**Validação Semântica de Normas em Projetos AEC Utilizando Modelos de Linguagem de Grande Escala e Metodologia RASE**

1. **Resumo**

A validação de normas em projetos de Arquitetura, Engenharia e Construção (AEC) apresenta desafios devido à complexidade das regulamentações regionais. Com a adoção do Building Information Modeling (BIM), ferramentas como o Revit permitem armazenar dados detalhados dos projetos, mas não garantem automaticamente a conformidade normativa. A metodologia RASE (Requirements, Applicability, Selection, Exception) possibilita a conversão dessas normas em uma linguagem computável, permitindo automação na validação. Este artigo propõe o uso de Modelos de Linguagem de Grande Escala (LLMs), como o LLaMA 3.0, para a extração automática de normas e sua conversão para o formato RASE. Diversas técnicas de comparação semântica são analisadas, incluindo Distância de Levenshtein, Similaridade do Cosseno, SBERT, BERTimbau, MPNET e WMD, com o objetivo de identificar a abordagem mais eficiente para validação de dados extraídos.

1. **Introdução**

A validação de normas em projetos de Arquitetura, Engenharia e Construção (AEC) enfrenta desafios significativos devido às especificidades normativas de cada região, tornando a verificação de conformidade um processo complexo e suscetível a erros. A interpretação manual dessas normas, além de ser trabalhosa, está sujeita a variações de entendimento, o que pode comprometer a precisão na aplicação dos regulamentos.

Com a adoção do Building Information Modeling (BIM), ferramentas como o Revit passaram a armazenar vastas quantidades de informações detalhadas sobre os projetos. Entretanto, assegurar que todas as normas regionais estejam corretamente implementadas nesses modelos tornou-se um desafio considerável, exigindo soluções automatizadas para uma verificação e validação eficientes.

A metodologia RASE (Requirements, Applicability, Selection, Exception) oferece uma abordagem para transformar normas de projetos AEC em uma linguagem computável, permitindo que programas automatizados realizem a validação dessas regras. Contudo, por se basear em textos interpretativos, a criação dessa linguagem é geralmente manual, o que pode levar a variações semânticas entre usuários, mesmo quando o significado pretendido é o mesmo.

A validação de grandes projetos utilizando BIM para diversas regras de engenharia é uma tarefa árdua e minuciosa, exigindo a análise detalhada de cada elemento. Nesse contexto, Modelos de Linguagem de Grande Escala (LLMs), como o LLaMA 3.0, podem ser explorados para a leitura e interpretação de regras em projetos AEC, convertendo-as para o formato RASE de maneira mais eficiente.

A aplicação de LLMs possibilita a automação eficiente na extração de informações a partir de documentos complexos, retornando os dados processados em formatos estruturados, como JSON. No entanto, é crucial validar se os dados extraídos pelos modelos são semanticamente equivalentes às informações originais. Para essa validação, existem diversas técnicas de comparação semântica, incluindo a Distância de Levenshtein, Similaridade do Cosseno, SBERT, BERTimbau, MPNET e WMD.

Este artigo propõe identificar a técnica mais eficaz entre as mencionadas para a validação semântica das métricas retornadas por LLMs no formato JSON, avaliando a confiabilidade dos resultados obtidos. Serão analisadas as similaridades semânticas entre os textos, determinando quais campos podem utilizar técnicas mais leves com resultados satisfatórios e quais requerem abordagens mais robustas.

1. **Referencial Teórico**
   1. **Building Information Modeling (BIM)**

O Building Information Modeling (BIM) é um processo baseado em modelos inteligentes que proporciona uma visão digital integrada de projetos AEC. Ele permite a modelagem precisa de edificações, sistemas estruturais e instalações, gerenciando grandes volumes de dados em um ambiente colaborativo. O uso do BIM facilita a análise de conformidade, mas também impõe desafios devido à complexidade das informações contidas no modelo.

O BIM possui sete dimensões (3D, 4D, 5D, 6D, 7D), abrangendo desde a modelagem tridimensional até estimativas de custo, cronogramas de construção, sustentabilidade e gestão do ciclo de vida dos empreendimentos.

* 1. **Metodologia RASE**

A metodologia RASE é empregada para converter normas e regulamentos em regras computáveis, permitindo a verificação automatizada da conformidade dos projetos. A metodologia estrutura normas em quatro componentes:

* **Requisito**: O que deve ser atendido.
* **Aplicabilidade**: Contexto onde a norma se aplica.
* **Seleção**: Critérios ou parâmetros necessários para a aplicação.
* **Exceção**: Situações onde a norma não se aplica.

