IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication 04, 2023

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Pipeline de disponibilização dos Relatórios Obrigatórios no mercado de Capitais Brasileiro

SILVA, T.[1]
[1]Programa de Pós Graduação em Engenharia Eletrônica e de Computação, Instituto Tecnológico de Aeronáutica - ITA
Silva, T.(email: thiagotcns@ita.br)

ABSTRACT A complexidade envolvida na coleta e processamento de dados não estruturados como relatórios financeiros motivou a proposta de um pipeline de captura, processamento e disponibilização de dados, o que possibilita o uso de técnicas avançadas de aprendizado de máquina e mineração de dados. Este projeto tem como objetivo permitir a análise e tomada de decisões informadas no mercado financeiro, contribuindo para o avanço dos estudos de finanças quantitativas com a integração de dados não estruturados. O método proposto envolve a captura, pré-processamento dos relatórios financeiros disponibilizados pela Comissão de Valores Mobiliários (CVM) em um catálogo contendo metadados e informações relevantes.

INDEX TERMS CVM, finanças quantitativas, dados não estruturados

I. INTRODUÇÃO

A. CONTEXTUALIZAÇÃO

Com o aumento significativo de empresas listadas no mercado acionário e a modernização dos sistemas de informação, houve um aumento também no volume de informações disponíveis para análises de empresas. A maioria destes dados não estruturados em formato de textos e muitas vezes têm informações valiosas, estudos anteriores indicam que os relatórios financeiros além de refletir informações do passado, também contém informações sobre o desempenho futuro (Zhang et al 2018).

A utilização de dados textuais para melhorar a modelagem da dinâmica do mercado financeiro tem sido a tradição da prática comercial.Neste trabalho proponho uma estrutura de extração de dados não estruturados como relatórios financeiros, press releases, atas de assembleias entre outros apresentada pelas à Comissão de Valores Mobiliários (CVM) pelas empresas listadas na B3.

B. MOTIVAÇÃO TÉCNICA

A disponibilização de relatórios financeiros é essencial tanto para empresas quanto para investidores. No entanto, a tarefa de coletar, processar e disponibilizar esses relatórios pode ser complexa e exigente. Além disso, lidar com formatos de dados heterogêneos e informações incompletas ou inconsistentes dificulta ainda mais a análise. Nesse contexto, é fundamental estabelecer uma estrutura robusta e automatizada para a captura, processamento e disponibilização precisa, ágil e confiável desses dados.

Neste projeto, apresentamos uma estrutura abrangente para a disponibilização de relatórios financeiros fornecidos pelas empresas à Comissão de Valores Mobiliários (CVM), com o objetivo de facilitar a análise e o acesso a informações financeiras relevantes. A estrutura proposta é composta por várias etapas interconectadas, que abrangem desde a coleta inicial dos dados financeiros (por meio de um pipeline de captura) até a sua disponibilização final para os usuários, utilizando a indexação de metadados por meio do Apache Solr.

A primeira etapa consiste na coleta dos dados brutos fornecidos pela CVM. Em seguida, os dados passam por um processo de processamento e transformação, para serem convertidos em um formato adequado para análise posterior. Além disso, são organizados e agrupados por instituição financeira e ano de publicação. Após o pipeline de captura, a próxima etapa inclui a geração de tokens e tensores. Nessa fase, podem ser empregados algoritmos avançados de aprendizado de máquina e mineração de dados para descobrir padrões e tendências relevantes.

A última etapa da estrutura consiste na disponibilização dos relatórios financeiros processados e analisados para os usuários finais. Isso é realizado por meio da indexação de metadados em uma estrutura robusta e ágil, permitindo recursos avançados de pesquisa. Esses metadados incluem infor-

mações como o nome da empresa, descrição do relatório, data de referência, ano de publicação e diretório de disponibilidade do relatório bruto e pré-processado, entre outros. É de suma importância garantir que os relatórios sejam apresentados de forma clara, intuitiva e facilmente compreensível, para auxiliar os usuários na tomada de decisões informadas e estratégicas.

