UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Matheus Vasconcelos Batalha

Detecção de contexto de citação baseada em aprendizado profundo

São Luís - MA

2023

Matheus Vasconcelos Batalha

Detecção de contexto de citação baseada em aprendizado profundo

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Curso de Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Antônio de Abreu Batista Jr

São Luís - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Batalha, Matheus Vasconcelos.
Detecção de contexto de citação baseada em aprendizado profundo / Matheus Vasconcelos Batalha. - 2024.
31 p.

Orientador(a): Antônio de Abreu Batista Jr.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2024.

1. ACL-ARC. 2. Detecção de Contexto de Citação. 3. Recomendação de Citação. 4. SciBERT. 5. SVM. I. Batista Jr, Antônio de Abreu. II. Título.

Matheus Vasconcelos Batalha

Detecção de contexto de citação baseada em aprendizado profundo

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho Aprovado, em 22 de Janeiro de 2024:

Prof. Dr. Antônio de Abreu Batista Jr
Orientador
Universidade Federal do Maranhão

Prof. Dr. Fabio Castro Gouveia
Examinador Externo
Fundação Osvaldo Cruz (FIOCRUZ)

Prof. Dr. Jesús Pascual Mena Chalco
Examinador Externo
Universidade Federal do ABC (UFABC)

Prof. Dr. Alexandre Cesar Muniz de Oliveira
Examinador Interno

São Luís - MA
2023

Agradecimentos

À medida que concluo esta etapa significativa da minha jornada acadêmica, é com profundo respeito e gratidão que reconheço aqueles que foram fundamentais neste processo. Em primeiro lugar, expresso minha sincera gratidão ao meu orientador, Prof. Antônio de Abreu Batista Jr., por sua orientação exemplar, paciência e dedicação. Sua capacidade de extrair o melhor de mim e seu compromisso com a excelência acadêmica foram cruciais para a realização deste trabalho.

Agradeço imensamente a meus pais, Alfredo e Tyciana, pilares inabaláveis de amor e apoio. Vocês sempre almejaram ver-me crescer, proporcionando-me a base sólida sobre a qual construí meus sonhos e aspirações. A cada passo desta jornada, senti a força de seu encorajamento e a ternura de suas crenças em minhas capacidades.

Um agradecimento especial à minha namorada, Thércia, cuja presença em minha vida tem sido uma fonte de inspiração e força. Sua capacidade de iluminar os caminhos mais obscuros e trazer clareza aos meus objetivos tem sido inestimável. Você é verdadeiramente a metade que soma em 200%.

Expresso minha gratidão a Deus por me guiar com sabedoria e força ao longo deste caminho e por iluminar meu percurso com fé e esperança.

Também agradeço ao Ministro Mauro por suas orações e apoio espiritual, que foram fontes de conforto e orientação durante os momentos mais desafiantes desta jornada.

Encerro estes agradecimentos com um sentimento de humildade e apreço por todos aqueles que, direta ou indiretamente, contribuíram para a realização deste trabalho. Cada palavra de incentivo, cada gesto de apoio, cada momento de compreensão foram essenciais para este feito.

"Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer."

(Alan Turing)

Resumo

Detecção automática do contexto de citações é uma tarefa crucial para a pesquisa e desenvolvimento acadêmico, pois reforça a integridade e o rigor científico. Este trabalho apresenta um método inovador para essa tarefa, combinando as capacidades do modelo SciBERT com a precisão da Máquina de Vetores de Suporte (SVM). O modelo SciBERT, pré-treinado em um vasto corpus de textos científicos, pode extrair características linguísticas que auxiliam na classificação de trechos como aqueles requerendo citações. A SVM, uma técnica de classificação bem estabelecida, então combina essas características para gerar uma classificação definitiva. O método proposto foi avaliado em um conjunto de dados de trechos de texto científico, alcançando uma melhoria de desempenho de 3% em relação ao estado da arte, demonstrando um avanço na identificação do contexto de citações. Esses resultados sugerem a eficácia de combinar um modelo de linguagem avançado com uma técnica de classificação bem estabelecida para a detecção automática do contexto de citações. Este método tem o potencial de aprimorar a qualidade e a confiabilidade dos trabalhos científicos.

Palavras-chave: Detecção de Contexto de Citação, Recomendação de Citação, SciBERT, SVM, ACL-ARC.

