7536 | Knowledge-informed semantic alignment and rule interpretation for automated compliance checking                                 | Zheng, Zhe and Zhou, Yu-Cheng and Lu, Xin-Zheng and Lin, Jia-Rui                                                                                                            | 2022 | N  | N  | 0  |
| 7537 | Hybrid Semantic Annotation: Rule-based and Manual Annotation of the Open American National Corpus with Top-Level Ontology       | de Andrade, Guidson Coelho and Oliveira, Alcione de Paiva and Moreira, Alexandra                                                                                            | 2019 | N  | N  | 0  |
| 7538 | Automatic Generation of Dictionaries: The Journalistic Lexicon Case                                                             | Cristani, Matteo and Tomazzoli, Claudio and Zorzi, Margherita                                                                                                               | 2019 | N  | N  | 0  |
| 7539 | ENRICHING LARGE DOCUMENT STORES WITH INTELLIGENT METADATA: A FRAMEWORK FOR EFFECTIVE KNOWLEDGE MANAGEMENT AND APPLIED ANALYTICS | Ivanov, Penko and Pavlova, Elitsa                                                                                                                                           | 2023 | N  | N  | 0  |
| 7540 | An adaptable, high-performance relation extraction system for complex sentences                                                 | Thomas, Anu and Sivanesan, Sangeetha                                                                                                                                        | 2022 | N  | N  | 0  |
| 7541 | A computational system based on ontologies to automate the mapping process of medical reports into structured databases         | Oliva, Jefferson Tales and Lee, Huei Diana and Spolaor, Newton and Resende Takaki, Weber Shoity and Rodrigues Coy, Claudio Saddy and Fagundes, Joao Jose and Wu, Feng Chung | 2019 | N  | N  | 0  |
| 7542 | Construction of the knowledge service model of a port supply chain enterprise in a big data environment                         | Yang, Bo and Yang, Meifang                                                                                                                                                  | 2021 | N  | N  | 0  |

| ID   | Title                                                                                                              | Authors                                                                                                                              | Ano  | A1 | A2 | BD |
|------|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 7543 | Analysing Customers' Opinions Towards Product Characteristics Using Social Media                                   | Cotfas, Livu-Adrian and Delcea, Camelia and Nica, Ionut                                                                              | 2020 | N  | N  | 0  |
| 7544 | A Semantic Approach for Automating Knowledge in Policies of CyberInsurance Services                                | Joshi, Ketki and Joshi, Karuna Pande and Mittal, Sudip                                                                               | 2019 | N  | N  | 0  |
| 7545 | Information Processing and Retrieval from CSV File by Natural Language                                             | Tapsai, Chalermpol                                                                                                                   | 2018 | N  | N  | 0  |
| 7546 | An ensemble clustering approach for topic discovery using implicit text segmentation                               | Memon, Muhammad Qasim and Lu, Yu and Chen, Penghe and Memon, Aasma and Pathan, Muhammad Salman and Zardari, Zulfiqar Ali             | 2021 | N  | N  | 0  |
| 7547 | Research on the Similarity between Nodes with Hypernymy/Hyponymy Relations based on IC and Taxonomical Structure   | Zhang, Xiaogang and Sun, Lili                                                                                                        | 2022 | N  | N  | 0  |
| 7548 | A Domain-Adaptable Heterogeneous Information Integration Platform: Tourism and Biomedicine Domains                 | Gil, Rafael Munoz and de Buenaga Rodriguez, Manuel and Galisteo, Fernando Aparicio and Paez, Diego Gachet and Garcia-Cuesta, Esteban | 2021 | N  | N  | 0  |
| 7549 | Extracting and Understanding Call-to-actions of Push-Notifications                                                 | Esteves, Beatriz and Fraser, Kieran and Kulkarni, Shridhar and Conlan, Owen and Rodriguez-Doncel, Victor                             | 2022 | N  | N  | 0  |
| 7550 | A systematic knowledge graph-based smart management method for operations: A case study of standardized management | Wen, Peihan and Zhao, Yiming and Liu, Jin                                                                                            | 2023 | N  | N  | 0  |
| 7551 | Big Data-Assisted Word Sense Disambiguation for Sign Language                                                      | Naranjo-Zeledon, Luis and Fernandez, Antonio and Peral, Jesus and Chacon-Rivas, Mario                                                | 2019 | N  | N  | 0  |
| 7552 | Weighting-based semantic similarity measure based on topological parameters in semantic taxonomy                   | Saif, Abdulgabbar and Zainodin, Ummi Zakiah and Omar, Nazlia and Ghareb, Abdullah Saeed                                              | 2018 | N  | N  | 0  |
| 7553 | DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization                    | Abulaish, Muhammad and Parwez, Md Aslam and Jahiruddin                                                                               | 2019 | N  | N  | 0  |
| 7554 | Transcription profiling and identification of infection-related genes in Phytophthora cactorum                     | Chen, Xiao-Ren and Huang, Shen-Xin and Zhang, Ye and Sheng, Gui-Lin and Zhang, Bo-Yue and Li, Qi-Yuan and Zhu, Feng and Xu, Jing-You | 2018 | N  | N  | 0  |
| 7555 | Cross-environment activity recognition using word embeddings for sensor and activity representation                | Azkune, Gorka and Almeida, Aitor and Agirre, Eneko                                                                                   | 2020 | N  | N  | 0  |
| 7557 | Scorecards, gateways and rankings: remuneration and conduct in financial services                                  | Sheedy, Elizabeth and Zhang, Le and Steffan, Dominik                                                                                 | 2022 | N  | N  | 1  |

| ID   | Title                                                                                                           | Authors                                                                                                         | Ano  | A1 | A2 | BD |
|------|-----------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 7558 | Toward Scalable Artificial Intelligence in Finance                                                              | Sanz, Jorge L. C. and Zhu, Yada                                                                                 | 2021 | N  | N  | 1  |
| 7559 | Detection of fraud risks in retailing sector using MLP and SVM techniques                                       | Pehlivanli, Davut and Eken, Suleyman and Ayan, Ebubekir                                                         | 2019 | N  | N  | 1  |
| 7560 | Time for a paradigm change: Problems with the financial industry's approach to operational risk                 | Butler, Tom and Brooks, Robert                                                                                  | 2023 | N  | N  | 1  |
| 7561 | Does executive accountability enhance risk management and risk culture?                                         | Sheedy, Elizabeth and Canestrari-Soh, Dominic S. B.                                                             | 2023 | N  | N  | 1  |
| 7562 | Robotic process automation in accounting systems                                                                | Harrast, Steven A.                                                                                              | 2020 | N  | N  | 1  |
| 7563 | The Role of Risk Climate and Ethical Self-interest Climate in Predicting Unethical Pro-organisational Behaviour | Sheedy, Elizabeth and Garcia, Patrick and Jepsen, Denise                                                        | 2021 | N  | N  | 1  |
| 7564 | A Structural and Navigational Method for Integrated Organizational Information Privacy Protection               | Abdullah, Hanifa                                                                                                | 2018 | N  | N  | 1  |
| 7565 | AIRMan: An Artificial Intelligence (AI) Risk Management System                                                  | Tjoa, Simon and Kieseberg, Peter and Temper, Marlies and Zanol, Jakob and Wagner, Markus and Holzinger, Andreas | 2022 | N  | N  | 1  |
| 7622 | Bibliometric Analysis of Using IT-GRC for Corporate Resilience and Sustainability                               | Mahendra, Irfan and Prabowo, Harjanto and Warnars, Harco Leslie Hendric Spits and Tjong, Yanti                  | 2024 | N  | N  | 3  |
| 7623 | A Structural and Navigational Method for Integrated Organizational Information Privacy Protection               | Abdullah, Hanifa                                                                                                | 2018 | N  | N  | 3  |
| 7624 | Toward Scalable Artificial Intelligence in Finance                                                              | Sanz, Jorge L. C. and Zhu, Yada                                                                                 | 2021 | N  | N  | 3  |
| 7625 | Financial Markets Prediction with Deep Learning                                                                 | Wang, Jia and Sun, Tong and Liu, Benyuan and Cao, Yu and Wang, Degang                                           | 2018 | N  | N  | 3  |
| 7626 | Understanding US Regulatory Bodies                                                                              | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7627 | The FFIEC: An Introduction                                                                                      | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7628 | Risk Assessments                                                                                                | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7629 | Cybersecurity and the Board of Directors                                                                        | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7630 | Cybersecurity Frameworks                                                                                        | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7631 | The Challenging Role of the Regulator                                                                           | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7632 | The Landscape of Cybersecurity                                                                                  | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7633 | Risk Management                                                                                                 | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |
| 7634 | International Cybersecurity Laws and Regulations                                                                | Edwards, Jason and Weaver, Griffin                                                                              | 2024 | N  | N  | 3  |

| ID   | Title                                                                                                             | Authors                                                                                                                                      | Ano  | A1 | A2 | BD |
|------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 7635 | Information Security Governance and Compliance                                                                    | Chapple, Mike                                                                                                                                | 2022 | N  | N  | 3  |
| 8147 | The temporal dynamics of enterprise risk management                                                               | Klein Jr., Vitor Hugo ; Reilley, Jacob T.                                                                                                    | 2021 | N  | N  | 7  |
| 8148 | Harmful events and misconducts in financial organizations: Human biases and root causes                           | Faugere, Christophe ; Stul, Olivier                                                                                                          | 2021 | N  | N  | 7  |
| 8149 | Analysis of Robotics, Artificial intelligence and Machine learning in the field of healthcare sector              | Sarath Kumar Boddu, Raja ; Ahamad, Shahanawaj ; Pradeep Kumar, K.V. ; Ramalingam, Mritha ; Kirana Pallathadka, Laxmi ; Peniero Tupas, Fernan | 2022 | N  | N  | 7  |
| 8150 | Advancing the concept of cybersecurity as a public good                                                           | Kianpour, Mazaher ; Kowalski, Stewart James ; Øverby, Harald                                                                                 | 2022 | N  | N  | 7  |
| 8151 | Convergence of business, innovation, and sustainability at the tipping point of the sustainable development goals | Azmat, Fara ; Lim, Weng Marc ; Moyeen, Abdul ; Voola, Ranjit ; Gupta, Girish                                                                 | 2023 | N  | N  | 7  |
| 8152 | Harnessing GPT-4 for generation of cybersecurity GRC policies: A focus on ransomware attack mitigation            | McIntosh, Timothy ; Liu, Tong ; Susnjak, Teo ; Alavizadeh, Hooman ; Ng, Alex ; Nowrozy, Raza ; Watters, Paul                                 | 2023 | N  | N  | 7  |
| 8153 | Index                                                                                                             | Jahankhani, Hamid ; O'Dell, Liam M. ; Bowen, Gordon ; Hagan, Daniel ; Jamal, Arshad                                                          | 2021 | N  | N  | 7  |
| 8154 | A cybersecurity risk quantification and classification framework for informed risk mitigation decisions           | Zadeh, Amir ; Lavine, Brandon ; Zolbanin, Hamed ; Hopkins, Donald                                                                            | 2023 | N  | N  | 7  |
| 8155 | 11 - Accounting, Accountability, and Auditing                                                                     | Purpura, Philip P. ; Purpura, Philip P.                                                                                                      | 2019 | N  | N  | 7  |
| 8156 | Software optimization for LGPD compliance using Paraconsistent Evidential Annotated Logic Et                      | Sakamoto, Liliam Sayuri ; Alves, Davis ; Abe, Jair M. ; De Souza, Jonatas S. ; De Souza, Nilson A. ; Martinez, Angel A.G.                    | 2021 | N  | N  | 7  |
| 8157 | The emerging Cloud Dilemma: Balancing innovation with cross-border privacy and outsourcing regulations            | Gozman, Daniel ; Willcocks, Leslie                                                                                                           | 2019 | N  | N  | 7  |
| 8158 | A survey towards an integration of big data analytics to big insights for value-creation                          | Saggi, Mandeep Kaur ; Jain, Sushma                                                                                                           | 2018 | N  | N  | 7  |
| 8159 | Chapter 8 - Information and communication technologies: a curse or blessing for SMEs?                             | Mohammed, Anne-Marie ; Idris, Bochra ; Saridakis, George ; Benson, Vladlena ; Benson, Vladlena ; Mcalaney, John                              | 2020 | N  | N  | 7  |
| 8160 | Chapter 3 - Statutory and regulatory GRC                                                                          | Johnson, Leighton ; Johnson, Leighton                                                                                                        | 2020 | N  | N  | 7  |
| 8161 | Responding to a corruption crisis through disclosure and remedial action: The case of Petrobras                   | Whitehead, Martin ; Belghitar, Yacine                                                                                                        | 2022 | N  | N  | 7  |
| 8162 | Chapter Nine - Continuous automation approach for                                                                 | Lupo, Guy ; Vo, Bao Quoc ; Locke, Natania ; Wu, Peggy ;                                                                                      | 2024 | N  | N  | 7  |