C. OBJETIVO

O objetivo principal deste projeto consiste em propor uma estrutura eficiente para a disponibilização de dados financeiros não estruturados, visando facilitar sua análise e apoiar a tomada de decisões. Como desdobramento deste objetivo geral, foram estabelecidos os seguintes objetivos específicos:

i. Desenvolver um pipeline para a captura e disponibilização de dados não estruturados, estabelecendo um fluxo eficiente e automatizado que permita a coleta e a organização dessas informações.

ii. Disponibilizar os metadados dos documentos em uma estrutura robusta e confiável, possibilitando que os usuários realizem pesquisas e obtenham uma compreensão detalhada dos documentos disponíveis.

iii. Contribuir para o avanço dos estudos em finanças quantitativas, por meio da integração de dados não estruturados. Isso permitirá explorar dados não convencionais, que poderão aprimorar a precisão dos modelos de previsão e proporcionar o desenvolvimento de estratégias de investimentos mais sofisticadas.

iv. Contribuir para a pesquisa de técnicas de processamento natural aplicadas à análise de relatórios financeiros, buscando identificar abordagens eficazes para extrair insights valiosos desses documentos, com o intuito de auxiliar na interpretação e compreensão das informações contidas neles.

Ao alcançar esses objetivos, espera-se proporcionar uma base sólida para a utilização de dados não estruturados no contexto financeiro, ampliando as possibilidades de análise e contribuindo para a tomada de decisões informadas e estratégicas.

II. REVISÃO BIBLIOGRÁFICA

A disponibilização e análise de dados não estruturados tem se tornado cada vez mais relevantes para as organizações, especialmente no contexto dos relatórios financeiros. Com o avanço da tecnologia no setor também torna se possível de obter e processar esses dados de forma automatizada, estudos indicam que 71 das instituições financeiras usam análise de big data para gerar vantagem competitiva para as suas organizações (Pejić et al. 2019). Pesquisas recentes mostram que apenas explorar dados históricos se tornou mais difícil, de acordo com Zhang et al. (2017), usando o expoente de Hurst, a correlação entre os retornos diários do Dow Jones e seus dados históricos recuou a partir da década de 1990. Neste contexto a análise de dados não estruturados como relatórios financeiros têm se tornado uma vantagem competitiva.

A. RELATÓRIOS FINANCEIROS

Os relatórios financeiros são os documentos mais representativo para analisar as comunicações das empresas sobre seu desempenho, saúde financeira e perspectivas futuras, são obrigatórios e produzidos anualmente contém informações para compartilhar com o mercado, às vezes, eles escondem pistas sobre desempenho futuro e informações valiosas as quais as empresas tentam esconder dos investidores (Masson et al. 2020). Sinais da mudança da posição financeira das empresas podem aparecer nos relatórios antes que possamos identificar nos números financeiros.

No Brasil a Comissão de Valores Mobiliários (CVM) tem um papel fundamental na regulação e supervisão desses relatórios, garantindo a transparência e a confiabilidade das informações divulgadas. Faz parte das suas atribuições estabelecer normas e diretrizes para a divulgação dos relatórios financeiros pelas das empresas (CVM, 2023). Estudos realizados com as empresas listadas na B3 indicam que os relatórios relacionados às Assembleias, Valores Mobiliários negociados e retidos, e Comunicados ao Mercado são os mais divulgados em termos de quantidade. A periodicidade de divulgação varia conforme o segmento da empresa na B3, empresas classificadas como Nível 1 de governança corporativa tem uma divulgação mais frequente devido a exigências regulatórias (Checon et al. 2021).

B. PREPROCESSAMENTO DE TEXTO

As etapas de pré-processamento por meio de vários métodos têm um impacto substancial nas técnicas de mineração de texto (Pejić et. al 2019). Envolve técnicas e procedimentos aplicados para preparar os dados textuais brutos para análise e extração de informações significativas. Em geral os pipelines de pré-processamento de texto consistem nas seguintes etapas :

1° Importação do texto bruto e formatação adequada.