Abstract

Automatic detection of citation context is a crucial task for academic research and development, as it reinforces scientific integrity and rigor. This work presents an innovative method for this task, combining the capabilities of the SciBERT model with the accuracy of the Support Vector Machine (SVM). The SciBERT model, pre-trained on a vast corpus of scientific texts, can extract linguistic features that help classify excerpts as citations. SVM, a well-established classification technique, then combines these features to generate a definitive classification. The proposed method was evaluated on a dataset of scientific text excerpts, achieving a performance improvement of 3% in relation to the state of the art, demonstrating an advance in identifying the context of citations. These results suggest the effectiveness of combining an advanced language model with a well-established classification technique for automatic citation context detection. This method has the potential to improve the quality and reliability of scientific work.

Keywords: Cite-Worthy Context Detection, Citation Recommendation, SciBERT, SVM, ACL-ARC.

Lista de ilustrações

Figura 1 – Arquitetura de aprendizado profundo usada para a tarefa de classificação de contexto de citação 20
Figura 2 – Distribuição entre as classes do dataset ACL-ARC. 22

Lista de tabelas

Tabela 1 – Parâmetros de configuração do modelo LinearSVC 23
Tabela 2 – Resultados da categorização de sentenças quanto à citabilidade no conjunto de dados ACL-ARC (ordenados pelo F1 Score, decrescente). . 24
Tabela 3 – Avaliação do Método 29
Tabela 4 – Matriz de confusão 29
Tabela 5 – Matrizes de Confusão por Capítulos da Monografia. 30
Tabela 6 – Avaliação do autor por capítulo. 30

Lista de abreviaturas e siglas

SVM	*Máquina de Vetores de Suporte (Support Vector Machine)*
ACL-ARC	*Corpus de Referência da Antologia da Associação por Linguística Computacional (Association for Computational Linguistics Anthology Reference Corpus)*
MLP	*Perceptron Multicamadas* (Multilayer Perceptron)
RNN	*Rede Neural Recorrente (Recurrent Neural Network)*
CNN	*Rede Neural Convolucional (Convolutional Neural Network)*
BERT	*Codificador de Representações Bidirecionais a partir de Transformer (Bidirectional Encoder Representations from Transformer)*
RoBERTa	*BERT Robustamente Otimizado (Robustly Optimized BERT)*
MLM	*Modelagem de Linguagem Mascarada (Masked Language Modelling)*
NSP	*Predição da Próxima Sentença (Next Sentence Prediction)*
PLN	*Processamento de Linguagem Natural*
BiLSTM	*Memória de Longo e Curto Prazo Bidirecional (Bidirectional Long Short-Term Memory)*

Sumário

1 INTRODUÇÃO 12

2 FUNDAMENTAÇÃO TEÓRICA 14
2.1 Recomendação de Citação 14
2.2 Detecção de Contexto de Citação 14
2.3 BERT 15
2.4 SciBERT 16
2.5 Máquina de Vetores de Suporte 16

3 TRABALHOS RELACIONADOS 18

4 MÉTODO PROPOSTO 19

5 AVALIAÇÃO DO MÉTODO PROPOSTO 21

6 CONCLUSÃO 25
6.1 Trabalhos futuros 25

REFERÊNCIAS 26

7 APÊNDICE B -AVALIAÇÃO DO DESEMPENHO DO MODELO PRÉ-TREINADO PROPOSTO NO TEXTO DESTA MONOGRAFIA 29

1 Introdução

A utilização apropriada de citações em artigos científicos é importante para assegurar transparência, confiabilidade e veracidade nas alegações do autor. Esta prática é essencial para fundamentar afirmações, atribuir créditos e aprofundar o entendimento de conceitos e ideias (TEUFEL; SIDDHARTHAN; TIDHAR, 2009).

A complexidade dessa tarefa não reside apenas na obrigação de citar todas as fontes relevantes, mas também em incorporar citações que enriqueçam a compreensão do leitor sobre os objetivos do trabalho.

Com o aumento exponencial no número de publicações científicas a cada ano (BORNMANN; MUTZ, 2015), as demandas associadas à elaboração de citações de qualidade tornam-se substancialmente mais complexas. Manter-se atualizado com as inovações em uma determinada área torna-se uma tarefa desafiadora, considerando a constante evolução do conhecimento.

Em resposta à questão fundamental de como lidar com a abundância de informações disponíveis, surge a recomendação automática de citações. Recomendação de citação refere-se à tarefa de recomendar citações apropriadas, auxiliando o usuário a fundamentar uma passagem de texto dentro de um documento. A frase textual a ser fundamentada pode variar em comprimento — de uma palavra até um parágrafo — e é chamada de contexto de citação (JEBARI; HERRERA-VIEDMA; COBO, 2023).

Encontrar o equilíbrio certo entre o uso excessivo e a escassez de citações demanda tempo e experiência em escrita científica. Esta necessidade de equilíbrio se estende à tarefa de recomendação automática de citações (Bornmann & Mutz, 2015).