| ID   | Title                                                                                                                                                      | Authors                                                                               | Ano  | A1 | A2 | BD |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|------|----|----|----|
|      | autonomous ethics-based audit of AI systems                                                                                                                | Salpukas, Michael ; Wu, Hsin-Fu ; Ellsworth, Shannon                                  |      |    |    |    |
| 8163 | Fostering economic prosperity through fintech-driven banking expansion and financial inclusion: Strategies and innovations                                 | Banna, Hasanul ; Rana, Md Sohel                                                       | 2023 | N  | N  | 7  |
| 8164 | Appendix C - List of smart cities and critical infrastructure policies and financing standards                                                             | Vacca, John R. ; R.Vacca, John                                                        | 2022 | N  | N  | 7  |
| 8165 | Investigating technology development in the energy sector and its implications for Indonesia                                                               | Sambodo, Maxensius Tri ; Silalahi, Mesnan ; Firdaus, Nur                              | 2024 | N  | N  | 7  |
| 8166 | List of abbreviations                                                                                                                                      | Rijcken, Claudia                                                                      | 2019 | N  | N  | 7  |
| 8167 | A deeper look into cybersecurity issues in the wake of Covid-19: A survey                                                                                  | Alawida, Moatsum ; Omolara, Abiodun Esther ; Abiodun, Oludare Isaac ; Al-Rajab, Murad | 2022 | N  | N  | 7  |
| 8168 | A Novel Approach for Optimizing Governance, Risk management and Compliance for Enterprise Information security using DEMATEL and FoM                       | Ramalingam, Dharmalingam ; Arun, Shivasankarappa ; Anbazhagan, Neelamegam             | 2018 | N  | N  | 7  |
| 8169 | Compliance-innovation: A quality-based route to sustainability                                                                                             | Doyle, Eleanor ; McGovern, Damien ; McCarthy, Stephen ; Perez-Alaniz, Mauricio        | 2019 | N  | N  | 7  |
| 8170 | A pathway model to five lines of accountability in cybersecurity governance                                                                                | Slapničar, Sergeja ; Axelsen, Micheal ; Bongiovanni, Ivano ; Stockdale, David         | 2023 | N  | N  | 7  |
| 8171 | Impact of dynamic capabilities on digital transformation and innovation to improve banking performance: A TOE framework study                              | Abdurrahman, Abdurrahman ; Gustomo, Aurik ; Prasetyo, Eko Agus                        | 2024 | N  | N  | 7  |
| 8172 | Illuminating and bridging the vortex between tacit and explicit knowledge: Counterbalancing information asymmetry in high-value low-frequency transactions | Hoksbergen, Mark ; Chan, Johnny ; Peko, Gabrielle ; Sundaram, David                   | 2021 | N  | N  | 7  |
| 8173 | Identity and access management in cloud environment: Mechanisms and challenges                                                                             | Indu, I. ; Anand, P.M. Rubesh ; Bhaskar, Vidhyacharan                                 | 2018 | N  | N  | 7  |
| 8174 | Customer relationships and interdependences in the internationalization process of the firm                                                                | Limp, André ; Rezende, Sérgio Fernando Loureiro ; Versiani, Ângela França             | 2018 | N  | N  | 7  |
| 8175 | On distributed ledgers security and illegal uses                                                                                                           | Moubarak, Joanna ; Chamoun, Maroun ; Filoli, Eric                                     | 2020 | N  | N  | 7  |
| 8176 | Moving from a ‘human-as-problem’ to a ‘human-as-solution’ cybersecurity mindset                                                                            | Zimmermann, Verena ; Renaud, Karen                                                    | 2019 | N  | N  | 7  |
| 8177 | The significance of securing as a critical component of                                                                                                    | Burdon, Mark ; Coles-Kemp, Lizzie                                                     | 2019 | N  | N  | 7  |

| ID   | Title                                                                                                                                     | Authors                                                                                                        | Ano  | A1 | A2 | BD |
|------|-------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|------|----|----|----|
|      | information security: An Australian narrative                                                                                             |                                                                                                                |      |    |    |    |
| 8178 | CloudFNF: An ontology structure for functional and non-functional features of cloud services                                              | Al-Sayed, Mustafa M. ; Hassan, Hesham A. ; Omara, Fatma A.                                                     | 2020 | N  | N  | 7  |
| 8179 | The interdiscursive appeal of risk matrices: Collective symbols, flexibility normalism and the interplay of 'risk' and 'uncertainty'      | Jordan, Silvia ; Mitterhofer, Hermann ; Jørgensen, Lene                                                        | 2018 | N  | N  | 7  |
| 8180 | Factors influencing effective use of big data: A research framework                                                                       | Surbakti, Feliks P. Sejahtera ; Wang, Wei ; Indulska, Marta ; Sadiq, Shazia                                    | 2020 | N  | N  | 7  |
| 8181 | Business analytics and AGI in corporate management systems                                                                                | Sukhobokov, Artem A.                                                                                           | 2018 | N  | N  | 7  |
| 8182 | Pandemics and environmental shocks: What aviation managers should learn from COVID-19 for long-term planning                              | Linden, Erik                                                                                                   | 2021 | N  | N  | 7  |
| 8183 | Paradoxical tensions in the implementation of digital security governance: Toward an ambidextrous approach to governing digital security  | Schinagl, Stef ; Shahim, Abbas ; Khapova, Svetlana                                                             | 2022 | N  | N  | 7  |
| 8184 | Evaluation of Commercial Next-Generation Sequencing Bioinformatics Software Solutions                                                     | Gullapalli, Rama R.                                                                                            | 2020 | N  | N  | 7  |
| 8185 | Validation of a socio-technical management process for optimising cybersecurity practices                                                 | Malatji, Masike ; Marnewick, Annlizé ; von Solms, Suné                                                         | 2020 | N  | N  | 7  |
| 8186 | Impact of comprehensive information security awareness and cognitive characteristics on security incident management – an empirical study | Thangavelu, Manisekaran ; Krishnaswamy, Venkataraman ; Sharma, Mayank                                          | 2021 | N  | N  | 7  |
| 8187 | A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly                                              | Yao, Yifan ; Duan, Jinhao ; Xu, Kaidi ; Cai, Yuanfang ; Sun, Zhibo ; Zhang, Yue                                | 2024 | N  | N  | 7  |
| 8188 | Analyzing pedagogical approaches used in second auditing courses                                                                          | Reinstein, Alan ; Churyk, Natalie Tatiana ; Tate, Stefanie L.                                                  | 2018 | N  | N  | 7  |
| 8189 | Chapter 2 - Cornell BEST: Keys to successful institutionalization of career and professional development programming                      | Varvayanis, Susi ; Schaffer, Chris B. ; August, Avery ; Infante Lara, Lorena ; Daniel, Laura ; Chalkley, Roger | 2020 | N  | N  | 7  |
| 8190 | 1 - Information and Data Management                                                                                                       | Ott, Florence ; Ott, Florence                                                                                  | 2021 | N  | N  | 7  |
| 8191 | The management of moral hazard through the                                                                                                | San-Jose, Leire ; Gonzalo, Jose Felix ; Ruiz-Roqueñi, Maite                                                    | 2022 | N  | N  | 7  |

| ID    | Title                                                                                                            | Authors                                                                                                                                                                                                                          | Ano  | A1 | A2 | BD |
|-------|------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
|       | implementation of a Moral Compliance Model (MCM)                                                                 |                                                                                                                                                                                                                                  |      |    |    |    |
| 8192  | The role of informational mechanisms in the adoption of Green IS to achieve eco-sustainability in municipalities | Butler, Tom ; Hackney, Ray                                                                                                                                                                                                       | 2021 | N  | N  | 7  |
| 8193  | 11 - Will the new security trends achieve the skin in the game? (Lesson learned from recent IOCs)                | Bottazzi, Giovanni ; Me, Gianluigi ; Rutigliano, Giuseppe Giulio ; Perrone, Pierluigi ; Capone, Luciano ; Jahankhani, Hamid ; O'Dell, Liam M. ; Bowen, Gordon ; Hagan, Daniel ; Jamal, Arshad                                    | 2021 | N  | N  | 7  |
| 8194  | Chapter 6 - Smart Sensor and Big Data Security and Resilience                                                    | Venticinque, Salvatore ; Amato, Alba ; Ficco, Massimo ; Palmieri, Francesco                                                                                                                                                      | 2018 | N  | N  | 7  |
| 10493 | Research on Semantic Verification Method of AIXM Data Based on SBVR                                              | Wang, Xiaowen ; Tian, Yungang ; Fu, Shenghao ; Musila, Charity Muthoni ; Liang, Qilian ; Wang, Wei ; Mu, Jiasong ; Liu, Xin ; Na, Zhenyu                                                                                         | 2023 | N  | N  | 8  |
| 10494 | Human-Centred Automated Reasoning for Regulatory Reporting via Knowledge-Driven Computing                        | Thilakarathne, Dilhan J. ; Al Haider, Newres ; Bosman, Joost ; Fujita, Hamido ; Fournier-Viger, Philippe ; Ali, Moonis ; Sasaki, Jun                                                                                             | 2020 | N  | N  | 8  |
| 10495 | Detecting Constraints and Their Relations from Regulatory Documents Using NLP Techniques                         | Winter, Karolin ; Rinderle-Ma, Stefanie ; Panetto, Hervé ; Debruyne, Christophe ; Proper, Henderik A. ; Ardagna, Claudio Agostino ; Roman, Dumitru ; Meersman, Robert                                                            | 2018 | N  | N  | 8  |
| 10496 | Towards a Framework for Writing Executable Natural Language Rules                                                | Bampis, Konstantinos ; Kolovos, Dimitrios ; Hingorani, Justin ; Pierantonio, Alfonso ; Trujillo, Salvador                                                                                                                        | 2018 | N  | N  | 8  |
| 10497 | Legal Patterns for Different Constitutive Rules                                                                  | Ceci, Marcello ; Butler, Tom ; O'Brien, Leona ; Al Khalil, Firas ; Pagallo, Ugo ; Palmirani, Monica ; Casanovas, Pompeu ; Sartor, Giovanni ; Villata, Serena                                                                     | 2018 | N  | N  | 8  |
| 10498 | Challenges in the Implementation of Privacy Enhancing Semantic Technologies (PESTs) Supporting GDPR              | Saniei, Rana ; Rodríguez-Doncel, Víctor ; Palmirani, Monica ; Araszkiewicz, Michał ; Casanovas, Pompeu ; Pagallo, Ugo ; Sartor, Giovanni                                                                                         | 2021 | N  | N  | 8  |
| 10499 | Execution of Multi-perspective Declarative Process Models                                                        | Ackermann, Lars ; Schönig, Stefan ; Petter, Sebastian ; Schützenmeier, Nicolai ; Jablonski, Stefan ; Panetto, Hervé ; Debruyne, Christophe ; Proper, Henderik A. ; Ardagna, Claudio Agostino ; Roman, Dumitru ; Meersman, Robert | 2018 | N  | N  | 8  |
| 10500 | Logic Based Look-Ahead for the Execution of Multi-perspective Declarative Processes                              | Käppel, Martin ; Schützenmeier, Nicolai ; Schönig, Stefan ; Ackermann, Lars ; Jablonski, Stefan ; Reinhartz-Berger, Iris                                                                                                         | 2019 | N  | N  | 8  |

| ID    | Title                                                                                                                      | Authors                                                                                                                                                                             | Ano  | A1 | A2 | BD |
|-------|----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
|       |                                                                                                                            | Zdravkovic, Jelena ; Gulden, Jens ; Schmidt, Rainer                                                                                                                                 |      |    |    |    |
| 10501 | A Case Study on Modeling and Validating Financial Regulations Using (Semi-) Automated Compliance Framework                 | Roychoudhury, Suman ; Sunkle, Sagar ; Choudhary, Namrata ; Kholkar, Deepali ; Kulkarni, Vinay ; Buchmann, Robert Andrei ; Karagiannis, Dimitris ; Kirikova, Marite                  | 2018 | N  | N  | 8  |
| 10502 | A Model-Driven Engineering Approach to Complex Performance Indicators: Towards Self-Service Performance Management (SS-PM) | Giunta, Benito ; Cherfi, Samira ; Perini, Anna ; Nurcan, Selmin                                                                                                                     | 2021 | N  | N  | 8  |
| 10503 | Collaborative Business Process Solution Considering an Ontological Dimension of Process Models                             | Sebu, Maria Laura ; Ciocarlie, Horia ; Balas, Valentina Emilia ; Jain, Lakhmi C. ; Balas, Marius Mircea                                                                             | 2018 | N  | N  | 8  |
| 10504 | Deep Learning for the Identification of Decision Modelling Components from Text                                            | Goossens, Alexandre ; Claessens, Michelle ; Parthoens, Charlotte ; Vanthienen, Jan ; Moschoyiannis, Sotiris ; Peñaloza, Rafael ; Vanthienen, Jan ; Soylu, Ahmet ; Roman, Dumitru    | 2021 | N  | N  | 8  |
| 10505 | KnIGHT: Mapping Privacy Policies to GDPR                                                                                   | Mousavi Nejad, Najmeh ; Scerri, Simon ; Lehmann, Jens ; Faron Zucker, Catherine ; Ghidini, Chiara ; Napoli, Amedeo ; Toussaint, Yannick                                             | 2018 | N  | N  | 8  |
| 10506 | Comparing the Performance of GPT-3 with BERT for Decision Requirements Modeling                                            | Goossens, Alexandre ; De Smedt, Johannes ; Vanthienen, Jan ; Sellami, Mohamed ; Vidal, Maria-Ester ; van Dongen, Boudewijn ; Gaaloul, Walid ; Panetto, Hervé                        | 2024 | N  | N  | 8  |
| 10507 | Automated Consistency Management in BPMN Based Business Process Models                                                     | Ishaq, Mamoon ; Choudhary, M. Abbas ; Bajwa, Imran Sarwar ; Kamareddine, Fairouz ; Costa, Anna                                                                                      | 2019 | N  | N  | 8  |
| 10508 | Automating Requirements Elicitation of Cloud-Based ERPs                                                                    | Abd Elmonem, Mohamed A. ; Nasr, Eman S. ; Gheith, Mervat H. ; Hassanien, Aboul Ella ; Shaalan, Khaled ; Gaber, Tarek ; Tolba, Mohamed F.                                            | 2018 | N  | N  | 8  |
| 10509 | Empowering Business-Level Blockchain Users with a Rules Framework for Smart Contracts                                      | Astigarraga, Tara ; Chen, Xiaoyan ; Chen, Yaoliang ; Gu, Jingxiao ; Hull, Richard ; Jiao, Limei ; Li, Yuliang ; Novotny, Petr ; Pahl, Claus ; Vukovic, Maja ; Yin, Jianwei ; Yu, Qi | 2018 | N  | N  | 8  |
| 10510 | Legislation-driven development of a Gift Aid system using Event-B                                                          | Williams, David M. ; Darwish, Salaheddin ; Schneider, Steve ; Michael, David R.                                                                                                     | 2020 | N  | N  | 8  |
| 10511 | Toward automated regulatory compliance                                                                                     | Kulkarni, Vinay ; Sunkle, Sagar ; Kholkar, Deepali ; Roychoudhury, Suman ; Kumar, Rahul ; Raghunandan, M.                                                                           | 2021 | N  | N  | 8  |