2° Conversão do texto para letras minúsculas.

3° Geração de tokens, que divide o texto em partes menores, chamadas de tokens.

4° Remoção de stopwords (palavras comuns que geralmente não agregam informações importantes).

5° Marcação de partes de fala (POS - Part-of-Speech) do texto nos tokens.

6° Aplicação de técnicas de stemming ou lematização para reduzir as palavras à sua forma base.

É importante ressaltar que essas etapas podem variar dependendo da aplicação específica. Existem três abordagens principais para um pipeline de pré-processamento de texto (Ferrario et al., 2020):

1° Abordagem clássica: utiliza técnicas tradicionais de processamento de linguagem natural, como "bag-of-words" e "bag-of-part-of-speech".

2° Abordagem moderna: emprega técnicas de "word embeddings", como o Word2Vec, GloVe, ELMo e FastText, que capturam as representações distribuídas das palavras.

3° Abordagem contemporânea: introduz o uso de redes neurais recorrentes (RNNs) e redes convolucionais (CNNs)

e modelos baseados em Transformers, como o Longformer. A abordagem contemporânea busca automatizar a descoberta das representações necessárias para a tarefa de classificação, reduzindo a necessidade de etapas de pré-processamento manual (Ferrario et al., 2020). No entanto, as tarefas de geração de tokens e vetores continuam sendo essenciais para alimentar os modelos de processamento de texto.

A geração de tokens é uma etapa crítica do pré-processamento de texto, que envolve a segmentação do texto em unidades semânticas menores. Os tokens podem representar palavras, frases, sentenças ou até mesmo caracteres individuais, dependendo do nível de granularidade desejado (Borggreve, 2022). Essa etapa é crucial para preservar a estrutura e o contexto do texto original durante a análise e modelagem subsequentes. No Longformer, a geração de tensores e vetores ocorre durante o processo de pré-processamento da seguinte forma (Beltagy, I et al 2020):

Geração de tokens: Durante a geração de tokens, o texto é dividido em unidades menores, como palavras ou subpalavras, dependendo do tokenizador utilizado. O Longformer utiliza uma abordagem baseada em WordPiece para quebrar palavras em subpalavras com base em um vocabulário pré-definido. Essa técnica permite capturar a flexibilidade e a generalização das palavras, especialmente quando se trata de palavras compostas, neologismos ou idiomas com estruturas morfológicas complexas.

Tokenização de documentos longos: Uma característica crucial do Longformer é sua capacidade de lidar com documentos longos, que podem exceder o limite fixo de tokens usado em modelos de Transformers convencionais. Para isso, o Longformer emprega uma estratégia chamada "sliding window". Nessa abordagem, o texto é dividido em segmentos menores e sobrepostos, chamados de janelas, que são alimentados no modelo de forma sequencial. Dessa forma, o Longformer pode processar sequências longas sem perder a conexão contextual entre as partes do texto.

Geração de vetores densos: Após a geração de tokens, os textos são transformados em vetores densos em um espaço de alta dimensionalidade. Cada token é atribuído a um índice único e representado por um vetor binário com todas as dimensões zeradas, exceto a dimensão correspondente ao seu índice, que é definida como 1. Essa representação binária permite que o modelo identifique e diferencie cada token durante o processamento.

C. CATÁLOGO DE DADOS E INDEXAÇÃO DE METADADOS

A indexação de metadados é uma etapa fundamental para a disponibilização e organização, especialmente em ambientes com grandes volumes de informações para que os dados sejam reutilizados (Nakandala et al. 2015). Metadados são informações descritivas que fornecem contexto e estrutura aos dados, facilitando sua organização e recuperação. Eles descrevem características como o tipo de dado, formato, autor, data de criação, entre outros atributos relevantes.

O Apache Solr é uma ferramenta amplamente utilizada para indexação de metadados e busca de informações. É uma plataforma de busca e indexação de texto de código aberto baseada na biblioteca Lucene, que oferece recursos poderosos para o gerenciamento eficiente de grandes volumes de dados (Apache 2021).