No entanto, as abordagens existentes para a recomendação de citações muitas vezes não abordam explicitamente a questão fundamental de determinar se um contexto específico, como uma sentença, realmente "necessita"de citações (ALI et al., 2020). Essa distinção, denominada "contexto que requer citação"(*cite-worthy context*), é o tema explorado nesta monografia como detecção de contexto de citação.

Esta etapa é vista como precursora à recomendação efetiva de citações relevantes, sendo que abordagens anteriores têm tratado essa tarefa como uma classificação binária (GOSANGI et al., 2021; ZENG; ACUNA, 2020; BONAB et al., 2018; FäRBER; JATOWT, 2018; SUGIYAMA et al., 2010).

Neste trabalho, propomos uma abordagem de aprendizado profundo que utiliza o modelo pré-treinado SciBERT (BELTAGY; LO; COHAN, 2019), para eficientemente resolver o problema da detecção de contexto de citação. Embasamos nossa escolha no

argumento de que *embeddings* - representações numéricas que capturam relações semânticas das palavras - obtidos com SciBERT podem capturar nuances de documentos científicos, proporcionando resultados competitivos com modelos simples de aprendizado de máquina. (MAHESHWARI; SINGH; VARMA, 2021)

Através de experimentos rigorosos, demonstramos a superioridade de nossa proposta na classificação de contexto de citação. Além disso, exploramos como nossa abordagem contribui para a eficácia na recomendação de citações, preenchendo uma lacuna na literatura e destacando a importância da detecção precisa de contextos de citação para aprimorar a qualidade e relevância das citações.

Objetivos

O objetivo geral deste trabalho consiste em propor um novo método de detecção de contexto de citação baseado em aprendizado profundo.

Objetivos Específicos

- Realizar a preparação de um dataset de documentos científicos para que seja detectada o contexto de citação.
- Desenvolver um novo método de detecção de contexto de citação baseado em aprendizado profundo que seja superior às abordagens existentes.
- Aplicar o novo método de detecção de contexto de citação a um conjunto de dados de documentos científicos e avaliar sua precisão.

2 Fundamentação Teórica

Neste capítulo serão abordadas algumas noções necessárias para uma melhor compreensão do que é apresentado na metodologia deste trabalho, o processo de deep learning e redes neurais, assim como o funcionamento das arquiteturas de redes utilizadas neste trabalho.

2.1 Recomendação de Citação

Sistemas de recomendação de citações evoluíram consideravelmente, abrangendo abordagens como Filtragem Baseada em Conteúdo, Filtragem Colaborativa e métodos Baseados em Grafos. Avanços tecnológicos, como MLP (Multilayer Perceptron - Perceptron Multicamadas), RNN (Recurrent Neural Network - Rede Neural Recorrente), CNN (Convolutional Neural Network - Rede Neural Convolucional) e embeddings, têm sido empregados nesses sistemas(ALI et al., 2020). Uma inovação recente é a aplicação de metodologias de multitarefa que utilizam estruturas de embedding, atenção e BiLSTM, direcionando a saída para várias tarefas de classificação, incluindo detecção de contexto de citação, antes de alimentar um recomendador de citações (VARANASI; GHOSAL; KORDONI, 2021), porém a precisão especificamente para a detecção de citações não é apontada, apenas para a tarefa de recomendação.

2.2 Detecção de Contexto de Citação

A detecção do contexto de citação é importante como etapa preparatória para a recomendação de citações. As técnicas evoluíram do uso de SVM com vetores de características baseados em unigramas e bigramas para o uso de CNNs com embeddings GLoVe(PENNINGTON; SOCHER; MANNING, 2014). Avanços subsequentes incluíram a combinação de CNNs com RNNs e a incorporação de embeddings BERT, como RoBERTa(LIU et al., 2019) e SciBERT, para processar contextos além de simples sentenças.

Sistematicamente, dado um conjunto de dados representado como um conjunto de sentenças em um artigo científico, denotado por $d = \{s_1, s_2, ..., s_n\}$, onde cada s_i representa a i-ésima sentença. Detecção de Contexto de Citação envolve a atribuição de uma de duas possíveis etiquetas a cada sentença s_i, sendo L = $\{l_c, l_n\}$. Aqui, l_c indica que a sentença requer uma citação, enquanto l_n indica o oposto, ou seja, que a sentença não requer uma citação.