| ID    | Title                                                                                                                                                                    | Authors                                                                                       | Ano  | A1 | A2 | BD |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------|----|----|----|
| 10512 | AI-driven streamlined modeling: experiences and lessons learned from multiple domains                                                                                    | Sunkle, Sagar ; Saxena, Krati ; Patil, Ashwini ; Kulkarni, Vinay                              | 2022 | N  | N  | 8  |
| 10513 | Modeling compliance specifications in linear temporal logic, event processing language and property specification patterns: a controlled experiment on understandability | Czepa, Christoph ; Amiri, Amirali ; Ntentos, Evangelos ; Zdun, Uwe                            | 2019 | N  | N  | 8  |
| 10514 | Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions           | Tyagi, Nemika ; Bhushan, Bharat                                                               | 2023 | N  | N  | 8  |
| 10515 | What is a process model composed of?                                                                                                                                     | Adamo, Greta ; Ghidini, Chiara ; Di Francescomarino, Chiara                                   | 2021 | N  | N  | 8  |
| 10516 | Logical English meets legal English for swaps and derivatives                                                                                                            | Kowalski, Robert ; Datoo, Akber                                                               | 2021 | N  | N  | 8  |
| 10517 | Evaluating cloud database migration options using workload models                                                                                                        | Ellison, Martyn ; Calinescu, Radu ; Paige, Richard F.                                         | 2018 | N  | N  | 8  |
| 10518 | From enterprise models to low-code applications: mapping DEMO to Mendix; illustrated in the social housing domain                                                        | Krouwel, Marien R. ; Op 't Land, Martin ; Proper, Henderik A.                                 | 2024 | N  | N  | 8  |
| 10519 | A framework for interoperability between models with hybrid tools                                                                                                        | Braun, Germán ; Fillottrani, Pablo Rubén ; Keet, C. Maria                                     | 2023 | N  | N  | 8  |
| 10520 | EANM'23 Abstract Book Congress Sep 9-13, 2023                                                                                                                            |                                                                                               | 2023 | N  | N  | 8  |
| 10521 | Suitability Assessment and Carbon Mitigation Potential of Plantations on India's Railway Land                                                                            | Chaturvedi, Rajiv Kumar ; Jose, Karun ; Shruthi, B. V. R. ; Kariya, Karan Piyush ; Garg, Amit | 2022 | N  | N  | 8  |
| 10522 | Formal Model of Business Processes Integrated with Business Rules                                                                                                        | Kluza, Krzysztof ; Nalepa, Grzegorz J.                                                        | 2019 | N  | N  | 8  |
| 10523 | A Unified Approach Toward Security Audit and Compliance in Cloud Computing                                                                                               | Rajesh, Y. S. ; Kumar, V. G. Kiran ; Poojari, Asmita                                          | 2024 | N  | N  | 9  |
| 10524 | Integrating governance, risk and compliance? A multi-method analysis of the new Three Lines Model                                                                        | Seidenfuss, Kai-Uwe ; Young, Angus ; Datwani, Mohan                                           | 2023 | N  | N  | 9  |
| 10525 | Regtech: steering the regulatory spaceship in the right direction?                                                                                                       | Papantoniou, Alexandros A.                                                                    | 2022 | N  | N  | 9  |
| 10526 | Dialectic tensions in the financial markets: a longitudinal study of pre- and                                                                                            | Currie, Wendy L. ; Gozman, Daniel P. ; Seddon, Jonathan J. M.                                 | 2018 | N  | N  | 9  |

| ID           | Title                                                                                                           | Authors                                                                                | Ano  | A1 | A2 | BD |
|--------------|-----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|------|----|----|----|
|              | post-crisis regulatory technology                                                                               |                                                                                        |      |    |    |    |
| <b>10527</b> | Presenting Cloud Business Performance for Manufacturing Organizations                                           | Chang, Victor                                                                          | 2020 | N  | N  | 9  |
| <b>10528</b> | Ant: a process aware annotation software for regulatory compliance                                              | Gyory, Raphaël ; Restrepo Amariles, David ; Lewkowicz, Gregory ; Bersini, Hugues       | 2023 | N  | N  | 9  |
| <b>10529</b> | Ethical and legal responsibility for Artificial Intelligence                                                    | Henz, Patrick                                                                          | 2021 | N  | N  | 9  |
| <b>10530</b> | What could possibly go wrong? A multi-panel Delphi study of organizational social media risk                    | Di Gangi, Paul M. ; Johnston, Allen C. ; Worrell, James L. ; Thompson, Samuel C.       | 2018 | N  | N  | 9  |
| <b>10531</b> | Is green FinTech reshaping the finance sphere? Unravelling through a systematic literature review               | Geetha, Sreelekshmi ; Biju, Ajithakumari Vijayappan Nair                               | 2024 | N  | N  | 9  |
| <b>10532</b> | Ethical and preventive legal technology                                                                         | Stathis, Georgios ; van den Herik, Jaap                                                | 2024 | N  | N  | 9  |
| <b>10533</b> | The role of CIO in digital transformation: an exploratory study                                                 | Chawla, Raghu Nandan ; Goyal, Praveen ; Saxena, Deepak Kumar                           | 2023 | N  | N  | 9  |
| <b>10534</b> | How to Achieve Swift Resilience: the Role of Digital Innovation Enabled Mindfulness                             | Ye, Dandan ; Liu, Martin J. ; Luo, Jun ; Yannopoulou, Natalia                          | 2024 | N  | N  | 9  |
| <b>10535</b> | Multilevel determinants of MNC corruption risk                                                                  | Markscheffel, Juliane ; Plouffe, Michael                                               | 2022 | N  | N  | 9  |
| <b>10536</b> | AI-driven streamlined modeling: experiences and lessons learned from multiple domains                           | Sunkle, Sagar ; Saxena, Krati ; Patil, Ashwini ; Kulkarni, Vinay                       | 2022 | N  | N  | 9  |
| <b>10537</b> | ISF Editorial 2020                                                                                              | Ramesh, R. ; Rao, H. R.                                                                | 2020 | N  | N  | 9  |
| <b>10538</b> | Decentralizing corporate governance? A praxeological inquiry                                                    | Mitchell, Scott L. ; Packard, Mark D. ; Clark, Brent B.                                | 2022 | N  | N  | 9  |
| <b>10539</b> | Agile governance theory: operationalization                                                                     | Luna, Alexandre J. H. de O. ; Marinho, Marcelo L. M. ; de Moura, Hermano P.            | 2020 | N  | N  | 9  |
| <b>10540</b> | The Role of Risk Climate and Ethical Self-interest Climate in Predicting Unethical Pro-organisational Behaviour | Sheedy, Elizabeth ; Garcia, Patrick ; Jepsen, Denise                                   | 2021 | N  | N  | 9  |
| <b>10609</b> | Evaluating the Impact of NIST 800.53 Security Control Violations                                                | Hamdani, Syed Waqas and Brealey, Chris and Kontogiannis, Kostas and Giammaria, Alberto | 2022 | N  | N  | 11 |
| <b>10610</b> | Architecture design decision maps for software sustainability                                                   | Lago, Patricia                                                                         | 2019 | N  | N  | 11 |
| <b>10611</b> | Using Cloud Control Matrix to evaluate trust in cloud providers                                                 | Maria Bruma, Livia                                                                     | 2021 | N  | N  | 11 |
| <b>10612</b> | Aggregation of security metrics for decision making: a reference architecture                                   | Ahmed, Yussuf and Naqvi, Syed and Josephs, Mark                                        | 2018 | N  | N  | 11 |
| <b>10613</b> | Factors Affecting the Implementation of Remote                                                                  | Gilberta, Gisela and Widuri, Rindang and Kasenda, Faris                                | 2022 | N  | N  | 11 |

| ID           | Title                                                                                                                                 | Authors                                                                                         | Ano  | A1 | A2 | BD |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|------|----|----|----|
|              | Audit in the Audit Process in the COVID-19 Pandemic                                                                                   |                                                                                                 |      |    |    |    |
| <b>10614</b> | Using Robust Data Governance to Mitigate the Impact of Cybercrime                                                                     | Onwujekwe, Gerald and Thomas, Manoj and Osei-Bryson, Kweku-Muata                                | 2019 | N  | N  | 11 |
| <b>10615</b> | Enabling Graph Neural Networks for Semi-Supervised Risk Prediction in Online Credit Loan Services                                     | Tang, Hao and Wang, Cheng and Zheng, Jianguo and Jiang, Changjun                                | 2024 | N  | N  | 11 |
| <b>10616</b> | Raccoon: automated verification of guarded race conditions in web applications                                                        | Koch, Simon and Sauer, Tim and Johns, Martin and Pellegrino, Giancarlo                          | 2020 | N  | N  | 11 |
| <b>10617</b> | Digital Retail Challenges within the EU: Fulfillment of Holistic Customer Journey Post GDPR                                           | Nabbosa, Veronica L. and Iftikhar, Rehan                                                        | 2019 | N  | N  | 11 |
| <b>10618</b> | Show Me the Whole World: Towards Entire Item Space Exploration for Interactive Personalized Recommendations                           | Song, Yu and Sun, Shuai and Lian, Jianxun and Huang, Hong and Li, Yu and Jin, Hai and Xie, Xing | 2022 | N  | N  | 11 |
| <b>10619</b> | Integrating the common variability language with multilanguage annotations for web engineering                                        | Horcas, Jose-Miguel and Corti\~{n}as, Alejandro and Fuentes, Lidia and Luaces, Miguel R.        | 2018 | N  | N  | 11 |
| <b>10620</b> | A framework and benchmark for deep batch active learning for regression                                                               | Holzm\"{u}ller, David and Zaverkin, Viktor and K\"{a}stner, Johannes and Steinwart, Ingo        | 2024 | N  | N  | 11 |
| <b>10621</b> | Research on Privacy Preserving of Searchable Encryption                                                                               | Feng, Tao and He, Weiyou                                                                        | 2018 | N  | N  | 11 |
| <b>10622</b> | Enhancing the extraction of SBVR business vocabularies and business rules from UML use case diagrams with natural language processing | Danenas, Paulius and Skersys, Tomas and Butleris, Rimantas                                      | 2019 | N  | N  | 12 |
| <b>10623</b> | Report on the 3rd International Joint Conference on Rules and Reasoning (RuleML+RR 2019)                                              | Calvanese, Diego and Fodor, Paul and Montali, Marco                                             | 2020 | N  | N  | 12 |
| <b>10624</b> | A web-based tool for analysing normative documents in english                                                                         | Camilleri, John J. and Haghshenas, Mohammad Reza and Schneider, Gerardo                         | 2018 | N  | N  | 12 |
| <b>10625</b> | SBVR-based Business Rule Creation for Legacy Programs using Variable Provenance                                                       | Chittimalli, Pavan Kumar and Bhattacharyya, Abhidip                                             | 2019 | Y  | N  | 12 |
| <b>10626</b> | An Approach to Mine SBVR Vocabularies and Rules from Business Documents                                                               | Chittimalli, Pavan Kumar and Prakash, Chandan and Naik, Ravindra and Bhattacharyya, Abhidip     | 2020 | N  | N  | 12 |
| <b>10627</b> | Open Information Extraction Using Dependency Parser for Business Rule Mining in SBVR Format                                           | Prakash, Chandan and Chittimalli, Pavan Kumar and Naik, Ravindra                                | 2021 | N  | N  | 12 |
| <b>10628</b> | BuRRiT0: a framework to extract, specify, verify and analyze business rules                                                           | Chittimalli, Pavan Kumar and Anand, Kritika and Pradhan, Shrishti and Mitra, Sayandee           | 2020 | N  | N  | 12 |

| ID           | Title                                                                                                           | Authors                                                                                                                                                                          | Ano  | A1 | A2 | BD |
|--------------|-----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
|              |                                                                                                                 | and Prakash, Chandan and Shere, Rohit and Naik, Ravindra                                                                                                                         |      |    |    |    |
| <b>10629</b> | Legislation-driven development of a Gift Aid system using Event-B                                               | Williams, David M. and Darwish, Salaheddin and Schneider, Steve and Michael, David R.                                                                                            | 2020 | N  | N  | 12 |
| <b>10630</b> | Relation Identification in Business Rules for Domain-specific Documents                                         | Bhattacharyya, Abhidip and Chittimalli, Pavan Kumar and Naik, Ravindra                                                                                                           | 2018 | N  | N  | 12 |
| <b>10631</b> | Generating UML Use Case and Activity Diagrams Using NLP Techniques and Heuristics Rules                         | M. Maatuk, Abdelsalam and A. Abdelnabi, Esra                                                                                                                                     | 2021 | N  | N  | 12 |
| <b>10632</b> | ERS-TOOL: hybrid model for software requirements elicitation in Spanish language                                | Apaza, Ren\'{a}n Dar\'{i}o Gonzales and Barrios, Jhon Edilberto Monroy and Becerra, Diego Alonso Iquira and Quispe, Jos\'{e} Alfredo Herrera                                     | 2018 | N  | N  | 12 |
| <b>10633</b> | Natural Language Processing in Business Process Identification and Modeling: A Systematic Literature Review     | de Almeida Bordignon, Ana Cl\'{a}udia and Thom, Lucin\'{e}ia Heloisa and Silva, Thanner Soares and Dani, Vinicius Stein and Fantinato, Marcelo and Ferreira, Renato Cesar Borges | 2018 | N  | N  | 12 |
| <b>10634</b> | Nominal Recursors as Epi-Recursors                                                                              | Popescu, Andrei                                                                                                                                                                  | 2024 | N  | N  | 12 |
| <b>10635</b> | DocToModel: Automated Authoring of Models from Diverse Requirements Specification Documents                     | Rajbhoj, Asha and Nistala, Padmalata and Kulkarni, Vinay and Soni, Shivani and Pathan, Ajim                                                                                      | 2023 | N  | N  | 12 |
| <b>10636</b> | Minimizing the ambiguities in medical devices regulations based on software requirement engineering techniques  | Alsaadi, Mohmood and Lisitsa, Alexei and Qasaimeh, Malik                                                                                                                         | 2019 | N  | N  | 12 |
| <b>10637</b> | Combining OCL and natural language: a call for a community effort                                               | Cabot, Jordi and Delgado, David and Burgue\~{n}o, Lola                                                                                                                           | 2022 | N  | N  | 12 |
| <b>10638</b> | 3D sketching for interactive model retrieval in virtual reality                                                 | Giunchi, Daniele and James, Stuart and Steed, Anthony                                                                                                                            | 2018 | N  | N  | 12 |
| <b>10639</b> | The Impact of SBVR Adoption on Booking Intention in Ciputra Hotel Jakarta                                       | Christopher, Christopherr and Karimun, Vincent Kuonadi and Adiatii, Maria Pia and Rosman, Dendy                                                                                  | 2023 | N  | N  | 13 |
| <b>10640</b> | MatGap: A Systematic Approach to Perform Match and Gap Analysis among SBVR-Based Domain Specific Business Rules | Mitra, Sayandeepr and Prakash, Chandan and Chakraborty, Shayak and Chittimalli, Pavan Kumar                                                                                      | 2018 | Y  | Y  | 13 |
| <b>10641</b> | The Semantic of Business Vocabulary and Business Rules: An Automatic Generation From Textual Statements         | Haj, Abdellatif and Jarrar, Abdessamad and Balouki, Youssef and Gadir, Taoufiq                                                                                                   | 2021 | N  | N  | 13 |
| <b>10642</b> | Usability of the Business Rules Specification Languages                                                         | Hnatkowska, Bogumila and Hnatkowska, Anna                                                                                                                                        | 2020 | N  | N  | 13 |
| <b>10643</b> | Deontic Rule of Rule-Based Service Choreographies                                                               | Abidin, Nor Najihah Zainanl and Manaf, Nurulhuda A. and                                                                                                                          | 2021 | N  | N  | 13 |