Ao utilizar o Apache Solr para indexar metadados, os dados são estruturados e organizados em índices que permitem uma recuperação eficiente e rápida das informações. Após a indexação dos metadados, é possível realizar consultas avançadas no Apache Solr para recuperar os documentos que correspondem a determinados critérios de busca.

O Solr oferece recursos poderosos de pesquisa, incluindo pesquisa textual avançada, filtragem de resultados, classificação por relevância e paginamento, entre outros recursos. Por exemplo, é possível filtrar os resultados por autor, data de criação ou qualquer outro campo relevante presente nos metadados indexados.

A utilização do Apache Solr na indexação de metadados traz vantagens significativas para a recuperação de informações em grandes conjuntos de dados. Sua escalabilidade, eficiência e recursos avançados de busca tornam-no uma escolha popular para a implementação de sistemas de busca e recuperação de dados.

Diversos estudos e trabalhos científicos têm explorado o uso do Apache Solr para indexação de metadados em diferentes contextos e domínios de aplicação. Essas pesquisas têm demonstrado a eficácia e a utilidade do Solr como uma solução robusta e flexível para a indexação e recuperação eficiente de metadados (K. Guntupally et al. 2020).

III. MÉTODO PROPOSTO

O regulador responsável pelo mercado de capitais brasileiro é a Comissão de Valores Mobiliários (CVM). As empresas nela cadastradas são obrigadas a divulgarem regularmente relatórios financeiros e não financeiros relativos ao seu exercício. Esses relatórios estão arquivados na CVM (Checon et al. 2021). A disponibilização destes relatórios pela CMV pode estar em caminhos (APIs) e formatos distintos. Aqui é proposta uma estrutura para captura, pré-processamento dos relatórios de forma eficiente e automatizado. Na figura 1, apresenta-se o diagrama de fluxo do método proposto para fazer a captura dos relatórios financeiros e disponibiliza-los para o consumo em seu formato bruto (pdf). Nesta etapa é possível fazer análises, ainda de forma artesanal, para auxiliar na tomada de decisões financeiras.

Na etapa seguinte, exibida na Figura 2, é proposta a conversão dos relatórios em formato PDF em um arquivo texto. Durante este processo, espera-se que uma pequena quantidade de dados pode ser descartada devido a formatação dos relatórios, como por exemplo, arquivos que contenham imagens, dados corrompidos, tipos ou fontes de textos muito difíceis de reconhecer em PDF. Na próxima etapa será criada os tokens do texto, ou seja, cada palavra do arquivo é armazenada como um valor em um vetor, posteriormente os tokens são transformados em tensores. A disponibilização

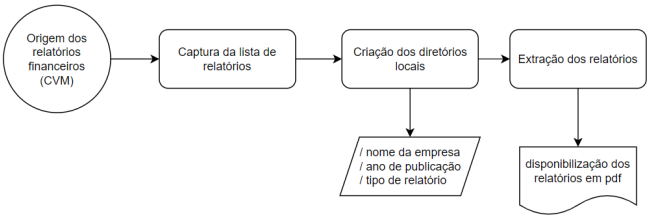

FIGURE 1. Etapa 1: Pipeline de captura dos relatórios financeiros.

dos dados finais possibilitará a partir de processamento de linguagem natural fazer análises de desempenho, análise de risco entre outros. (Masson et al. 2020).

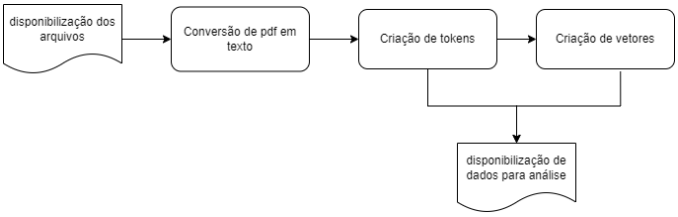

FIGURE 2. Etapa 2: pré-processamento: geração de tokens e vetores.