Este problema de classificação binária é crucial para a compreensão da importância

das sentenças em um contexto acadêmico e serve como a base para a construção de um modelo capaz de automatizar a identificação de trechos que demandam referências bibliográficas. Essa definição orienta a pesquisa na busca por soluções eficazes e eficientes para a detecção precisa de contextos de citação em documentos científicos.

2.3 BERT

BERT (Codificador de Representações Bidirecionais a partir de Transformer)(DEVLIN et al., 2016) é um codificador bidirecional baseado em Transfomer de várias camadas. Essa arquitetura, característica dos Transformers, é fundamental para o funcionamento do BERT. Dispensando recorrências e convoluções, o Transformer utiliza mecanismos de atenção, incluindo a autoatenção, para capturar relações de longo alcance em sequências de texto. No contexto do BERT, essa abordagem permite que a rede aprenda representações contextualizadas e complexas, adaptando-se a nuances semânticas em dados textuais. O empilhamento de camadas no Transformer possibilita a aprendizagem hierárquica, contribuindo para a eficácia do BERT em tarefas de processamento de linguagem natural e classificação, destacando-se pela capacidade de lidar eficientemente com dependências de longo alcance em sequências.

Durante o pré-treinamento, o BERT utiliza duas tarefas principais: Modelagem de Linguagem Mascarada (MLM-*Masked LM*) e Predição da Próxima Sentença (NSP - *Next Sentence Prediction*). A MLM envolve mascarar aleatoriamente tokens de entrada e predizer esses tokens mascarados, enquanto o NSP visa antecipar se uma sentença segue a outra. Essas tarefas são fundamentais para promover a natureza bidirecional do modelo, permitindo que ele capture informações contextuais tanto à esquerda quanto à direita em sequências de texto.

O modelo é pré-treinado em um vasto conjunto de dados textuais, como o BooksCorpus (800 milhões de palavras) e a Wikipedia em inglês (2,5 bilhões de palavras). Essa escolha de um corpus generalista é respaldada por estudos que evidenciam a eficácia do pré-treinamento não supervisionado em grandes conjuntos de dados antes do ajuste fino para tarefas específicas. A generalidade do corpus permite ao BERT aprender padrões e características linguísticas úteis em uma variedade de domínios. (PHANG; FÉVRY; BOWMAN, 2018)

O BERT, com sua arquitetura avançada e o treinamento em grandes datasets, destaca-se como um modelo de referência para tarefas de processamento de linguagem natural. Sua capacidade de ajuste fino para tarefas específicas, como classificação, é uma demonstração clara de sua compreensão profunda das nuances semânticas do texto. Esse modelo prova ser uma ferramenta valiosa, eficaz e versátil em uma ampla gama de aplicações.

2.4 SciBERT

O progresso recente no campo do Processamento de Linguagem Natural (PLN) tem sido impulsionado pela adoção de modelos neurais profundos. Contudo, o treinamento desses modelos frequentemente demanda grandes volumes de dados rotulados. Em domínios gerais, a obtenção de dados extensos é possível por meio de crowd sourcing, porém, em domínios científicos, a coleta de dados anotados é desafiadora e dispendiosa devido à expertise exigida para a anotação de qualidade.

O SciBERT, conforme apresentado em Beltagy, Lo e Cohan (2019), representa um marco significativo no campo do Processamento de Linguagem Natural (PLN) científica. Esta variação do BERT adota a mesma arquitetura robusta, porém distingue-se pelo seu treinamento específico em textos científicos. Um aspecto notável do SciBERT é o seu vocabulário especializado, o SCIVOCAB, meticulosamente desenvolvido a partir de um vasto corpus científico. Em comparação com o vocabulário geral do BERT, o SCIVOCAB compartilha apenas 42% de seus tokens, refletindo uma diferença substancial no léxico frequentemente empregado em textos científicos e de domínio geral. Essa divergência ressalta a capacidade do SCIVOCAB em captar com precisão a linguagem e terminologia especializadas da ciência, essencial para adaptar o SciBERT às nuances linguísticas particulares do discurso científico.

Graças ao êxito do pré-treinamento não supervisionado em grandes corpora, o SciBERT oferece embeddings contextualizados altamente eficazes, resultando em um aprimoramento considerável na performance de diversas tarefas de PLN. Sua integração em arquiteturas neurais, mesmo aquelas de menor complexidade, para tarefas específicas, demonstra uma melhoria notável em comparação com abordagens gerais de PLN. A especialização e o desempenho aprimorado do SciBERT sublinham sua importância para o avanço contínuo no campo da PLN científica. Sua habilidade singular de extrair características linguísticas relevantes de textos científicos estabelece o SciBERT como uma ferramenta indispensável, não apenas para a compreensão e análise de linguagem em contextos científicos, mas também como um catalisador para inovações futuras na área.