| ID    | Title                                                                                                                                   | Authors                                                                                                                                                       | Ano  | A1 | A2 | BD |
|-------|-----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
|       |                                                                                                                                         | Moschoyiannis, Sotiris and Jamaludin, Nur Amalina                                                                                                             |      |    |    |    |
| 10644 | BuRRITo: A Framework to Extract, Specify, Verify and Analyze Business Rules                                                             | Chittimalli, Pavan Kumar and Anand, Kritika and Pradhan, Shrishti and Mitra, Sayandeep and Prakash, Chandan and Shere, Rohit and Naik, Ravindra               | 2019 | N  | N  | 13 |
| 10645 | Trace-Based Verification of Rule-Based Service Choreographies                                                                           | Moschoyiannis, Sotiris and Maglaras, Leandros and Manaf, Nurulhuda A.                                                                                         | 2018 | N  | N  | 13 |
| 10646 | Fine-Grained Video Retrieval With Scene Sketches                                                                                        | Zuo, Ran and Deng, Xiaoming and Chen, Keqi and Zhang, Zhengming and Lai, Yu-Kun and Liu, Fang and Ma, Cuixia and Wang, Hao and Liu, Yong-Jin and Wang, Hongan | 2023 | N  | N  | 13 |
| 10647 | Fine-Grained Instance-Level Sketch-Based Video Retrieval                                                                                | Xu, Peng and Liu, Kun and Xiang, Tao and Hospedales, Timothy M. and Ma, Zhanyu and Guo, Jun and Song, Yi-Zhe                                                  | 2021 | N  | N  | 13 |
| 10648 | A Double Sensing Scheme With Selective Bitline Voltage Regulation for Ultralow-Voltage Timing Speculative SRAM                          | Yang, Jun and Ji, Hao and Guo, Yichen and Zhu, Jizhe and Zhuang, Yuan and Li, Zhi and Liu, Xinning and Shi, Longxing                                          | 2018 | N  | N  | 13 |
| 10649 | Smartphone-Based Virtual Reality Systems (SBVRs) as a Promotion Tools for Foodservice Industry                                          | Wahyuningtyas, Dianka and Rosman, Dendy and Levyta, Farah and Putra, Eka Diraksa and Maulana, Ryonaldy                                                        | 2022 | N  | N  | 13 |
| 10650 | Towards the adoption of OMG standards in the development of SOA-based IoT systems                                                       | Costa, Bruno ; Pires, Paulo F. ; Delicato, Flávia C.                                                                                                          | 2020 | N  | N  | 14 |
| 10651 | Extracting SBVR business vocabularies and business rules from UML use case diagrams                                                     | Skersys, Tomas ; Danenas, Paulius ; Butleris, Rimantas                                                                                                        | 2018 | N  | N  | 14 |
| 10652 | Chapter 14 - Rental Cars-R-Us case study                                                                                                | Harmon, Paul ; Harmon, Paul                                                                                                                                   | 2019 | N  | N  | 14 |
| 10653 | Tackling rapid technology changes by applying enterprise engineering theories                                                           | Dvořák, Ondřej ; Pergl, Robert                                                                                                                                | 2022 | N  | N  | 14 |
| 10654 | BIGOWL4DQ: Ontology-driven approach for Big Data quality meta-modelling, selection and reasoning                                        | Barba-González, Cristóbal ; Caballero, Ismael ; Varela-Vaca, Ángel Jesús ; Cruz-Lemus, José A. ; Gómez-López, María Teresa ; Navas-Delgado, Ismael            | 2024 | N  | N  | 14 |
| 10655 | Characterizing and evaluating the quality of software process modeling language: Comparison of ten representative model-based languages | García-García, J.A. ; Enríquez, J.G. ; Domínguez-Mayo, F.J.                                                                                                   | 2019 | N  | N  | 14 |
| 10656 | Natural language processing-enhanced extraction of SBVR business vocabularies and business rules from UML use case diagrams             | Danenas, Paulius ; Skersys, Tomas ; Butleris, Rimantas                                                                                                        | 2020 | N  | N  | 14 |

| ID    | Title                                                                                                                                                      | Authors                                                                                      | Ano  | A1 | A2 | BD |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|------|----|----|----|
| 10657 | Chapter 3 - Key Concepts                                                                                                                                   | McGilvray, Danette ; McGilvray, Danette                                                      | 2021 | N  | N  | 14 |
| 10658 | Importing participatory practices of the socio-environmental systems community to the process system engineering community: An application to supply chain | Roth, Anastasia ; Pinta, François ; Negny, Stéphane ; Montastruc, Ludovic                    | 2021 | N  | N  | 14 |
| 10659 | A comprehensive review on resolving ambiguities in natural language processing                                                                             | Yadav, Apurwa ; Patel, Aarshil ; Shah, Manan                                                 | 2021 | N  | N  | 14 |
| 10660 | An integrated socio-technical enterprise modelling: A scenario of healthcare system analysis and design                                                    | Fayoumi, Amjad ; Williams, Richard                                                           | 2021 | N  | N  | 14 |
| 10661 | On the declarative paradigm in hybrid business process representations: A conceptual framework and a systematic literature study                           | Abbad Andaloussi, Amine ; Burattin, Andrea ; Slaats, Tijs ; Kindler, Ekkart ; Weber, Barbara | 2020 | N  | N  | 14 |
| 10662 | AuMixDw: Towards an automated hybrid approach for building XML data warehouses                                                                             | Ouaret, Zoubir ; Boukraa, Doulkifli ; Boussaid, Omar ; Chalal, Rachid                        | 2019 | N  | N  | 14 |
| 10663 | Integration of Business Applications with the Blockchain: Odoo and Hyperledger Fabric Open Source Proof of Concept                                         | Belhi, Abdelhak ; Gasmi, Houssem ; Bouras, Abdelaziz ; Aouni, Belaid ; Khalil, Ibrahim       | 2021 | N  | N  | 14 |
| 16995 | Natural language ambiguity resolution by intelligent semantic annotation of software requirements                                                          | Ashfaq, Fariha ; Bajwa, Imran Sarwar                                                         | 2021 | N  | N  | 15 |
| 16996 | Automatic extraction of SBVR based business vocabulary from natural language business rules                                                                | Haj, Abdellatif ; Balouki, Youssef ; Gadi, Taoufiq                                           | 2019 | N  | N  | 15 |
| 16997 | Developing SBVR vocabularies and business rules from OWL2 ontologies                                                                                       | Bernotaityte, Gintare ; Nemuraite, Lina ; Butkiene, Rita ; Paradauskas, Bronius              | 2013 | N  | N  | 15 |
| 16998 | A preliminary approach to the automatic extraction of business rules from unrestricted text in the banking industry                                        | Martínez-Fernández, José L. ; González, José C. ; Villena, Julio ; Martínez, Paloma          | 2008 | N  | N  | 15 |
| 16999 | Automated generation of terminological dictionary from textual business rules                                                                              | Haj, Abdellatif ; Balouki, Youssef ; Gadi, Taoufiq                                           | 2021 | N  | N  | 15 |
| 17000 | SBVR Business Rules Generation from Natural Language Specification                                                                                         | Bajwa, Imran S ; Lee, Mark G ; Bordbar, Behzad                                               | 2010 | N  | N  | 15 |
| 17001 | From UML/OCL to SBVR specifications: A challenging transformation                                                                                          | Cabot, Jordi ; Pau, Raquel ; Raventós, Ruth                                                  | 2010 | N  | N  | 15 |
| 17002 | From Natural Language to SBVR Model Authoring Using                                                                                                        | Roychoudhury, Suman ; Sunkle, Sagar ; Kholkar, Deepali ; Kulkarni, Vinay                     | 2017 | N  | N  | 15 |

| ID    | Title                                                                                                              | Authors                                                                     | Ano  | A1 | A2 | BD |
|-------|--------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|------|----|----|----|
|       | Structured English for Compliance Checking                                                                         |                                                                             |      |    |    |    |
| 17003 | A model transformation from NL to SBVR                                                                             | Ramzan, Shabana ; Bajwa, Imran Sarwar ; Ul Haq, Ikram ; Naeem, M. Asif      | 2014 | N  | N  | 15 |
| 17004 | Automated Identification of Semantic Similarity between Concepts of Textual Business Rules                         | Haj, Abdellatif ; Balouki, Youssef ; Gadi, Taoufiq                          | 2020 | N  | N  | 15 |
| 17005 | From Legal Contracts to Formal Specifications: A Systematic Literature Review                                      | Soavi, Michele ; Zeni, Nicola ; Mylopoulos, John ; Mich, Luisa              | 2022 | N  | N  | 15 |
| 17006 | Combining Natural Language Processing Approaches for Rule Extraction from Legal Documents                          | Dragonì, Mauro ; Villata, Serena ; Rizzi, Williams ; Governatori, Guido     | 2018 | N  | N  | 15 |
| 17007 | Automated checking of conformance to SBVR structured english notation                                              | Haj, Abdellatif ; Balouki, Youssef ; Gadi, Taoufiq                          | 2019 | N  | N  | 15 |
| 17008 | Toward Better Mapping between Regulations and Operations of Enterprises Using Vocabularies and Semantic Similarity | Sunkle, Sagar ; Kholkar, Deepali ; Kulkarni, Vinay                          | 2015 | N  | N  | 15 |
| 17009 | Explanation of proofs of regulatory (Non-)compliance using semantic vocabularies                                   | Sunkle, Sagar ; Kholkar, Deepali ; Kulkarni, Vinay                          | 2015 | N  | N  | 15 |
| 17010 | A frame of reference for research of integrated Governance, Risk and Compliance (GRC)                              | Racz, Nicolas ; Weippl, Edgar ; Seufert, Andreas                            | 2010 | N  | N  | 15 |
| 17011 | Transforming BPMN Processes to SBVR Process Rules with Deontic Modalities                                          | Skersys, Tomas ; Danenas, Paulius ; Mickeviciute, Egle ; Butleris, Rimantas | 2022 | N  | N  | 15 |
| 17012 | An automated detection of inconsistencies in SBVR-based business rules using many-sorted logic                     | Anand, Kritika ; Chittimalli, Pavan Kumar ; Naik, Ravindra                  | 2018 | N  | N  | 15 |
| 17013 | On Rule Extraction from Regulations                                                                                | Wyner, Adam ; Peters, Wim                                                   | 2011 | N  | N  | 15 |
| 17014 | Deontic Ambiguities in Legal Reasoning                                                                             | Governatori, Guido ; Rotolo, Antonino                                       | 2023 | N  | N  | 15 |
| 17015 | Domain adaptation for an automated classification of deontic modalities in software engineering contracts          | Joshi, Vivek ; Anish, Preethu Rose ; Ghaisas, Smita                         | 2021 | N  | N  | 15 |
| 17016 | From Legal Information to Open Legal Data: A Case Study in U.S. Federal Legal Information                          | Casellas, Núria ; Vallbé, Joan-Josep ; Bruce, Thomas Robert                 | 2011 | N  | N  | 15 |
| 17017 | A Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations                          | Joshi, Karuna Pande ; Saha, Srishty                                         | 2020 | N  | N  | 15 |
| 17018 | Domain-specific knowledge graphs: A survey                                                                         | Abu-Salih, Bilal                                                            | 2021 | N  | N  | 15 |

| ID    | Title                                                                                                                | Authors                                                                                                                                                                               | Ano  | A1 | A2 | BD |
|-------|----------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|----|----|----|
| 17019 | Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey                                                       | Agrawal, Garima ; Kumarage, Tharindu ; Alghami, Zeyad ; Liu, Huan                                                                                                                     | 2023 | N  | N  | 15 |
| 17020 | Linked legal data: Improving access to regulations                                                                   | Casellas, Nuria ; Bruce, Thomas R. ; Frug, Sara S. ; Bouwman, Sarah ; Dias, Dallas ; Lin, Jie ; Marathe, Sharvari ; Rai, Krithi ; Singh, Ankit ; Sinha, Debraj ; Venkataraman, Sanjna | 2012 | N  | N  | 15 |
| 17021 | Standards for Knowledge Graphs in the Financial Sector                                                               | Bennett, Michael G.                                                                                                                                                                   | 2021 | N  | N  | 15 |
| 17022 | Experimental investigation of transformations from SBVR business vocabularies and business rules to owl 2 ontologies | Karpovič, Jaroslav ; Ablonskis, Linas ; Nemuraitė, Lina ; Paradauskas, Bronius                                                                                                        | 2016 | N  | N  | 15 |
| 17023 | An Approach to Mine SBVR Vocabularies and Rules from Business Documents                                              | Kumar, Pavan ; Prakash, Chandan ; Naik, Ravindra ; Bhattacharyya, Abhidip                                                                                                             | 2020 | N  | N  | 15 |
| 17024 | Interpreting Regulations with SBVR                                                                                   | Abi-Lahoud, Elie ; Butler, Tom ; Chapin, Donald ; Hall, John                                                                                                                          | 2013 | N  | N  | 15 |
| 17025 | Patterns for legal compliance checking in a decidable framework of linked open data                                  | Francesconi, Enrico ; Governatori, Guido                                                                                                                                              | 2023 | N  | N  | 15 |
| 17026 | Automated Generation of OCL Constraints: NL based Approach vs Pattern Based Approach                                 | Bajwa, Imran Sarwar ; Shahzada, Muhammad Anwar                                                                                                                                        | 2017 | N  | N  | 15 |
| 17027 | An approach to mine business rule intents from domain-specific documents                                             | Bhattacharyya, Abhidip ; Chittimalli, Pavan Kumar ; Naik, Ravindra                                                                                                                    | 2017 | N  | N  | 15 |
| 17028 | Open Information Extraction Using Dependency Parser for Business Rule Mining in SBVR Format                          | Prakash, Chandan ; Chittimalli, Pavan Kumar ; Naik, Ravindra                                                                                                                          | 2021 | N  | N  | 15 |
| 17029 | Identifying and extracting hierarchical information from business PDF documents                                      | Shere, Rohit Prakash ; Chittimalli, Pavan Kumar ; Naik, Ravindra                                                                                                                      | 2022 | N  | N  | 15 |
| 17030 | A Systematic Review of Methods for Consistency Checking in SBVR-based Business Rules.                                | Mitra, Sayandeep ; Chittimalli, Pavan Kumar                                                                                                                                           | 2017 | N  | N  | 15 |
| 17031 | Domain-independent method of detecting inconsistencies in SBVR-based business rules                                  | Chittimalli, Pavan Kumar ; Anand, Kritika                                                                                                                                             | 2016 | N  | N  | 15 |
| 17032 | Identifying Anomalies in SBVR-based Business Rules using Directed Graphs and SMT-LIBv2.                              | Mitra, S ; Anand, K ; (2), PK Chittimalli - ICEIS ; 2018, undefined                                                                                                                   | 2018 | N  | N  | 15 |
| 17033 | Semantic search and query over SBVR-based business rules using SMT based approach and information retrieval method   | Anand, Kritika ; Mitra, Sayandeep ; Chittimalli, Pavan Kumar                                                                                                                          | 2019 | N  | N  | 15 |
| 17034 | Ontology matching: State of the art and future challenges                                                            | Shvaiko, Pavel ; Euzenat, Jérôme                                                                                                                                                      | 2013 | N  | N  | 15 |

Fonte: Elaborado pelo autor.