Os textos extraídos das normas são convertidos em JSON, permitindo sua avaliação automática.

* 1. **Modelos de Linguagem de Grande Escala (LLM)**

Os Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) são redes neurais profundas treinadas em grandes volumes de dados textuais, capazes de processar, compreender e gerar texto de forma autônoma. Esses modelos são baseados em arquiteturas como o Transformer, introduzido pelo artigo "Attention Is All You Need" (Vaswani et al., 2017), que revolucionou o campo do Processamento de Linguagem Natural (NLP) ao permitir a modelagem eficiente de contextos extensos em textos.

Os LLMs utilizam aprendizado de máquina, especificamente aprendizado profundo, para mapear padrões complexos no uso da linguagem, capturando relações semânticas e contextuais entre palavras e frases. Durante o pré-treinamento, esses modelos aprendem a prever palavras ausentes em frases (tarefa de modelagem de linguagem) ou a responder perguntas baseadas em conhecimento extraído de dados não estruturados.

Atualmente, os LLMs são amplamente utilizados para diversas aplicações, incluindo: Geração de textos, tradução automática, assistentes virtuais, análise semântica e classificação de textos e extração de informações estruturadas.

Apesar das vantagens, os LLMs apresentam desafios como vieses presentes nos dados de treinamento, custo computacional elevado e a necessidade de técnicas avançadas para avaliação da precisão e consistência dos resultados.

* 1. **Modelo LLaMA 3.0**

O LLaMA 3.0 (Large Language Model Meta AI) é uma das versões mais recentes dos modelos desenvolvidos pela Meta AI, projetado para fornecer uma alternativa otimizada aos modelos de linguagem existentes, como GPT e PaLM. O LLaMA 3.0 aprimora sua capacidade de geração de texto, compreensão semântica e eficiência computacional em comparação com seus predecessores, tornando-se uma opção viável para aplicações que exigem alta precisão na interpretação de linguagem natural.

Principais características do LLaMA 3.0:

* **Eficiência Computacional:** Comparado a outros modelos de grande escala, o LLaMA 3.0 foi otimizado para consumir menos recursos computacionais sem comprometer a qualidade das respostas.
* **Melhoria no Aprendizado Contextual:** A arquitetura foi aprimorada para melhor capturar dependências de longo alcance entre palavras e frases, tornando-o altamente eficaz em tarefas que envolvem análise de texto técnico, como normas e regulamentos.
* **Suporte a Multilínguas:** Embora tenha sido treinado predominantemente em inglês, o modelo apresenta boas capacidades de processamento em múltiplos idiomas, incluindo português.
* **Geração Estruturada de Dados:** O LLaMA 3.0 tem sido explorado para converter informações extraídas de textos em formatos estruturados, como JSON ou XML, possibilitando sua aplicação em tarefas de automação e validação semântica.
* **Aprimoramentos no Filtro de Viés:** Em comparação com versões anteriores, foram implementadas técnicas mais avançadas para mitigar vieses presentes nos dados de treinamento, reduzindo respostas tendenciosas.

No contexto da metodologia RASE, o LLaMA 3.0 é utilizado para extrair e estruturar normas e regulamentos, convertendo textos em representações computáveis. Sua capacidade de entender e segmentar informações semânticas complexas permite que ele automatize o processo de transformação de normas em regras específicas, melhorando a eficiência da verificação de conformidade em projetos AEC.

* 1. **Métodos de Validação Semântica**

A avaliação das respostas dos LLMs para RASE pode ser realizada utilizando diferentes métodos de comparação semântica:

* **Distância de Levenshtein**: Mede a similaridade entre cadeias de caracteres com base no número mínimo de operações necessárias para transformar uma string em outra. A biblioteca FuzzyWuzzy em Python é amplamente utilizada para esse tipo de análise, facilitando a identificação de erros de digitação e correspondência aproximada de strings.​
* **Similaridade do Cosseno e TF-IDF**: A Similaridade do Cosseno é utilizada para medir a proximidade entre vetores em um espaço multidimensional. Quando combinada com a técnica TF-IDF (Term Frequency - Inverse Document Frequency), a análise de similaridade torna-se mais robusta, atribuindo pesos diferenciados às palavras e evitando que termos comuns tenham alta relevância.​
* **SBERT (Sentence-BERT)**: Modelo baseado no BERT, otimizado para calcular a similaridade entre sentenças. Ele gera embeddings fixos para sentenças, tornando a comparação de textos mais eficiente e escalável.​
* **BERTimbau**: Versão do BERT treinada especificamente para o português, com variantes especializadas, como o BERTimbau STS, voltada para análise de similaridade entre sentenças.​
* **MPNET (Masked and Permuted Pre-training for Language Understanding)**: Modelo que combina elementos do BERT e XLNet, aprimorando a captura de dependências entre palavras e a eficiência no pré-treinamento.​
* **WMD (Word Mover's Distance)**: Métrica que mede a distância semântica entre textos utilizando embeddings de palavras. Calcula o "caminho mínimo" necessário para transformar um texto em outro, considerando o significado das palavras.