2.5 Máquina de Vetores de Suporte

O SVM é uma técnica de aprendizado de máquina supervisionado, não sendo uma rede neural, utilizada para tarefas de classificação e regressão. O objetivo principal do SVM é encontrar um hiperplano que separe eficientemente instâncias de diferentes classes em um espaço multidimensional. Em um problema de classificação binária, o SVM busca um hiperplano que divide o espaço de características em duas regiões, uma para cada classe. Esse hiperplano é escolhido de forma a maximizar a margem entre as instâncias mais próximas de ambas as classes, chamadas de vetores de suporte.

Durante o treinamento, pesos, incluindo o vetor de pesos e o termo de polarização (bias), para encontrar o hiperplano de separação ideal.

Em espaços de alta dimensão, é mais provável que os dados sejam linearmente separáveis, o que é benéfico para os SVMs. SVMs podem ser computacionalmente eficientes, mesmo em espaços de alta dimensão.

Uma variante particularmente relevante para tarefas de classificação: o C-SVC, ou Classificação por Máquina de Vetores de Suporte com parâmetro C. Essencialmente, o C-SVC é uma formulação específica do SVM tradicional, projetada para lidar de forma eficiente com problemas de classificação. O elemento distintivo do C-SVC reside na introdução do parâmetro de regularização C, que desempenha um papel crucial no equilíbrio entre a precisão do modelo e a sua capacidade de generalização. O valor de C determina a tolerância do modelo a erros de classificação durante o treinamento, com valores menores de C aumentando a regularização e, consequentemente, incentivando uma maior generalização, enquanto valores maiores buscam aperfeiçoar a precisão do modelo nos dados de treinamento. A habilidade do C-SVC de ajustar esse equilíbrio torna-o uma ferramenta poderosa e flexível, adaptável a uma ampla gama de cenários de classificação, incluindo aqueles com alto grau de complexidade e variação nos dados.

3 Trabalhos Relacionados

Esta seção revisita trabalhos anteriores na área de detecção de contexto de citação, enfatizando as técnicas utilizadas e como elas se comparam à abordagem proposta neste estudo, que combina a SVM com o SciBERT no dataset ACL-ARC.

Sugiyama et al. (2010) aplicaram SVM para analisar a "citabilidade", usando características como unigramas e bigramas. No entanto, sua abordagem não considerou embeddings avançados e não explorou a comparação com outras metodologias, limitando-se a discutir a acurácia.

Bonab et al. (2018) demonstraram a eficácia dos embeddings GLoVe na melhoria dos resultados de classificação com CNNs, corroborando a importância de embeddings pré-treinados, um princípio que também observamos ao utilizar BERT em nosso estudo.

Wright e Augenstein (2021) avançou na área com um dataset especializado e um método baseado em Longformer (BELTAGY; PETERS; COHAN, 2020), ressaltando o valor do pré-treinamento de embeddings. Este estudo alinha-se com a nossa percepção sobre a importância de embeddings robustos.

Zeng e Acuna (2020) enriqueceram as características BERT com contexto sentencial adicional e uma camada de atenção. Embora tenham avançado na redução da dimensionalidade, não exploraram a combinação com técnicas de classificação tradicionais.

Outros estudos, como os de Maheshwari, Singh e Varma (2021) e Visser e Dunaiski (2022), integraram modelos derivados do BERT com redes neurais diversas, mas não investigaram a eficácia de abordagens de classificação mais simples.

Ao investigar classificação de textos da área médica, Magna et al. (2020) mostram que a combinação de embeddings BERT-like com SVM pode ser eficaz, uma direção que nosso trabalho expande para a área de detecção de citações.

Diferenciando-se dos trabalhos anteriores, nossa pesquisa apresenta uma abordagem inovadora ao combinar SciBERT com SVM no dataset ACL-ARC, alcançando um desempenho de estado da arte, com melhorias significativas em F1 score e Recall em comparação com metodologias anteriores. A escolha de reintegrar uma técnica de classificação tradicional, como SVM, com embeddings avançados, reflete nossa visão de que o potencial completo de técnicas clássicas ainda pode ser explorado com ferramentas modernas de processamento de linguagem natural. Esta abordagem representa um avanço significativo na detecção de contexto de citação, demonstrando a eficácia de combinar técnicas tradicionais e contemporâneas.

4 Método Proposto

A presente pesquisa aborda a detecção de contexto de citação em artigos científicos, uma tarefa fundamental para avaliar a relevância e a necessidade de referências em um texto acadêmico.