A cada artigo foi atribuído um identificador único e arbitrário para resumir o Quadro 36 que apresenta o resultado da aplicação dos critérios de seleção e exclusão e está organizado da seguinte maneira:

- ID: contém o número sequencial do estudo, independente da base consultada;
- A1: Indica o resultado da avaliação 1, onde Y – indica que foi aceito, N – indica que foi rejeitado;
- A2: Indica o resultado da avaliação 2, onde Y – indica que foi aceito, N – indica que foi rejeitado;

Após aplicar os critérios de inclusão e exclusão estabelecidos nos 211 trabalhos, foram aceitos 65 na primeira avaliação, 37 na segunda. O Quadro 36 apresenta os detalhes da avaliação.

Quadro 37 - Principais revistas e congressos consultados

| Nome                                                                                                     |
|----------------------------------------------------------------------------------------------------------|
| 2018 25th Asia-Pacific Software Engineering Conference (APSEC)                                           |
| 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA)                          |
| 2018 INTERNATIONAL CONFERENCE ON ADVANCES IN BIG DATA, COMPUTING AND DATA COMMUNICATION SYSTEMS (ICABCD) |
| 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)                                |
| 2021 2nd International Conference on Computing and Data Science (CDS)                                    |
| 2021 IEEE INTERNATIONAL CONFERENCE ON SERVICES COMPUTING (SCC 2021)                                      |
| 2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)           |
| AI Open                                                                                                  |
| AUTOMATION IN CONSTRUCTION                                                                               |

Fonte: Elaborado pelo autor

O Quadro 37 resume as principais revistas e congressos extraídos dos artigos.

## Artefatos

Os artefatos criados para essa revisão estão disponíveis no repositório de códigos<sup>73</sup> sob a pasta “rsl”.

---

<sup>73</sup> Endereço <https://github.com/asantos2000/master-degree-santos-anderson>

## **APÊNDICE B – FORNECEDORES DE SOLUÇÕES PARA A GRC**

A metodologia utilizada foi a consulta as publicações disponíveis pelos fornecedores usando a ferramenta de busca Google, procurando pelas palavras-chave: (nome do fornecedor OR nome do produto) AND (“Artificial Intelligence” OR “AI” OR “LLM” OR “ChatGPT” OR “GenAI” OR Generative AI” OR “Machine Learning” OR “ML”). O “APÊNDICE B – FORNECEDORES DE SOLUÇÕES PARA A GRC” contém a lista dos fornecedores com o resultado da pesquisa.

Não foi surpreendente que a maioria dos fornecedores anuncie algum tipo de IA em seus produtos, porém, apenas dois fornecedores (1, 7) descreveram o uso de IA para inferência sobre as regras de negócio, semelhante ao proposto nessa pesquisa, embora não foi possível obter mais informações das suas publicações on-line. Os demais se dividiram meio-a-meio entre uso do LLM para perguntas e respostas, geração de texto e resumo e o outro grupo com uso mais tradicional de IA, com classificação e detecção de anomalias, resumidos no Quadro 38.

Quadro 38 - Uso de IA pelos produtos

| <b>Uso</b>            | <b>Produtos</b> |
|-----------------------|-----------------|
| Extração de regras    | 2               |
| Classificação         | 4               |
| Detecção de anomalias | 2               |
| Geração de texto      | 1               |
| Perguntas e respostas | 4               |
| Resumo                | 2               |

Fonte: Elaborado pelo autor

O Quadro 39 contém a lista dos fornecedores de solução para GRC avaliados em 2024-04-08. Ela é utilizada na seção “3.1.1 Visão geral das soluções comerciais para a GRC”.

Quadro 39 - Fornecedores de GRC

| ID | Fabricante | Produto                  | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                | Uso        | Processo | Referência (URL)                                                                                                                                                                      |
|----|------------|--------------------------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | Archer     | Archer                   | Sim | <b>Archer Regulatory Content Analysis</b><br>allows you to more effectively identify regulatory changes and ensure compliance while minimizing resource-intensive manual processes. Incorporating patent-pending technology, it utilizes natural language processing and machine learning to analyze how an organization maps existing regulations to controls. | Mapeamento | Manual   | <a href="https://www.archerirm.com/_files/ugd/0fc6af_5b220b9998f4450e86864c1fb74b36b7.pdf">https://www.archerirm.com/_files/ugd/0fc6af_5b220b9998f4450e86864c1fb74b36b7.pdf</a>       |
| 2  | AuditBoard | AuditBoard Risk Platform | Não | Não está claro, não foi encontrado na documentação do produto, apenas em um anúncio.                                                                                                                                                                                                                                                                            | N/A        | N/A      | <a href="https://www.auditboard.com/blog/auditboard-announces-ai-and-analytics-capabilities/">https://www.auditboard.com/blog/auditboard-announces-ai-and-analytics-capabilities/</a> |

| ID | Fabricante | Produto      | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Uso                   | Processo | Referência (URL)                                                                                                                                                    |
|----|------------|--------------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 3  | Camms      | Camms Risk   | Sim | Machine-learning (ML) helps focus Operational Risk Management (ORM) resources on situations that require human intervention by replacing rules-driven alerting systems. ML can also analyse and identify emerging threats more sensitively than rules-based triggers.; it can also apply unsupervised techniques to spot instances of fraud and analyse new customers” and trading partners” profiles faster and in greater depth. Natural-language processing can be applied to call surveillance. | Detecção de anomalias | N/A      | <a href="https://cammsgroup.com/wp-content/uploads/2023/11/OP-Risk-eBook-Global.pdf">https://cammsgroup.com/wp-content/uploads/2023/11/OP-Risk-eBook-Global.pdf</a> |
| 4  | Deligent   | Deligent One | Sim | Identify key provisions, generate contract summaries and help your team manage review with machine learning powered analysis. Hundreds of pre-trained clause models to leverage on day one; Rapidly and easily train Diligen to recognise new concepts.                                                                                                                                                                                                                                             | Resumo, classificação | Manual   | <a href="https://www.diligent.com/">https://www.diligent.com/</a>                                                                                                   |

| ID | Fabricante             | Produto                                             | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Uso                   | Processo | Referência (URL)                                                                                                            |
|----|------------------------|-----------------------------------------------------|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|----------|-----------------------------------------------------------------------------------------------------------------------------|
| 5  | Fusion Risk Management | Fusion Framework System, Fusion Resilient Framework | Sim | <p>Simplify your complicated business processes, and automatically generate executive reporting. Expand your team with a Fusion taught AI-powered Analyst at your fingertips. Copilot will boost team engagement through user-guided support within the Fusion Framework® System™ to answer questions or make suggestions on a self-service basis. Amplify your dynamic scenario response evaluation with insightful data analysis. Fusion Copilot™ can quickly parse through your program data, assess relevant information, and provide its analysis and/or recommendations instantly, something that would take a client much longer to do.</p> | Perguntas e respostas | Manual   | <a href="https://www.fusionrm.com/solutions/resilience-copilot/">https://www.fusionrm.com/solutions/resilience-copilot/</a> |

| ID | Fabricante | Produto       | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Uso                                           | Processo | Referência (URL)                                                                                        |
|----|------------|---------------|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------|----------|---------------------------------------------------------------------------------------------------------|
| 6  | IBM        | IBM OpenPages | Sim | Forneça respostas em tempo real para perguntas comuns com o Watson Assistant. Analise a documentação do usuário e consulte as perguntas frequentes para tornar o conhecimento do GRC disponível a todos os usuários. Integre-se ao Natural Language Understanding para selecionar e mapear uma lista pré-selecionada de categorizações de risco para controles específicos, incluindo incidentes de TI ou categorias de tipo de evento da Basileia. Integre-se ao Catálogo de conhecimento para obter o valor total dos metadados. Mantenha uma lista de ativos de dados e projetos com informações confidenciais ou sensíveis. | Resumo, classificação e perguntas e respostas | Manual   | <a href="https://www.ibm.com/br-pt/products/openpages">https://www.ibm.com/br-pt/products/openpages</a> |

| ID | Fabricante | Produto    | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Uso                                      | Processo | Referência (URL)                                                                                                                                                |
|----|------------|------------|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 7  | LogicGate  | Risk Cloud | Sim | <p>LogicGate has partnered with Ascent, a pioneer in AI-driven regulatory knowledge automation, to create Risk Cloud's Regulatory Compliance Powered By Ascent Application. This Application takes the guessing game and manual work out of your compliance program. With access to over 125 regulators and fueled by AI that delivers obligations and rule changes targeted to your firm, this pre-built Application gives you confidence that you have eliminated all regulatory gaps in your process. Automatically reducing thousands of regulations down to a manageable size allows your teams to supercharge their productivity, be more accurate, and pivot to more strategic work.</p> | Catálogo de regras e gerência de mudança | Manual   | <a href="https://www.logicgate.com/blog/how-ai-can-improve-regulatory-compliance/">https://www.logicgate.com/blog/how-ai-can-improve-regulatory-compliance/</a> |

|   |              |         |     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                     |
|---|--------------|---------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|   |              |         |     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                     |
| 8 | MetricStream | AiSPIRE | Sim | <p>MetricStream AiSPIRE is the industry's first AI-powered, knowledge-centric GRC. AiSPIRE will leverage large language models, GRC ontology-based knowledge graphs, and generative AI capabilities to unlock the full potential of an organization's existing GRC and transactional data. By providing cognitive insights across all aspects of enterprise GRC, AiSPIRE enables greater efficiency, as well as predictive and data-driven decision-making. AiSPIRE connects with multiple systems and data sources and creates cognitive insights from underlying GRC data to improve the performance of existing GRC programs. Unlike other GRC tools that rely on manually defined rules and workflows, AiSPIRE harnesses the power of a customer's data to train advanced</p> | <p>Perguntas e respostas e Detecção de anomalias</p> <p>Automático</p> <p><a href="https://www.metricstream.com/products/AiSPIRE.html">https://www.metricstream.com/products/AiSPIRE.html</a>,<br/> <a href="https://www.metricstream.com/blog/grc-ai-risk-audit-compliance-management.html">https://www.metricstream.com/blog/grc-ai-risk-audit-compliance-management.html</a></p> |

| ID | Fabricante | Produto | IA | O que eles dizem                                                                                                                                                                                                                                                                                                                | Uso | Processo | Referência (URL) |
|----|------------|---------|----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|----------|------------------|
|    |            |         |    | machine learning models and AI. It then continuously senses risk, audit, and control deficiencies, duplicate risks and controls that can be removed, patterns of over-testing and under-testing of controls, and allows proactive planning and prioritization of risk assessments, control testing, issue, and action planning. |     |          |                  |

| ID | Fabricante | Produto   | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                      | Uso                   | Processo | Referência (URL)                                                                                                                                                                                            |
|----|------------|-----------|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 9  | NAVEX      | NAVEX One | Sim | NAVEX, the global leader in integrated risk and compliance management software, today announced the availability of NAVEX One Compliance Assistant. This innovative AI-driven solution delivers instant, secure, and accurate responses in natural language to employee questions regarding company policies and procedures. The functionality makes an organization's compliance program more accessible and personal for all users. | Perguntas e respostas | Manual   | <a href="https://www.navex.com/en-us/products/navex-ethics-compliance/ai-employee-compliance-assistant/">https://www.navex.com/en-us/products/navex-ethics-compliance/ai-employee-compliance-assistant/</a> |

| ID | Fabricante | Produto                        | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Uso           | Processo   | Referência (URL)                                                                                                                                                                                |
|----|------------|--------------------------------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 10 | OneTrust   | GRC & Security Assurance Cloud | Sim | <p>OneTrust Data Discovery: NER identifies exact named entities such as people, organizations, and locations within unstructured data.</p> <p>Optical character recognition (OCR) AI models extract characters from images, including printed or handwritten text.</p> <p>Security classifiers for API keys for AWS, Azure, and Google, plus encryption keys and secrets, passwords, and usernames reduce the risk of exposure through a data leak or breach.</p> <p>Regulatory intelligence enhances jurisdictional insights by correlating classified personally identifiable information (PII) with applicable laws and regulations, such as CPRA.</p> | Classificação | Automático | <a href="https://www.onetrust.com/news/onetrust-enhances-machine-learning-powered-classification/">https://www.onetrust.com/news/onetrust-enhances-machine-learning-powered-classification/</a> |

| ID | Fabricante | Produto                   | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                | Uso              | Processo | Referência (URL)                                                                                                                                                    |
|----|------------|---------------------------|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 11 | Resolver   | Risk Intellinget Platform | Não | Não está claro, não foi encontrado na documentação do produto.                                                                                                                                                                                                                                                                                                  | N/A              | N/A      | N/A                                                                                                                                                                 |
| 12 | Riskonnect | Riskonnect GRC            | Sim | Use artificial intelligence to create a risk record within Riskonnect.                                                                                                                                                                                                                                                                                          | Geração de texto | Manual   | <a href="https://riskonnect.com/governance-risk-compliance/integrating-ai-into-grc/">https://riskonnect.com/governance-risk-compliance/integrating-ai-into-grc/</a> |
| 13 | SAI360     | SAI360 GRC                | Sim | Our GRC software is powered by cutting-edge technology, featuring seamless workflow, robust integrations, advanced reporting, AI capabilities, top-tier security, and a cloud-based platform. Best practice modules and professional services further enhance its effectiveness, providing a comprehensive solution for streamlined and secure risk management. | N/A              | N/A      | <a href="https://www.sai360.com/solutions/integrated-grc">https://www.sai360.com/solutions/integrated-grc</a>                                                       |

| ID | Fabricante | Produto        | IA  | O que eles dizem                                                                                                                                                                                                                                                                                                                                                        | Uso           | Processo | Referência (URL)                                                                                                                                                                                                                                                                                      |
|----|------------|----------------|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 14 | ServiceNow | ServiceNow GRC | Sim | The Governance, Risk, and Compliance: Predictive Intelligence plugin uses artificial intelligence (AI) and machine learning (ML) to predict the related risk statements on the risk records. The Predictive Intelligence plugin uses the similarity capability in the Predictive Intelligence engine to compare the risk name and description with the risk statements. | Classificação | Manual   | <a href="https://docs.servicenow.com/bundle/washingtondc-governance-risk-compliance/page/product/grc-workspace-risk/concept/categories-of-risks.html">https://docs.servicenow.com/bundle/washingtondc-governance-risk-compliance/page/product/grc-workspace-risk/concept/categories-of-risks.html</a> |
| 15 | Workiva    | Workiva        | Não | Não está claro, não foi encontrado na documentação do produto.                                                                                                                                                                                                                                                                                                          | N/A           | N/A      | N/A                                                                                                                                                                                                                                                                                                   |

Fonte: Elaborado pelo autor.