Com a disponibilização dos dados brutos (relatórios financeiros disponibilizados em pdf) e os dados especializados (dados para análise) será necessário organizar as informações. Nesta última etapa, exibida na figura 3, é proposto a criação de um catálogo de dados onde será armazenado informações como metadados, data de publicação do relatório, nome e tipo do relatório, formato do arquivo, tamanho do arquivo, categoria, entre outros.

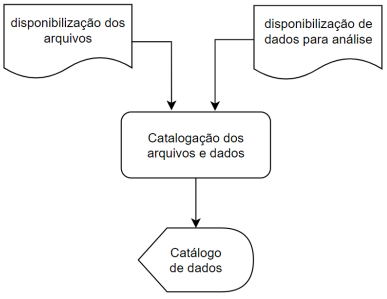

FIGURE 3. Etapa 3: Disponibilização dos metadados via Apache Solr.

A arquitetura proposta consiste em um componente central que orquestra as chamadas e supervisiona as etapas de captura e pré-processamento dos dados. Além disso, o componente central é responsável pela extração dos metadados, que são posteriormente indexados na etapa final do processo. A Figura 4 ilustra a arquitetura proposta, destacando o componente central e os três módulos envolvidos: captura de dados, pré-processamento e indexação de metadados.

Ao final deste trabalho, espero demonstrar a eficácia desta estrutura na disponibilização eficiente de relatórios financeiros, possibilitando o aumento da agilidade na análise de dados e redução nos erros humanos. Além disso, pretendo contribuir com o avanço nos estudos da finança quantitativa ao possibilitar uma abordagem integrada com dados não estruturados e que possa contribuir para minha pesquisa.

IV. RESULTADOS E DISCUSSÕES

O objetivo do projeto proposto foi desenvolver um componente para disponibilizar e indexar textos brutos e pré-processados provenientes dos relatórios financeiros disponibilizados pela Comissão de Valores Mobiliários (CVM), a fim de auxiliar na tomada de decisões financeiras. Obtivemos resultados significativos nos três principais módulos do componente: Captura de Dados da CVM, Pré-processamento (geração de tokens e vetores) e Indexação de Metadados no Apache Solr e também no teste integrado.

A. CAPTURA DE DADOS DA CVM

O módulo de captura de dados da CVM demonstrou-se eficiente e confiável na extração dos documentos. Nos testes realizados, conseguimos capturar diversos relatórios econômico-financeiros, como relatórios anuais, relatórios de análise gerencial e demonstrações financeiras anuais completas, publicados entre 2018 e 2022. O módulo também se mostrou eficiente na organização dos documentos em um sistema de arquivos local. Ao ser executado pela primeira vez, o módulo criou os diretórios locais, seguindo a estrutura: "caminho fornecido pelo usuário como parâmetro"/"nome da empresa"/"ano de publicação do relatório"/"tipo do relatório"/"código CVM"-"nome da empresa"-"assunto do relatório"-"data de publicação".pdf. Durante esse processo, disponibilizamos 1790 relatórios de cinco empresas do ramo financeiro listadas na bolsa de valores, conforme apresentado na Tabela 1.

B. PRÉ-PROCESSAMENTO: GERAÇÃO DE TOKENS E VETORES

Após a captura dos documentos da CVM, realizamos o pré-processamento dos textos para torná-los adequados para análises posteriores. A primeira etapa foi a conversão dos arquivos PDF em um dataframe no formato de texto, que serviu como entrada para o processo de geração de tokens.

Utilizamos a biblioteca LongformerTokenizer para realizar a geração de tokens dos documentos. Essa abordagem consiste em dividir o texto em unidades menores utilizando a técnica de Byte-Pair Encoding, que codifica pares de bytes. A biblioteca mostrou-se eficaz, permitindo a geração de um grande número de tokens. Por exemplo, o relatório anual integrado de 2022 resultou em 365.307 tokens, ultrapassando o limite de entrada de 4.096 tokens do modelo Longformer.