1. **Métodos**

O processo metodológico deste estudo envolveu três bases de dados principais, denominadas N1, N2 e N3, cada uma representando uma etapa fundamental na transformação das normas de projetos AEC para o formato RASE.

* 1. **Construção das Bases de Dados**

A primeira base de dados, N1, foi composta por um conjunto de normas regulamentadoras aplicáveis a projetos AEC. Essas normas foram analisadas e fragmentadas em trechos que contivessem apenas uma única regra computável. Essa conversão foi realizada com o objetivo de facilitar a identificação e extração dos componentes lógicos presentes em cada norma. Abaixo, segue um exemplo de conversão:

**Norma original:**

“As áreas de qualquer espaço ou edificação de uso público ou coletivo devem ser servidas de uma ou mais rotas acessíveis. As edificações residenciais multifamiliares, condomínios e conjuntos habitacionais necessitam ser acessíveis em suas áreas de uso comum. As unidades autônomas acessíveis devem estar conectadas às rotas acessíveis. Áreas de uso restrito, conforme definido em 3.1.38, como casas de máquinas, barriletes, passagem de uso técnico e outros com funções similares, não necessitam atender às condições de acessibilidade desta Norma.”

**Fragmentação em regras individuais:**

• As áreas de qualquer espaço ou edificação de uso público ou coletivo devem ser servidas de uma ou mais rotas acessíveis.  
• As edificações residenciais multifamiliares, condomínios e conjuntos habitacionais necessitam ser acessíveis em suas áreas de uso comum.  
• As unidades autônomas acessíveis devem estar conectadas às rotas acessíveis.  
• Áreas de uso restrito, conforme definido em 3.1.38, como casas de máquinas, barriletes, passagem de uso técnico e outros com funções similares, não necessitam atender às condições de acessibilidade desta Norma.

Após a extração e segmentação das normas na base N1, a próxima etapa foi a identificação dos operadores da metodologia RASE. Para isso, foi necessário classificar cada trecho da norma em seus respectivos operadores:

* 1. **Identificação dos Operadores RASE**

Para cada um dos trechos extraídos da base N1, foram atribuídos os seguintes operadores lógicos: Requisito, Aplicabilidade, Seleção e Exceção. Esta etapa foi realizada de forma manual e automática por meio de modelos de Processamento de Linguagem Natural (PLN). A seguir, um exemplo da extração desses operadores para um trecho específico:

• **Texto:** "As áreas de qualquer espaço ou edificação de uso público ou coletivo devem ser servidas de uma ou mais rotas acessíveis."

• **Requisito:** "devem ser servidas de uma ou mais rotas acessíveis."

• **Aplicabilidade:** "As áreas de qualquer espaço ou edificação."

• **Seleção:** "uso público ou coletivo."

• **Exceção:**

* 1. **Conversão para Estrutura JSON**

A terceira base de dados, N3, foi gerada a partir das informações estruturadas da base N2. Os textos e operadores extraídos foram convertidos em um formato padronizado JSON para facilitar a manipulação e comparação automática. O formato JSON gerado seguiu a estrutura:

{

"Texto": "devem ser servidas de uma ou mais rotas acessíveis.",

"Tipo": "requisito",

"Objeto": "espaço ou edificação",

"Propriedade": "uso",

"Comparação": "=",

"Alvo": "Verdadeiro"

}

Essa conversão permitiu que um sistema automático validasse os objetos e suas comparações, identificando se um projeto estava em conformidade com as normas estabelecidas.

* 1. **Validação dos Dados Extraídos**

Os dados extraídos pelo modelo LLaMA 3 foram comparados com os dados originais para avaliar a precisão da conversão semântica. Para isso, diferentes técnicas de similaridade foram empregadas:

* + - **Distância de Levenshtein**: Mede a similaridade entre cadeias de caracteres, quantificando o número mínimo de operações necessárias para transformar uma string em outra.
    - **Similaridade do Cosseno e TF-IDF**: Mede a proximidade vetorial entre textos, atribuindo pesos a termos e evitando que palavras comuns influenciem indevidamente a análise.
    - **SBERT (Sentence-BERT)**: Utiliza redes neurais profundas para calcular embeddings e medir a similaridade semântica entre sentenças.
    - **BERTimbau**: Versão do BERT treinada para o português, com capacidade aprimorada para análise semântica.
    - **MPNET**: Combina elementos do BERT e XLNet para capturar dependências contextuais complexas.
    - **WMD** (Word Mover’s Distance): Mede a distância entre palavras em embeddings pré-treinados, calculando a transformação mínima necessária entre textos.