## APÊNDICE C – ORGANIZAÇÃO DO EXPERIMENTO

Os notebooks com os algoritmos e validações estão disponíveis no repositório de códigos<sup>74</sup> sob a pasta “code/src” e “code/data”. A seguir são resumidos a organização do repositório, dos notebooks, dos dados utilizados pelos algoritmos e como reproduzir o experimento.

### Organização dos arquivos de código-fonte

Os arquivos contendo o código executável está no repositório de código do autor no github - <https://github.com/asantos2000/master-degree-santos-anderson> - ramificação: main.

O repositório está organizado da seguinte forma:

- grc-suppliers - Arquivos de suporte para esta dissertação com informações sobre os fornecedores de ferramentas de GRC avaliados em Março/2025;
- ontologies – Arquivos com as ontologias usadas ou avaliadas durante o desenvolvimento da dissertação;
- rsl – Arquivos de referências bibliométricas utilizadas na revisão sistemática da literatura desta dissertação;
- code – pasta raiz contendo o código-fonte, dados e scripts de suporte para o experimento;
  - data – Contém os dados usados e criados durante a execução do experimento:
    - checkpoints – arquivos do tipo json contendo os dados gerados pelo experimento; essa pasta é usada pelo algoritmo e seus arquivos são movidos para outras pastas, ver arquivos READEM.md em cada pasta checkpoint;
    - checkpoints\_classification – Arquivos que foram gerados pelo notebook (1 a 10) chap\_6\_semantic\_annotation\_rules\_classification.ipynb. O \*-0.json é a entrada para a execução desse notebook;

---

<sup>74</sup> Endereço: <https://github.com/asantos2000/master-degree-santos-anderson>

- checkpoints\_evaluation – Contém os arquivos resultantes da execução do notebook src/chap\_7\_validation\_rules\_transformation.ipynb;
- checkpoints\_extraction – Contém os arquivos resultantes da execução do notebook src/chap\_7\_validation\_elements\_extraction;
- checkpoints\_transform – Contém os arquivos resultantes da execução do notebook chap\_6\_nlp2sbvr\_transform.ipynb e o arquivo \*-0.json que é a entrada para a execução deste notebook.
- Arquivos de suporte – Arquivos de suporte a execução dos algoritmos são apresentados na seção “Arquivos de suporte a execução dos algoritmos”.
- labs – notebooks e scripts com experimentos e utilitários;
- src – Contém os arquivos executáveis do experimento:
  - app – Aplicação para apresentação dos resultados do experimento;
  - chap\_6\_cfr2sbvr\_modules.ipynb – Arquivos com os módulos compartilhados pelos notebook desenvolvido para o experimento;
  - chap\_6\_create\_kg.ipynb – Script para configuração inicial do banco de dados do KG;
  - chap\_6\_nlp2sbvr\_elements\_association\_creation.ipynb – Notebook que insere os elementos e metadados do checkpoint mais recente no KG. A versão chap\_6\_nlp2sbvr\_elements\_association\_creation\_best.ipynb utiliza os dados criados pelo chap\_7\_validation\_support.ipynb para selecionar os resultados com maior pontuação e criar o KG;
  - chap\_6\_nlp2sbvr\_transform.ipynb – Transforma as declarações em linguagem natural dos elementos para SBVR seguindo a taxonomia de Witt (2012);

- chap\_6\_semantic\_annotation\_elements\_extraction.ipynb –  
Extrai os elementos das seções da CFR;
- chap\_6\_semantic\_annotation\_rules\_classification.ipynb –  
Classifica os elementos extraídos segundo a taxonomia de  
Witt (2012);
- chap\_7\_validation\_elements\_extraction.ipynb – Avalia o  
processo de extração;
- chap\_7\_validation\_rules\_classification.ipynb – Avalia o  
processo de classificação;
- chap\_7\_validation\_rules\_transformation.ipynb – Avalia o  
processo de transformação;
- sbvr\_xsd\_to\_rdf.py – Transforma o xsd do SBVR em rdf.  
Experimental, usado para o desenvolvimento de ontologia  
SBVR;
- chap\_7\_validation\_support.ipynb – Contém scripts SQL  
para dupla verificação da consistência dos dados dos  
*checkpoints* e dos notebooks de validação e views para a  
aplicação cfr2sbvr\_inspect, entre as views, a  
RAW\_LLM\_VALIDATION\_BEST\_VW filtra os elementos  
com melhor resultado na avaliação;
- Demais pastas – As pastas: token\_estimator, checkpoint,  
configuration, llm\_query, logging\_setup,  
rules\_taxonomy\_provider, são módulos python criadas pelo  
notebook chap\_6\_cfr2sbvr\_modules.ipynb. Não devem ser  
editadas diretamente.
  - scripts – scripts para suporte ao banco de dados AllegroGraph;  
outputs – contém arquivos gerados durante a execução dos notebooks,  
na maioria planilhas excel;

### Como executar o projeto

Para executar o projeto faça um clone do repositório e siga as instruções no arquivo de README.md da pasta src/.

## Arquivos de suporte a execução dos algoritmos

Os arquivos de suporte estão na pasta data/ e são listados de acordo com sua função:

- classify\_subtypes.json e classify\_subtypes.yaml – Contém a taxonomia de Witt (2012) e os identificadores dos *templates* e exemplos. O arquivo json pode ser transformado no arquivo yaml usando o script json-to-yaml do notebook labs/lab\_3.ipynb.
- documents\_true\_table.json – Tabela verdade, contém os dados usados na avaliação do processo (notebooks chap\_7\*).
- witt\_examples.yaml – Contém os exemplos extraídos do livro do Witt (2012).
- witt\_subtemplates.yaml - Contém os *subtemplates* extraídos do livro do Witt (2012).
- witt\_template\_subtemplate\_relationship.yaml – Relaciona os *templates* a *subtemplates*.
- witt\_templates.yaml – Templates extraídos do livro de Witt (2012).
- US\_LegalReference.ttl – Parte da ontologia FRO-CFR carregada no grafo.
- Code\_Federal\_Regulations.ttl – Ontologia FRO-CFR carregada no grafo.
- sbvr-dtc-19-05-32-ontology-v1.ttl – SBVR ontologia experimental carregada no KG.
- prod-fibo-quickstart-2024Q3.ttl – Ontologia FIBO carregada no KG.
- FRO\_CFR\_Title\_17\_Part\_275.ttl – Seções da CFR seguindo a ontologia FRO.

## Avaliação da classificação da taxonomia de regras

Quadro 40 - Taxonomia de regras definidas por Witt (2012)

### 9.2 - Definitional rules (15)

9.2.1 - Formal term definitions

9.2.1.1 - Formal intensional definitions

9.2.1.2 - Formal extensional definitions

9.2.1.3 - Symbolic literal definitions

9.2.2 - Categorization scheme enumerations  
9.2.3 - Category transition constraints  
9.2.4 - Complex concept structure rules  
9.2.4.1 - Complex concept cardinality rules  
9.2.4.2 - Complex concept equivalence rules  
9.2.4.3 - Complex concept set constraints  
9.2.5 - Valid value definitions  
9.2.6 - Data calculation rules  
9.2.6.1 - Data calculation algorithms  
9.2.6.2 - Conversion factor definitions  
9.2.7 - Standard format definitions

### **9.3 - Data rules (36)**

9.3.1 - Data cardinality rules  
9.3.1.1 - Mandatory data rules  
9.3.1.1.1 - Mandatory data item rules  
9.3.1.1.2 - Mandatory option selection rules  
9.3.1.1.3 - Mandatory group rules  
9.3.1.2 - Prohibited data rules  
9.3.1.3 - Maximum cardinality rules  
9.3.1.4 - Multiple data rules  
9.3.1.5 - Dependent cardinality rules  
9.3.2 - Data content rules  
9.3.2.1 - Value set rules  
9.3.2.1.1 - Value set rules constraining single data items  
9.3.2.1.2 - Value set rules constraining combinations of data items  
9.3.2.2 - Range rules  
9.3.2.3 - Equality rules  
9.3.2.4 - Uniqueness constraints  
9.3.2.4.1 - Uniqueness constraints constraining single data items  
9.3.2.4.2 - Uniqueness constraints constraining combinations of data items  
9.3.2.4.3 - Uniqueness constraints constraining sets of data items  
9.3.2.5 - Data consistency rules  
9.3.2.5.1 - Data consistency rules constraining a combination of data items  
9.3.2.5.2 - Data consistency rules constraining a set function  
9.3.2.5.3 - Data consistency rules constraining a set  
9.3.2.6 - Temporal data constraints  
9.3.2.6.1 - Simple temporal data constraints  
9.3.2.6.2 - Temporal data non-overlap constraints  
9.3.2.6.3 - Temporal data completeness constraints  
9.3.2.6.4 - Temporal data inclusion constraints

9.3.2.6.5 - Temporal single record constraints

9.3.2.6.6 - Day type constraints

9.3.2.7 - Spatial data constraints

9.3.2.8 - Data item format rules

9.3.3 - Data update rules

9.3.3.1 - Data update prohibition rules

9.3.3.2 - State transition constraints

9.3.3.3 - Monotonic transition constraints

#### **9.4 - Activity rules (11)**

9.4.1 - Activity restriction rules

9.4.1.1 - Rules restricting when an activity can occur

9.4.1.1.1 - Activity time limit rules

9.4.1.1.2 - Activity exclusion period rules

9.4.1.1.3 - Activity obligation rule

9.4.1.2 - Activity pre-condition rules

9.4.1.3 - Activity prohibition rules

9.4.1.4 - Information retention rules

9.4.1.5 - Activity conflict rules

9.4.2 - Process decision rules

9.4.3 - Activity obligation rules

#### **9.5 - Party rules (4)**

9.5.1 - Party restriction rules

9.5.2 - Role separation and binding rules

9.5.3 - Information access rules

9.5.4 - Responsibility rules

Fonte: Elaborado pelo autor.

O Quadro 40 resume as 66 categorias da taxonomia de Witt (2012) usada nos algoritmos de classificação e transformação, essa taxonomia foi transformada em arquivos json e yaml, juntamente com exemplos de regras extraídas do livro do autor.

## APÊNDICE D – RESULTADO DA TRANSFORMAÇÃO

Neste apêndice estão listados os resultados da transformação de NL para SBVR. Apenas os elementos que obtiveram a pontuação mais alta em todas as medidas foram listados. Esses foram os elementos carregados no KG.