Conceitualmente, o conjunto de dados é representado como um conjunto de sentenças em um artigo científico, denotado por $d = \{s_1, s_2, ..., s_n\}$, onde cada s_i representa a i-ésima sentença.(GOSANGI et al., 2021)

Detecção de Contexto de Citação envolve a atribuição de uma de duas possíveis etiquetas a cada sentença s_i, sendo L = $\{l_c, l_n\}$. Aqui, l_c indica que a sentença requer uma citação, enquanto l_n indica o oposto, ou seja, que a sentença não requer uma citação.

Neste trabalho, propomos uma abordagem de aprendizado profundo para a tarefa de classificação de contexto de citação, conforme ilustrada na Figura 1. A abordagem consiste em duas etapas principais:

1. Representação de texto: A primeira etapa é a representação de texto, onde adotamos o modelo SciBERT. Esta escolha se destaca pela sua capacidade única de gerar representações vetoriais contextualizadas dos contextos de citação. O SciBERT é um modelo de linguagem pré-treinado com grande capacidade de captura de informações semânticas e sintáticas, sendo especialmente adequado para textos científicos. Essa etapa visa transformar as entradas em vetores de características, levando em consideração a complexidade e a especificidade da linguagem presente em documentos acadêmicos. Após passar por essa etapa, obtemos um embedding de palavras com 768 dimensões.

2. Classificação: A segunda etapa é a classificação, onde utilizamos uma SVM, a qual toma a entrada 768-dimensional e realiza um ajuste nos seus pesos para definir o melhor plano que divida as instâncias entre as classes objetivadas.

A combinação da representação de texto através do SciBERT e da classificação com SVCLinear [1] forma uma arquitetura coerente especificamente projetada para a detecção de contextos de citação em documentos científicos.

[1] LinearSVC é uma implementação linear do Classificador de Vetores de Suporte (SVC) no scikit-learn, uma biblioteca popular de aprendizado de máquina em Python. É especificamente projetado para tarefas de classificação linear em larga escala.

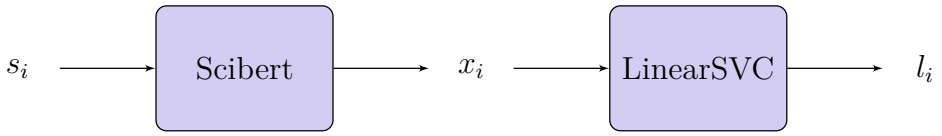

Figura 1 – Arquitetura de aprendizado profundo usada para a tarefa de classificação de contexto de citação

5 Avaliação do Método Proposto

Nesta seção, descreve-se o processo para avaliar o método proposto. A avaliação é executada sobre um conjunto de dados reais descrito na próxima seção, e a métrica usada para avaliar a sua acurácia é descrita na seção 5. Os hiperparâmetros do classificador LinearSVC são apresentados na seção 5. Os resultados são apresentados na seção 5.

Conjuntos de Dados

Para avaliar nossa abordagem, são necessários os conteúdos das publicações. Destaca-se que muitos conjuntos de dados acadêmicos disponíveis cobrem apenas os contextos de citação e não todas as sentenças das publicações ou abordam apenas meta-informações sobre as publicações, como a rede de citação.

Neste trabalho, seguimos a metodologia de Färber e Jatowt (2018), montando os dados de treinamento iterando sobre todas as sentenças em nossos dados de entrada (ou seja, texto simples de publicações), detectando se há um marcador de citação presente e rotulando as sentenças de acordo, antes de remover os marcadores de citação das sentenças. Após uma revisão dos conjuntos de dados acadêmicos disponíveis, decidimos utilizar o conjunto de dados da ACL-ARC, uma vez que ele vem sendo amplamente adotado em diversos artigos que investigam o mesmo problema de pesquisa abordado nesta monografia. ACL-ARC é um corpus reconhecido na área que consiste de publicações acadêmicas da área da linguística computacional.

Vale ressaltar que os conteúdos do ACL-ARC são extraídos de arquivos PDF, característica que confere ao corpus uma natureza especialmente ruidosa. Essa peculiaridade é considerada durante o processo de seleção, sendo importante para entendermos os desafios associados à manipulação desses dados. Dessa forma, optamos por incorporar o ACL-ARC em nossa análise, cientes de sua relevância e representatividade no contexto da pesquisa.