Para a geração de vetores, utilizamos a biblioteca otimizada PyTorch. Devido às limitações do modelo, tivemos que dividir o texto em partes de tamanho máximo definido como 4.096 tokens. Cada parte foi processada para gerar seus respectivos vetores usando um modelo pré-treinado Longformer. Para cada vetor final, calculamos sua média e o concatenamos em uma lista de vetores, resultando em

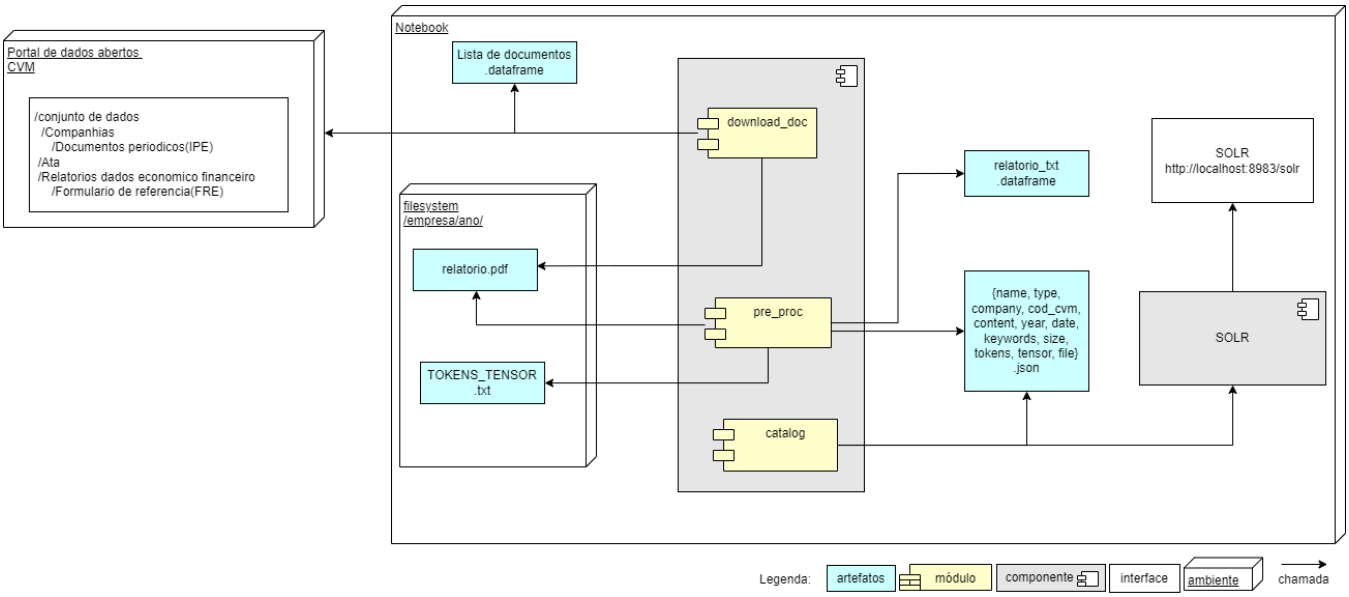

FIGURE 4. Arquitetura proposta: componente central e três módulos (captura de relatórios, pré-processamento e indexação de metadados

um único tensor ao longo de uma dimensão (dimensão 1). O relatório integrado de 2022 resultou em um tensor de tamanho 92.160 (1, 92.160).

Devido ao tempo de processamento, não foi possível executar o mesmo procedimento para todos os 1.760 arquivos coletados na etapa de teste de captura de dados. Para gerar o tensor de tamanho (1, 92.160), foram necessários 55 minutos em um Notebook com 16 GB de memória RAM e um processador i7-8750H CPU @2.20GHz. A geração de vetores foi essencial para capturar as relações semânticas entre as palavras e possibilitar análises posteriores.