Quadro 41 - Elementos transformados

| Elemento   | Seção     | Parágrafos | Título                                                 | Declaração (original)                                                                                                                                                                                                                                                           | Declaração (transformada)                                                                                                                                                                                                            |
|------------|-----------|------------|--------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fact_Types | § 275.0-2 | {'(a)'}    | Service of Process on Non-Resident Investment Advisers | A person may serve process, pleadings, or other papers on a non-resident investment adviser, or on a non-resident general partner or non-resident managing agent of an investment adviser by serving any or all of its appointed agents.                                        | A person may serve process, pleadings, or other papers on a non-resident investment adviser, or on a non-resident general partner or non-resident managing agent of an investment adviser by definition at most one appointed agent. |
| Fact_Types | § 275.0-2 | {'(a)(1)'} | Service of Process                                     | A person may serve a non-resident investment adviser, non-resident general partner, or non-resident managing agent by furnishing the Commission with one copy of the process, pleadings, or papers, for each named party, and one additional copy for the Commission's records. | A person may serve a non-resident investment adviser, non-resident general partner, or non-resident managing agent by definition exactly one copy of the process, pleadings, or papers for each named party.                         |
| Fact_Types | § 275.0-2 | {'(b)(1)'} | Managing Agent Definition                              | Managing agent means any person, including a trustee, who directs or manages, or who participates in directing or managing, the affairs of any unincorporated organization or association other than a partnership.                                                             | A managing agent is by definition a person who directs or manages, or who participates in directing or managing, the affairs of any unincorporated organization or association other than a partnership.                             |
| Fact_Types | § 275.0-2 | {'(b)(2)'} | Non-resident Definition                                | Non-resident means an individual who resides in any place not subject to the jurisdiction of the United States; a                                                                                                                                                               | A non-resident is by definition an individual who resides in any place not subject to the jurisdiction of the United                                                                                                                 |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>     | <b>Título</b>                                                                  | <b>Declaração (original)</b>                                                                                                                                                                                                                                                                                                                               | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                                                                    |
|-----------------|--------------|-----------------------|--------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 |              |                       |                                                                                | corporation that is incorporated in or that has its principal office and place of business in any place not subject to the jurisdiction of the United States; and a partnership or other unincorporated organization or association that has its principal office and place of business in any place not subject to the jurisdiction of the United States. | States, a corporation that is incorporated in or that has its principal office and place of business in any place not subject to the jurisdiction of the United States, or a partnership or other unincorporated organization or association that has its principal office and place of business in any place not subject to the jurisdiction of the United States. |
| Fact_Types      | § 275.0-2    | {'(b)(3)'}<br>{'(d)'} | Principal office and place of business definition<br>Definition of Application | Principal office and place of business has the same meaning as in § 275.203A-3(c) of this chapter.<br>For purposes of this rule, an application means any application for an order of the Commission under the Act other than an application for registration as an investment adviser.                                                                    | A principal office and place of business is by definition the same as in § 275.203A-3(c) of this chapter.<br>An application is by definition an application for an order of the Commission under the Act, other than an application for registration as an investment adviser.                                                                                      |
| Fact_Types      | § 275.0-7    | {'(a)'}               | Definition of Small Business or Small Organization                             | The term small business or small organization for purposes of the Investment Advisers Act of 1940 shall mean an investment adviser that has assets under management of less than \$25 million, or such higher amount as the Commission may by rule deem appropriate.                                                                                       | A small business or small organization is by definition an investment adviser that has assets under management of less than \$25 million, or such higher amount as the Commission may by rule deem appropriate.                                                                                                                                                     |
| Fact_Types      | § 275.0-7    | {'(a)(1)'}            | Assets Under Management                                                        | An investment adviser has assets under management, as defined under Section 203A(a)(3) of the Act and reported on its annual updating amendment to Form ADV, of less than \$25 million, or such higher                                                                                                                                                     | The assets under management of an investment adviser is by definition less than \$25 million, or such higher amount as the Commission may by rule deem appropriate.                                                                                                                                                                                                 |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>        | <b>Título</b>                                                      | <b>Declaração (original)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                   | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-----------------|--------------|--------------------------|--------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 |              |                          |                                                                    | amount as the Commission may by rule deem appropriate.                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Fact_Types      | § 275.0-7    | {'(a)(2)'}<br>{'(a)(3)'} | Investment Adviser Asset Requirement<br>Investment Adviser Control | An investment adviser did not have total assets of \$5 million or more on the last day of the most recent fiscal year.<br><br>An investment adviser does not control, is not controlled by, and is not under common control with another investment adviser that has assets under management of \$25 million or more, or any person (other than a natural person) that had total assets of \$5 million or more on the last day of the most recent fiscal year. | The total assets of an investment adviser is by definition no more than \$5 million on the last day of the most recent fiscal year.<br><br>An investment adviser does not control, is not controlled by, and is not under common control with another investment adviser that has assets under management of at least \$25 million, or any person (other than a natural person) that had total assets of at least \$5 million on the last day of the most recent fiscal year. |
| Fact_Types      | § 275.0-7    | {'(b)(1)'}               | Control Definition                                                 | Control means the power, directly or indirectly, to direct the management or policies of a person, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                                                                                                                         | Control is by definition the power to direct the management or policies of a person, either directly or indirectly, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                                                                                                                       |
| Fact_Types      | § 275.0-7    | {'(b)(1)(i)(A)'}         | Control Presumption                                                | A person is presumed to control a corporation if the person directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.                                                                                                                                                                                                                                                                                 | A person is by definition presumed to control a corporation that directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.                                                                                                                                                                                                                                                                                           |
| Fact_Types      | § 275.0-7    | {'(b)(1)(ii)'}           | Presumption of Control in Partnership                              | A person is presumed to control a partnership if the person has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the partnership.                                                                                                                                                                                                                                                                               | A person is by definition presumed to control a partnership that either has the right to receive upon dissolution or has contributed 25 percent or more of the capital of the partnership.                                                                                                                                                                                                                                                                                    |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>                 | <b>Título</b>                                               | <b>Declaração (original)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                         | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-----------------|--------------|-----------------------------------|-------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fact_Types      | § 275.0-7    | {'(b)(1)(iii)'}<br>{'(b)(1)(iv)'} | Presumption of Control in LLC<br>Control of Trust by Person | A person is presumed to control a limited liability company (LLC) if the person directly or indirectly has the right to vote 25 percent or more of a class of the interests of the LLC, has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the LLC, or is an elected manager of the LLC.<br>A person is presumed to control a trust if the person is a trustee or managing agent of the trust.                      | A person is by definition presumed to control a limited liability company (LLC) that either directly or indirectly has the right to vote 25 percent or more of a class of the interests of the LLC, has the right to receive upon dissolution, or has contributed 25 percent or more of the capital of the LLC, or is an elected manager of the LLC.<br>A person is by definition presumed to control a trust if the person is a trustee or managing agent of the trust.                                      |
| Fact_Types      | § 275.0-7    | {'(b)(2)'}<br>{'(a)(2)'}          | Total Assets Definition<br>Secretary of the Commission      | Total assets means the total assets as shown on the balance sheet of the investment adviser or other person described above under paragraph (a)(3) of this section, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated, whichever is larger.<br>The Secretary of the Commission is responsible for forwarding copies of process, pleadings, or other papers to each named party by registered or certified mail. | A total assets is by definition the total assets as shown on the balance sheet of the investment adviser or other person described above under paragraph (a)(3) of this section, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated, whichever is larger.<br>A Secretary of the Commission is by definition a person who is responsible for forwarding copies of process, pleadings, or other papers to each named party by registered or certified mail. |
| Names           | § 275.0-2    | {'(a)(3)'}                        | Secretary                                                   | The Secretary of the Commission is responsible for forwarding copies of process, pleadings, or other papers to each named party by registered or certified mail.                                                                                                                                                                                                                                                                                                     | A Secretary of the Commission is by definition a person who is responsible for forwarding copies of process, pleadings, or other papers to each named party by registered or certified mail.                                                                                                                                                                                                                                                                                                                  |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>            | <b>Título</b>                           | <b>Declaração (original)</b>                                                                                                                                                                                                                                                                                     | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                                       |
|-----------------|--------------|------------------------------|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Names           | § 275.0-5    | {'(a)'}{'(a)'}               | Federal Register                        | The official journal where the notice of initiation of the proceeding is published.                                                                                                                                                                                                                              | A Federal Register is by definition the official journal that publishes the notice of initiation of the proceeding.                                                                                                                                                                                                                    |
| Names           | § 275.0-5    | {'(a)', '(d)', '(c)', '(b)'} | Commission                              | The regulatory body responsible for issuing orders and conducting hearings.                                                                                                                                                                                                                                      | A Commission is by definition a regulatory body that is responsible for issuing orders and conducting hearings.                                                                                                                                                                                                                        |
| Names           | § 275.0-5    | {'(b)', '(d)', '(a)', '(c)'} | Commission                              | The regulatory body responsible for issuing orders and conducting hearings.                                                                                                                                                                                                                                      | A Commission is by definition a regulatory body that is responsible for issuing orders and conducting hearings.                                                                                                                                                                                                                        |
| Names           | § 275.0-5    | {'(d)'}                      | Act                                     | The legislation under which the Commission operates and applications are filed.                                                                                                                                                                                                                                  | An Act is by definition the legislation under which the Commission operates and applications are filed.                                                                                                                                                                                                                                |
| Operative_Rules | § 275.0-2    | {'(a)(2)'}                   | Forwarding Copies to Named Parties      | The Secretary of the Commission (Secretary) will promptly forward a copy to each named party by registered or certified mail at that party's last address filed with the Commission.                                                                                                                             | The Secretary of the Commission must forward a copy to each named party if the copy is by registered or certified mail at that party's last address filed with the Commission.                                                                                                                                                         |
| Operative_Rules | § 275.0-2    | {'(a)(3)'}                   | Evidence of Service Certification       | If the Secretary certifies that the Commission was served with process, pleadings, or other papers pursuant to paragraph (a)(1) of this section and forwarded these documents to a named party pursuant to paragraph (a)(2) of this section, this certification constitutes evidence of service upon that party. | The certification by the Secretary must constitute evidence of service upon a named party if the Secretary certifies that the Commission was served with process, pleadings, or other papers pursuant to paragraph (a)(1) of this section and forwarded these documents to a named party pursuant to paragraph (a)(2) of this section. |
| Operative_Rules | § 275.0-5    | {'(a)'}                      | Submission of Facts and Hearing Request | Any interested person may, within the period of time specified therein, submit to the Commission in writing any facts bearing upon the desirability of a hearing on the                                                                                                                                          | An interested person may submit to the Commission in writing any facts bearing upon the desirability of a hearing on the matter only if the period of time specified                                                                                                                                                                   |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>     | <b>Título</b>                   | <b>Declaração (original)</b>                                                                                                                                                                                                             | <b>Declaração (transformada)</b>                                                                                                                                                                                                                    |
|-----------------|--------------|-----------------------|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 |              |                       |                                 | matter and may request that a hearing be held, stating his reasons therefor and the nature of his interest in the matter.                                                                                                                | therein is met, and may request that a hearing be held, stating his reasons therefor and the nature of his interest in the matter.                                                                                                                  |
| Operative_Rules | § 275.0-5    | {'(b)'}<br>{'(b)'}    | Order Disposition               | An order disposing of the matter will be issued as of course following the expiration of the period of time referred to in paragraph (a) of this section, unless the Commission thereafter orders a hearing on the matter.               | An order disposing of the matter may be issued as of course following the expiration of the period of time referred to in paragraph (a) of this section only unless the Commission thereafter orders a hearing on the matter.                       |
| Operative_Rules | § 275.0-5    | {'(c)'}<br>{'(c)'}    | Hearing Order by the Commission | The Commission will order a hearing on the matter, if it appears that a hearing is necessary or appropriate in the public interest or for the protection of investors, upon the request of any interested person or upon its own motion. | A hearing on the matter may be ordered by the Commission only if it appears that a hearing is necessary or appropriate in the public interest or for the protection of investors, upon the request of any interested person or upon its own motion. |
| Terms           | § 275.0-2    | {'(a)(1)'}<br>{'(a)'} | Commission's records            | Records maintained by the Commission, which include copies of process, pleadings, or other papers served.                                                                                                                                | The Commission's records are by definition the records maintained by the Commission, which include copies of process, pleadings, or other papers served.                                                                                            |
| Terms           | § 275.0-2    | {'(a)(3)'}<br>{'(a)'} | Evidence of service             | The certification by the Secretary that the Commission was served with process, pleadings, or other papers and forwarded these documents to a named party.                                                                               | An evidence of service is by definition the certification by the Secretary that the Commission was served with process, pleadings, or other papers, or forwarded these documents to a named party.                                                  |
| Terms           | § 275.0-2    | {'(b)(1)'}<br>{'(b)'} | Managing agent                  | Any person, including a trustee, who directs or manages, or who participates in directing or managing, the affairs of any                                                                                                                | A managing agent is by definition a person who directs or manages, or who participates in directing or managing, the                                                                                                                                |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>                              | <b>Título</b>                | <b>Declaração (original)</b>                                                                                                                                                                                                   | <b>Declaração (transformada)</b>                                                                                                                                                                                                                               |
|-----------------|--------------|------------------------------------------------|------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                 |              |                                                |                              | unincorporated organization or association other than a partnership.                                                                                                                                                           | affairs of any unincorporated organization or association other than a partnership.                                                                                                                                                                            |
| Terms           | § 275.0-2    | {'(b)(2)'}<br>{'(b)'}                          | Non-resident                 | An individual, corporation, partnership, or other unincorporated organization or association that resides or has its principal office and place of business in any place not subject to the jurisdiction of the United States. | A non-resident is by definition an individual, corporation, partnership, or other unincorporated organization or association that resides or has its principal office and place of business in any place not subject to the jurisdiction of the United States. |
| Terms           | § 275.0-5    | {'(a)'}<br>{'(a)'}                             | Initiation of the proceeding | The process that begins when a notice is published in the Federal Register indicating the earliest date upon which an order disposing of the matter may be entered.                                                            | An initiation of the proceeding is by definition the process that begins when a notice is published in the Federal Register indicating the earliest date upon which an order disposing of the matter may be entered.                                           |
| Terms           | § 275.0-5    | {'(a)', '(b)'}<br>{'(a)', '(b)'}               | Period of time               | The timeframe specified in the notice during which interested persons can submit facts or request a hearing.                                                                                                                   | The period of time of a notice is by definition the timeframe during which interested persons can submit facts or request a hearing.                                                                                                                           |
| Terms           | § 275.0-5    | {'(a)', '(c)'}<br>{'(a)', '(c)'}               | Interested person            | Any person who may submit facts or request a hearing on the matter.                                                                                                                                                            | An interested person is by definition a person who may submit facts or request a hearing on the matter.                                                                                                                                                        |
| Terms           | § 275.0-5    | {'(a)', '(c)', '(b)'}<br>{'(a)', '(c)', '(b)'} | Hearing                      | A proceeding ordered by the Commission if necessary or appropriate in the public interest or for the protection of investors.                                                                                                  | A hearing is by definition a proceeding ordered by the Commission if necessary or appropriate in the public interest or for the protection of investors.                                                                                                       |
| Terms           | § 275.0-5    | {'(a)', '(d)', '(b)'}<br>{'(a)', '(d)', '(b)'} | Order                        | A decision issued by the Commission disposing of the matter.                                                                                                                                                                   | An Order is by definition a decision that is issued by the Commission disposing of the matter.                                                                                                                                                                 |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>     | <b>Título</b>           | <b>Declaração (original)</b>                                                                                                                                                                                                     | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                |
|-----------------|--------------|-----------------------|-------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Terms           | § 275.0-5    | {'(b)', '(a)'}        | Period of time          | The timeframe specified in the notice during which interested persons can submit facts or request a hearing.                                                                                                                     | The period of time is by definition the timeframe specified in the notice during which interested persons can submit facts or request a hearing.                                                                                                                                                                |
| Terms           | § 275.0-5    | {'(b)', '(a)', '(c)'} | Hearing                 | A proceeding ordered by the Commission if necessary or appropriate in the public interest or for the protection of investors.                                                                                                    | A hearing is by definition a proceeding ordered by the Commission if necessary or appropriate in the public interest or for the protection of investors.                                                                                                                                                        |
| Terms           | § 275.0-5    | {'(b)', '(d)', '(a)'} | Order                   | A decision issued by the Commission disposing of the matter.                                                                                                                                                                     | An Order is by definition a decision that is issued by the Commission disposing of the matter.                                                                                                                                                                                                                  |
| Terms           | § 275.0-5    | {'(c)'}               | Public interest         | A consideration for ordering a hearing if it appears necessary or appropriate.                                                                                                                                                   | A public interest is by definition a consideration for ordering a hearing that appears necessary or appropriate.                                                                                                                                                                                                |
| Terms           | § 275.0-5    | {'(d)'}               | Order of the Commission | An order issued by the Commission under the Act.                                                                                                                                                                                 | An order is by definition the order of the Commission that is issued under the Act.                                                                                                                                                                                                                             |
| Terms           | § 275.0-7    | {'(a)'}               | Small organization      | An investment adviser with assets under management of less than \$25 million, or such higher amount as the Commission may deem appropriate, and other specific conditions.                                                       | The assets under management of an investment adviser is by definition less than \$25 million, or such higher amount as the Commission may deem appropriate.                                                                                                                                                     |
| Terms           | § 275.0-7    | {'(a)(2)', '(b)(2)'}  | Total assets            | The total assets as shown on the balance sheet of the investment adviser or other person described, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated, whichever is larger. | The total assets of an investment adviser or other person is by definition calculated as the larger of the total assets shown on the balance sheet of the investment adviser or other person described, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated. |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafos</b>               | <b>Título</b>                   | <b>Declaração (original)</b>                                                                                                                                                                                                                                                                                                              | <b>Declaração (transformada)</b>                                                                                                                                                                                                                                                                                                                                          |
|-----------------|--------------|---------------------------------|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Terms           | § 275.0-7    | {'(a)(3)', '(b)(1)'}            | Control                         | The power, directly or indirectly, to direct the management or policies of a person, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                  | A control is by definition a power that directly or indirectly directs the management or policies of a person, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                        |
| Terms           | § 275.0-7    | {'(b)(1)', '(a)(3)'}            | Control                         | The power, directly or indirectly, to direct the management or policies of a person, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                  | Control is by definition the power that directly or indirectly directs the management or policies of a person, whether through ownership of securities, by contract, or otherwise.                                                                                                                                                                                        |
| Terms           | § 275.0-7    | {'(b)(1)(i)(A)', '(b)(1)(iii)'} | Right to vote                   | A person is presumed to control a corporation if the person directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.                                                                                                                                                            | A person is by definition presumed to control a corporation if the person directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.                                                                                                                                                                              |
| Terms           | § 275.0-7    | {'(b)(1)(ii)'}                  | Partnership                     | A person is presumed to control a partnership if the person has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the partnership.                                                                                                                                                          | A partnership is by definition a person that is presumed to control a partnership if the person has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the partnership.                                                                                                                                                      |
| Terms           | § 275.0-7    | {'(b)(1)(iii)'}                 | Limited liability company (LLC) | A person is presumed to control a limited liability company (LLC) if the person directly or indirectly has the right to vote 25 percent or more of a class of the interests of the LLC, has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the LLC, or is an elected manager of the LLC. | A limited liability company (LLC) is by definition a company that a person is presumed to control if the person directly or indirectly has the right to vote 25 percent or more of a class of the interests of the LLC, has the right to receive upon dissolution, or has contributed, 25 percent or more of the capital of the LLC, or is an elected manager of the LLC. |