Devido ao acentuado desbalanceamento dos dados neste conjunto de dados como pode ser visto na Figura 2, é necessário realizar algum tratamento para evitar enviesamentos no treinamento. Esse desequilíbrio é resolvido por oversampling para todas as Redes Neurais (NNs) e por undersampling para todas as abordagens SVM, seguindo a abordagem de Färber e Jatowt (2018). No nosso método, usamos uma SVC Linear para classificar os dados. Para evitar que o modelo seja tendencioso para a classe majoritária, usamos o random undersampling para equilibrar as distribuições das classes. O random undersampling é um algoritmo não heurístico que elimina aleatoriamente instâncias da classe majoritária. Estudos mostram que o random undersampling é aproximadamente equivalente ao oversampling

para classificadores baseados em SVM (MOHAMMED; RAWASHDEH; ABDULLAH, 2020). No nosso caso, retiramos uma amostra da classe majoritária na mesma quantidade da classe minoritária.

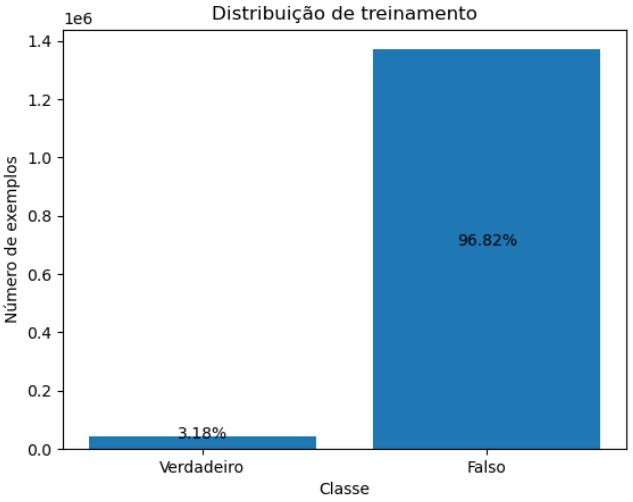

Figura 2 – Distribuição entre as classes do dataset ACL-ARC.

Métricas de Avaliação

Para avaliar o desempenho do nosso método, usamos as seguintes métricas:

Precisão: A precisão é a proporção das sentenças classificadas como citativas que realmente são citativas.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Em que:

TP é o número de sentenças classificadas como citativas que realmente são citativas FP é o número de sentenças classificadas como citativas que não são citativas

Recall: O recall é a proporção das sentenças citativas que foram corretamente classificadas como citativas.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Em que:

TP é o número de sentenças classificadas como citativas que realmente são citativas FN é o número de sentenças citativas que não foram classificadas como citativas

F1 Score: O F1 Score é uma média harmônica da precisão e do recall. Ele é uma medida mais equilibrada das duas métricas.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: A acurácia é a proporção de sentenças que foram classificadas corretamente, independentemente de serem citativas ou não.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Em que:

TP é o número de sentenças classificadas como citativas que realmente são citativas TN é o número de sentenças classificadas como não citativas que realmente não são citativas FP é o número de sentenças classificadas como citativas que não são citativas FN é o número de sentenças citativas que não foram classificadas como citativas

Configuração experimental

Os hiperparâmetros utilizados do classificador LinearSVC são mostrados na Tabela 1.

Tabela 1 – Parâmetros de configuração do modelo LinearSVC

Parâmetro	Valor
função de perda	squared_hinge
C	1.0
penalidade	L2
Número máximo de iterações	1000

Resultados e Discussão

Esta seção apresenta os resultados obtidos e os compara com os achados de outros trabalhos anteriores usados como referência.

Nosso método obteve resultados melhores na métrica de F1 Score do que o método do estado da arte, de Gosangi et al. (2021), no mesmo conjunto de dados (ACL-ARC), especialmente na métrica F1-score, como pode ser visto na Tabela 2. A métrica F1-score é a mais adequada para conjuntos de dados desequilibrados porque leva em consideração tanto a precisão quanto a recall.

O desempenho superior nas métricas F1 Score e Recall demonstra a capacidade superior do nosso método de identificar os contextos requerendo citações, que são uma

Tabela 2 – Resultados da categorização de sentenças quanto à citabilidade no conjunto de dados ACL-ARC (ordenados pelo F1 Score, decrescente).

Método	Precision	Recall	F1 Score	Accuracy
Nosso método	**0.766**	**0.761**	**0.760**	**0.761**
Gosangi et al. (2021)	0.813	0.662	0.730	0.770
Bonab et al. (2018)	0.4485	0.4056	0.4260	92.36%
Färber e Jatowt (2018)	0.196	0.269	0.227	0.941

classe rara, e a mais crítica. Essa importância é amplificada em cenários como o do conjunto de dados da ACL-ARC que adotamos, onde essa classe constitui apenas 3% dos dados.