Para exemplificar o uso, realizamos um teste de similaridade entre os relatórios anuais integrados dos bancos Pan, Santander e Itaú, utilizando uma medida de similaridade coseno com os vetores gerados. Obtivemos um grau de similaridade de 0,81 entre os relatórios do Itaú e Pan, 0,25 entre os relatórios do Itaú e Santander, e 0,30 entre os relatórios do Pan e Santander.

Em outro teste, utilizamos os tokens para um modelo de similaridade e conseguimos obter a expressão com o maior nível de similaridade ao fazer uma pergunta. Ao questionar o modelo "quem é o presidente?" de acordo com o relatório anual gerencial de 2022, obtivemos a seguinte resposta: "pel perante a sociedade . em um processo absolutamente planejado e conduzido dentro da governança estabelecida , em 2021 realizamos a transição da liderança do banco para o milton maluhy". Embora o modelo não tenha fornecido diretamente o nome do presidente Milton Maluhy como resposta, obtivemos uma expressão com alto grau de similaridade com a pergunta.

C. INDEXAÇÃO DE METADADOS NO APACHE SOLR

Após o pré-processamento dos documentos e a geração dos vetores, realizamos a indexação dos metadados no Apache Solr, um sistema de busca e indexação de alto desempenho. O Solr nos permitiu armazenar e pesquisar eficientemente os documentos pré-processados, facilitando a recuperação de informações relevantes para nossa análise.

Utilizamos a estrutura de indexação fornecida pelo Solr para os metadados obtidos nas etapas de captura de dados e pré-processamento. Atribuímos a cada documento um identificador único e indexamos os metadados, como nome do relatório (name), tipo do relatório (type), nome da empresa (company-name), código CVM (cod-cvm), assunto do relatório (content), ano de publicação (year), data de referência (date), palavras-chave (keywords), que podem variar de acordo com a disponibilidade da informação, tamanho do relatório em bytes (size), e os diretórios onde estão armazenados os tokens (tokens), vetores (tensor) e os relatórios (file). Essa indexação nos permitiu realizar consultas rápidas e precisas nos documentos de interesse.

D. COMPONENTE DE CAPTURA, PRÉ-PROCESSAMENTO E INDEXAÇÃO (CPI)

Os módulos mencionados são coordenados pelo componente CPI. Para garantir sua efetividade, os testes foram aplicados separadamente em cada módulo. Ao executar o pipeline do início ao fim, o tempo total gasto foi de 105 minutos para processar os 12 relatórios do primeiro semestre de 2023 da Magazine Luiza. Esses relatórios abrangiam demonstrações financeiras anuais completas, demonstrações financeiras em padrões internacionais, demonstrações financeiras intermediárias, press-release e relatório de agente fiduciário.

Em suma, os resultados obtidos evidenciam a efetividade do componente de captura de dados da CVM, o sucesso do pré-processamento dos documentos para a geração de tokens e vetores, e a eficiência da indexação no Solr para a busca e recuperação de informações. Esses avanços são essenciais

para impulsionar a pesquisa em análise financeira e a tomada de decisões embasada em dados.

V. CONCLUSÃO

Neste tarbalho, apresentamos uma proposta de pipeline para a captura, processamento e disponibilização de relatórios financeiros no mercado de capitais brasileiro. Por meio da integração de técnicas avançadas de aprendizado de máquina e mineração de dados, buscamos melhorar a análise e a tomada de decisões informadas no mercado financeiro.

Ao longo deste trabalho, realizamos a coleta de dados não estruturados fornecidos pela Comissão de Valores Mobiliários (CVM) e desenvolvemos um processo de pré-processamento e transformação para adequar os relatórios financeiros a um formato adequado para análise.

Além disso, organizamos e agrupamos os relatórios por instituição financeira e ano de publicação, além de indexar os metadados, catalogando os documentos e facilitando o acesso e a compreensão dos dados.

No entanto, este trabalho não está isento de desafios e limitações. A heterogeneidade dos formatos de dados e a presença de informações incompletas ou inconsistentes foram obstáculos enfrentados durante o processo de captura e processamento dos relatórios. Portanto, futuros trabalhos podem se concentrar em melhorar a robustez e a precisão do pipeline, a fim de lidar de forma mais eficiente com essas questões.