| Elemento | Seção     | Parágrafos                                        | Título                 | Declaração (original)                                                                                                                                                                                                                                                                | Declaração (transformada)                                                                                                                                                                                                                                                                             |
|----------|-----------|---------------------------------------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Terms    | § 275.0-7 | {'(b)(1)(iii)', '(b)(1)(i)(A)'}<br>{'(b)(1)(iv)'} | Right to vote<br>Trust | A person is presumed to control a corporation if the person directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.<br>A person is presumed to control a trust if the person is a trustee or managing agent of the trust. | A person is by definition presumed to control a corporation that directly or indirectly has the right to vote 25 percent or more of a class of the corporation's voting securities.<br>A person is by definition presumed to control a trust that is a trustee or managing agent of the trust.        |
| Terms    | § 275.0-7 | {'(b)(2)', '(a)(2)'}                              | Total assets           | The total assets as shown on the balance sheet of the investment adviser or other person described, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated, whichever is larger.                                                     | The total assets of an investment adviser or other person is by definition calculated as the larger of the total assets shown on the balance sheet of the investment adviser or other person, or the balance sheet of the investment adviser or such other person with its subsidiaries consolidated. |

Fonte: Elaborado pelo autor.

No Quadro 41 estão listados os elementos extraídos de cada seção e parágrafo que tiveram a maior pontuação de SemScore, *Similarity Score* e *Transformation Confidence*, ao lado da versão transformada, são 51 elementos: 16 tipos de fatos, 6 nomes, 5 regras operativas e 24 termos.

Tabela 10 - Pontuações de validação dos elementos

| Elementos  | Seção     | Parágrafos                  | SemScore     | Similarity Score | Transformation Accuracy | Grammar Syntax Accuracy | Transformation Confidence |
|------------|-----------|-----------------------------|--------------|------------------|-------------------------|-------------------------|---------------------------|
| Fact_Types | § 275.0-2 | {'(a)'}<br>{'(a)(1)'}       | 0,93<br>0,92 | 0,70<br>0,85     | 0,50<br>0,70            | 0,90<br>0,90            | 0,90<br>0,90              |
| Fact_Types | § 275.0-2 | {'(b)(1)'}<br>{'(b)(1)(i)'} | 0,87         | 0,95             | 0,95                    | 0,95                    | 0,90                      |

| Elementos       | Seção     | Parágrafos                   | SemScore | Similarity Score | Transformation Accuracy | Grammar Syntax Accuracy | Transformation Confidence |
|-----------------|-----------|------------------------------|----------|------------------|-------------------------|-------------------------|---------------------------|
| Fact_Types      | § 275.0-2 | {'(b)(2)'}                   | 0,92     | 0,95             | 0,95                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-2 | {'(b)(3)'}                   | 0,88     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-5 | {'(d)'}                      | 0,81     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(a)'}                      | 0,92     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(a)(1)'}                   | 0,86     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(a)(2)'}                   | 0,79     | 0,95             | 0,90                    | 1,00                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(a)(3)'}                   | 0,97     | 0,95             | 0,90                    | 1,00                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(1)'}                   | 0,82     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(1)(i)(A)'}             | 0,87     | 0,95             | 0,85                    | 0,90                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(1)(ii)'}               | 0,89     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(1)(iii)'}              | 0,94     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(1)(iv)'}               | 0,83     | 0,95             | 0,90                    | 1,00                    | 0,90                      |
| Fact_Types      | § 275.0-7 | {'(b)(2)'}                   | 0,93     | 0,95             | 0,80                    | 0,90                    | 0,90                      |
| Names           | § 275.0-2 | {'(a)(2)'}                   | 0,88     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Names           | § 275.0-2 | {'(a)(3)'}                   | 0,90     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Names           | § 275.0-5 | {'(a)'}                      | 0,90     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Names           | § 275.0-5 | {'(a)', '(d)', '(c)', '(b)'} | 0,83     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Names           | § 275.0-5 | {'(b)', '(d)', '(a)', '(c)'} | 0,83     | 0,95             | 0,95                    | 0,95                    | 0,90                      |
| Names           | § 275.0-5 | {'(d)'}                      | 0,91     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Operative_Rules | § 275.0-2 | {'(a)(2)'}                   | 0,91     | 0,85             | 0,80                    | 0,95                    | 0,90                      |
| Operative_Rules | § 275.0-2 | {'(a)(3)'}                   | 0,92     | 0,90             | 0,90                    | 0,95                    | 0,90                      |
| Operative_Rules | § 275.0-5 | {'(a)'}                      | 0,93     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Operative_Rules | § 275.0-5 | {'(b)'}                      | 0,93     | 0,90             | 0,80                    | 0,95                    | 0,90                      |
| Operative_Rules | § 275.0-5 | {'(c)'}                      | 0,91     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms           | § 275.0-2 | {'(a)(1)'}                   | 0,92     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms           | § 275.0-2 | {'(a)(3)'}                   | 0,87     | 0,85             | 0,80                    | 0,95                    | 0,90                      |

| Elementos | Seção     | Parágrafos                      | SemScore | Similarity Score | Transformation Accuracy | Grammar Syntax Accuracy | Transformation Confidence |
|-----------|-----------|---------------------------------|----------|------------------|-------------------------|-------------------------|---------------------------|
| Terms     | § 275.0-2 | {'(b)(1)'}                      | 0,86     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-2 | {'(b)(2)'}                      | 0,96     | 10,00            | 10,00                   | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(a)'}                         | 0,94     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(a)', '(b)'}                  | 0,91     | 0,90             | 0,85                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-5 | {'(a)', '(c)'}                  | 0,86     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(a)', '(c)', '(b)'}           | 0,94     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(a)', '(d)', '(b)'}           | 0,85     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(b)', '(a)'}                  | 0,90     | 0,95             | 0,90                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(b)', '(a)', '(c)'}           | 0,94     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(b)', '(d)', '(a)'}           | 0,85     | 0,95             | 0,95                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-5 | {'(c)'}                         | 0,89     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-5 | {'(d)'}                         | 0,85     | 0,90             | 0,80                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(a)'}                         | 0,81     | 0,85             | 0,80                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(a)(2)', '(b)(2)'}            | 0,92     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(a)(3)', '(b)(1)'}            | 0,86     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)', '(a)(3)'}            | 0,88     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)(i)(A)', '(b)(1)(iii)'} | 0,89     | 0,95             | 0,90                    | 1,00                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)(ii)'}                  | 0,92     | 0,80             | 0,70                    | 0,90                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)(iii)'}                 | 0,94     | 0,95             | 0,90                    | 0,95                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)(iii)', '(b)(1)(i)(A)'} | 0,86     | 0,90             | 0,70                    | 0,90                    | 0,90                      |
| Terms     | § 275.0-7 | {'(b)(1)(iv)'}                  | 0,89     | 0,70             | 0,60                    | 0,50                    | 0,80                      |

| Elementos | Seção     | Parágrafos           | SemScore | Similarity Score | Transformation Accuracy | Grammar Syntax Accuracy | Transformation Confidence |
|-----------|-----------|----------------------|----------|------------------|-------------------------|-------------------------|---------------------------|
| Terms     | § 275.0-7 | {'(b)(2)', '(a)(2)'} | 0,91     | 0,95             | 0,90                    | 0,95                    | 0,90                      |

Fonte: Elaborado pelo autor.

A Tabela 10 lista os valores de avaliação entre a declaração original e a declaração transformada, a pontuação de transformation\_confidence indica o quanto a declaração transformada se adequou ao *template*.

Quadro 42 - Classificação e *templates* usados na transformação dos elementos

| Elemento   | Seção     | Parágrafo        | transformation_template_ids | statement_classification_type | statement_classification_subtype |
|------------|-----------|------------------|-----------------------------|-------------------------------|----------------------------------|
| Fact_Types | § 275.0-2 | {'(a)'}          | {'T12'}                     | Definitional rules            | Complex concept structure rules  |
| Fact_Types | § 275.0-2 | {'(a)(1)'}       | {'T12'}                     | Definitional rules            | Complex concept structure rules  |
| Fact_Types | § 275.0-2 | {'(b)(1)'}       | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-2 | {'(b)(2)'}       | {'T8'}                      | Definitional rules            | Formal extensional definitions   |
| Fact_Types | § 275.0-2 | {'(b)(3)'}       | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-5 | {'(d)'}          | {'T8'}                      | Definitional rules            | Formal extensional definitions   |
| Fact_Types | § 275.0-7 | {'(a)'}          | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(a)(1)'}       | {'T15'}                     | Definitional rules            | Valid value definitions          |
| Fact_Types | § 275.0-7 | {'(a)(2)'}       | {'T15'}                     | Definitional rules            | Valid value definitions          |
| Fact_Types | § 275.0-7 | {'(a)(3)'}       | {'T12'}                     | Definitional rules            | Complex concept structure rules  |
| Fact_Types | § 275.0-7 | {'(b)(1)'}       | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(b)(1)(i)(A)'} | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(b)(1)(ii)'}   | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(b)(1)(iii)'}  | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(b)(1)(iv)'}   | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Fact_Types | § 275.0-7 | {'(b)(2)'}       | {'T8'}                      | Definitional rules            | Formal extensional definitions   |
| Names      | § 275.0-2 | {'(a)(2)'}       | {'T7'}                      | Definitional rules            | Formal term definitions          |

| Elemento        | Seção     | Parágrafo                    | transformation_template_ids | statement_classification_type | statement_classification_subtype |
|-----------------|-----------|------------------------------|-----------------------------|-------------------------------|----------------------------------|
| Names           | § 275.0-2 | {'(a)(3)'}                   | {'T7'}                      | Definitional rules            | Formal term definitions          |
| Names           | § 275.0-5 | {'(a)'}                      | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Names           | § 275.0-5 | {'(a)', '(d)', '(c)', '(b)'} | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Names           | § 275.0-5 | {'(b)', '(d)', '(a)', '(c)'} | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Names           | § 275.0-5 | {'(d)'}                      | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Operative_Rules | § 275.0-2 | {'(a)(2)'}                   | {'T56'}                     | Party rules                   | Responsibility rules             |
| Operative_Rules | § 275.0-2 | {'(a)(3)'}                   | {'T56'}                     | Party rules                   | Responsibility rules             |
| Operative_Rules | § 275.0-5 | {'(a)'}                      | {'T53'}                     | Party rules                   | Party restriction rules          |
| Operative_Rules | § 275.0-5 | {'(b)'}                      | {'T49'}                     | Activity rules                | Activity obligation rules        |
| Operative_Rules | § 275.0-5 | {'(c)'}                      | {'T50'}                     | Activity rules                | Activity pre-condition rules     |
| Terms           | § 275.0-2 | {'(a)(1)'}                   | {'T8'}                      | Definitional rules            | Formal extensional definitions   |
| Terms           | § 275.0-2 | {'(a)(3)'}                   | {'T8'}                      | Definitional rules            | Formal extensional definitions   |
| Terms           | § 275.0-2 | {'(b)(1)'}                   | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-2 | {'(b)(2)'}                   | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(a)'}                      | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(a)', '(b)'}               | {'T15'}                     | Definitional rules            | Valid value definitions          |
| Terms           | § 275.0-5 | {'(a)', '(c)'}               | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(a)', '(c)', '(b)'}        | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(a)', '(d)', '(b)'}        | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(b)', '(a)'}               | {'T15'}                     | Definitional rules            | Valid value definitions          |
| Terms           | § 275.0-5 | {'(b)', '(a)', '(c)'}        | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(b)', '(d)', '(a)'}        | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(c)'}                      | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-5 | {'(d)'}                      | {'T7'}                      | Definitional rules            | Formal intensional definitions   |
| Terms           | § 275.0-7 | {'(a)'}                      | {'T15'}                     | Definitional rules            | Valid value definitions          |
| Terms           | § 275.0-7 | {'(a)(2)', '(b)(2)'}         | {'T16'}                     | Definitional rules            | Data calculation algorithms      |

| <b>Elemento</b> | <b>Seção</b> | <b>Parágrafo</b>                | <b>transformation_template_ids</b> | <b>statement_classification_type</b> | <b>statement_classification_subtype</b> |
|-----------------|--------------|---------------------------------|------------------------------------|--------------------------------------|-----------------------------------------|
| Terms           | § 275.0-7    | {'(a)(3)', '(b)(1)'}            | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)', '(a)(3)'}            | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)(i)(A)', '(b)(1)(iii)'} | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)(ii)'}                  | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)(iii)'}                 | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)(iii)', '(b)(1)(i)(A)'} | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(1)(iv)'}                  | {'T7'}                             | Definitional rules                   | Formal intensional definitions          |
| Terms           | § 275.0-7    | {'(b)(2)', '(a)(2)'}            | {'T16'}                            | Definitional rules                   | Data calculation algorithms             |

Fonte: Elaborado pelo autor.

A transformação das declarações seguiu a taxonomia de Witt (2012) e o Quadro 42 relaciona cada elemento com sua classificação na taxonomia e o *template* utilizado.

Todos os dados apresentados neste apêndice estão disponíveis no repositório de código do autor.