Cada uma dessas técnicas foi ajustada para garantir que seus valores de similaridade fossem normalizados para a mesma escala de medição. Algumas técnicas retornam valores de 0 a 1, enquanto outras utilizam escalas de 1 a 100 ou classificações binárias. A padronização foi essencial para garantir uma comparação justa entre os métodos.

Além da similaridade, foi analisado o tempo de processamento de cada técnica. Isso permitiu avaliar quais abordagens eram mais eficientes tanto em qualidade de resposta quanto em desempenho computacional.

* 1. **Comparação e Análise dos Resultados**

Com os valores obtidos, foram identificadas correlações entre as técnicas de comparação semântica. Algumas abordagens apresentaram desempenho similar, enquanto outras se destacaram por fornecer análises mais precisas em determinados contextos. Os resultados foram analisados considerando:

* Precisão na identificação dos operadores RASE.
* Tempo de processamento para cada técnica.
* Diferenciação entre textos semanticamente semelhantes e distintos.

A partir dessa análise, foi possível identificar quais técnicas de validação semântica são mais eficazes na automação da conversão de normas AEC para o formato RASE, otimizando a implementação de verificações automáticas de conformidade em modelos BIM.

Essa metodologia permitiu uma avaliação aprofundada da viabilidade do uso de Modelos de Linguagem de Grande Escala na extração e interpretação de normas regulamentadoras, contribuindo para o avanço da automação na verificação de conformidade em projetos de Arquitetura, Engenharia e Construção.

1. **Resultados**

(Apresentar os resultados da comparação entre as diferentes abordagens e sua eficácia na validação dos dados extraídos pelos LLMs.)

1. **Discussão**

6.1 Considerações Finais (Analisar os desafios encontrados, limitações da metodologia e sugestões para trabalhos futuros.)

**7. Referências**

* FUZZY: [2201.13427] Fuzzy Segmentations of a String. [[2201.13427] Fuzzy Segmentations of a String](https://arxiv.org/abs/2201.13427?utm_source=chatgpt.com)
* TF-IDF: Análise de similaridade entre TF-IDF e modelos contextualizados de linguagem baseados em tokens. [Análise de similaridade entre TF-IDF e modelos contextualizados de linguagem baseados em tokens](https://repositorio.ufsc.br/bitstream/handle/123456789/243553/monografia.pdf?isAllowed=y&sequence=1&utm_source=chatgpt.com)
* SBERT: [1908.10084] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [[1908.10084] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](https://arxiv.org/abs/1908.10084?utm_source=chatgpt.com)
* BERTIMBAU: BERTimbau: Pretrained BERT Models for Brazilian Portuguese | Intelligent Systems.   
  [BERTimbau: Pretrained BERT Models for Brazilian Portuguese | Intelligent Systems](https://dl.acm.org/doi/10.1007/978-3-030-61377-8_28?utm_source=chatgpt.com)
* MPNET: [2004.09297] MPNet: Masked and Permuted Pre-training for Language Understanding. [[2004.09297] MPNet: Masked and Permuted Pre-training for Language Understanding](https://arxiv.org/abs/2004.09297?utm_source=chatgpt.com)
* WMD: Supervised Word Mover's Distance. [Supervised Word Mover's Distance](https://proceedings.neurips.cc/paper_files/paper/2016/file/10c66082c124f8afe3df4886f5e516e0-Paper.pdf)
* RASE: [Aplicação da metodologia RASE na norma de acessibilidade associada ao BIM](https://www.academia.edu/101456227/Aplica%C3%A7%C3%A3o_da_metodologia_RASE_na_norma_de_acessibilidade_associada_ao_BIM?utm_source=chatgpt.com)
* @inbook{hjelseth2015,
* title = {BIM-based model checking (BMC)},
* author = {Hjelseth, Eilif},
* year = 2015,
* month = {05},
* publisher = {Building Information Modeling: Applications and Practices},
* pages = {33--61},
* doi = {10.1061/9780784413982.ch02},
* isbn = {978-0-7844-1398-2}
* }