Além disso, uma área promissora para futuros estudos é a aplicação de técnicas de processamento de linguagem natural (NLP) e aprendizado de máquina (modelo baseados em transformers), para a análise semântica dos relatórios financeiros. Isso poderia permitir uma compreensão mais profunda do conteúdo dos relatórios, identificando sentimentos, tendências e insights que vão além da análise puramente quantitativa. Outro aspecto a ser explorado é a integração de dados externos, como informações de mercado e notícias financeiras, para enriquecer ainda mais a análise dos relatórios financeiros. Isso poderia fornecer um contexto mais amplo para as decisões de investimento e contribuir para uma visão mais abrangente do mercado de capitais.

Em resumo, A proposta apresentada neste projeto demonstra a importância da disponibilização eficiente de dados financeiros não estruturados por meio de um pipeline de captura, processamento e disponibilização de dados. Através dessa estrutura, é possível superar a complexidade envolvida na coleta e pré-processamento de relatórios financeiros, permitindo o uso de técnicas avançadas de aprendizado de máquina e mineração de dados. Essas áreas oferecem oportunidades para o avanço dos estudos em finanças quantitativas e aprimoramento das tomadas de decisão no mercado financeiro.

REFERENCES

[1] Apache Software Foundation, "Solr Reference Guide - Getting Started", Solr Apache, 2021. [Online]. Disponível: https://solr.apache.org/guide/solr/latest/getting-started/introduction.html. [Acessado em: 2 jul. 2023].

[2] Beltagy, Iz, Peters, Matthew E., Cohan, Arman. (2020). Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150. Disponível em: https://arxiv.org/abs/2004.05150

[3] Borggreve, L.A. (2022) "Effects of Annual Report Sentiment on Stock Returns." UNIVERSITY OF TWENTE STUDENT THESES

[4] Checon, Bianca Q., Santana, Verônica de F., "Relatórios Obrigatórios e Padrões de Divulgação no Mercado de Capitais." Artigo Cíclico n°8. FGV EASP Instituto de finanças

[5] CVM (2023). "O papel da CVM". Disponível em: https://www.gov.br/investidor/pt-br/investir/cuidados-ao-investir/o-papel-da-cvm: :text=A

[6] Ferrario, Andrea and Naegelin, Mara, The Art of Natural Language Processing: Classical, Modern and Contemporary Approaches to Text Document Classification (March 1, 2020). Available at SSRN: https://ssrn.com/abstract=3547887

[7] K. Guntupally, K. Dumas, W. Darnell, M. Crow, R. Devarakonda and P. Giri, "Automated Indexing of Structured Scientific Metadata Using Apache Solr," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 5685-5687, doi: 10.1109/BigData50022.2020.9378448.

[8] Masson, Corentin and Patrick Paroubek. "NLP Analytics in Finance with DoRe: A French 250M Tokens Corpus of Corporate Annual Reports." International Conference on Language Resources and Evaluation (2020).

[9] Pejić Bach M., Ž. Krstić, S. Seljan, and L. Turulja, "Text Mining for Big Data Analysis in Financial Sector: A Literature Review", Sustainability, vol. 11, no. 5, p. 1277, Feb. 2019, doi: 10.3390/su11051277.

[10] QUINTANA PELAYO, Guillermo; STUDENT, AI Master. "NLP approach to Annual Reports Analysis". 2020.

[11] Xing, Frank Z., et al. "Natural Language Based Financial Forecasting: A Survey." Artificial Intelligence Review, vol. 50, no. 1, June 2018, pp. 49–73. Version: Author's final manuscript

[12] Zhang, X., Tan, Y. (2018). "Deep Stock Ranker: A LSTM Neural Network Model for Stock Selection." In: Tan, Y., Shi, Y., Tang, Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science(), vol 10943. Springer, Cham.