O bom desempenho do SciBERT em tarefas de detecção de contexto de citação pode ser atribuído a uma combinação de fatores, incluindo:

- A multidimensionalidade do embedding gerado pelo SciBERT. Embeddings de alta dimensão são mais propensos a capturar as sutilezas do significado das palavras e frases. Isso pode ser particularmente importante para tarefas de linguagem natural, como a detecção de contexto de citação, onde é necessário entender o significado de palavras e frases em um contexto específico.
- O fato de que essas dimensões são treinadas para capturar características de textos científicos. Isso significa que o SciBERT é capaz de aprender a identificar palavras e frases que são comumente encontradas em citações.

Por último, é importante notar que o bom desempenho do método proposto em um conjunto de dados específico é um bom sinal, mas é preciso testá-lo em conjuntos de dados diferentes para confirmar sua superioridade. Além disso, outras configurações do BERT, como apresentadas em modelos derivados do mesmo, merecem ser testadas, mas não foram abordadas neste trabalho. Além disso, os parâmetros escolhidos na configuração do SVM foram baseados apenas na configuração padrão do mesmo, sendo interessante investigar diferentes configurações, especialmente quanto ao parâmetro C, para verificar a efetividade ou mesmo melhorar o método proposto. No entanto, o desempenho do método pode ser limitado devido à falta de uso de técnicas de busca de hiperparâmetro.

6 Conclusão

Este trabalho teve como objetivo principal desenvolver um método inovador para a detecção de contexto de citação em documentos científicos, empregando redes neurais artificiais. Através de uma abordagem que combina o poder do SciBERT com a eficiência da Máquina de Vetores de Suporte (SVM), este estudo conseguiu estabelecer um novo padrão de desempenho no dataset ACL-ARC. A integração inteligente e criteriosa entre um modelo de linguagem avançado e uma técnica de classificação robusta resultou em avanços na precisão da detecção de contexto de citação, um aspecto importante para a recomendação de citações.

Como todo estudo, este trabalho possui suas limitações. A abordagem desenvolvida foi testada especificamente no dataset ACL-ARC, o que sugere a necessidade de pesquisas futuras para avaliar sua aplicabilidade e eficácia em outros conjuntos de dados e contextos.

A capacidade de detectar com precisão o contexto de citação abre caminho para sistemas de recomendação de citação mais eficientes e precisos, contribuindo para a melhoria da gestão de informações científicas e facilitando o acesso a literatura relevante. Além disso, os métodos desenvolvidos neste trabalho podem ser adaptados para outras tarefas de PLN, demonstrando a flexibilidade e adaptabilidade das técnicas empregadas.

6.1 Trabalhos futuros

Olhando para o futuro, sugerimos a expansão desta pesquisa para incluir a aplicação do método em diferentes conjuntos de dados e a exploração de outros modelos de embeddings. Isso não só ajudará a testar a generalização da abordagem, mas também poderá revelar novas perspectivas e possibilidades dentro do campo do PLN. Acreditamos que este trabalho não apenas contribui para o campo do PLN, mas também abre novos caminhos para pesquisas futuras, incentivando a contínua busca por inovação e excelência em detecção de contexto de citação e áreas relacionadas.

Referências

ALI, Z.; KEFALAS, P.; MUHAMMAD, K.; ALI, B.; IMRAN, M. Deep learning in citation recommendation models survey. *Expert Systems with Applications*, Elsevier, v. 162, p. 113790, 2020. Citado 2 vezes nas páginas 12 e 14.

BELTAGY, I.; LO, K.; COHAN, A. Scibert: Pretrained language model for scientific text. In: *EMNLP*. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 12 e 16.

BELTAGY, I.; PETERS, M. E.; COHAN, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. Citado na página 18.

BONAB, H.; ZAMANI, H.; LEARNED-MILLER, E.; ALLAN, J. Citation worthiness of sentences in scientific reports. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2018. (SIGIR '18), p. 1061–1064. ISBN 9781450356572. Citado 3 vezes nas páginas 12, 18 e 24.

BORNMANN, L.; MUTZ, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, John Wiley & Sons, Inc., USA, v. 66, n. 11, p. 2215–2222, nov 2015. ISSN 2330-1635. Citado na página 12.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Bidirectional encoder representations from transformers. 2016. Citado na página 15.

FäRBER, A. T. M.; JATOWT, A. To cite, or not to cite? detecting citation contexts in text. In: PASI, G.; PIWOWARSKI, B.; AZZOPARDI, L.; HANBURY, A. (Ed.). *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*. [S.l.]: Springer, 2018. (Lecture Notes in Computer Science, v. 10772), p. 598–603. ISBN 978-3-319-76941-7. Citado 3 vezes nas páginas 12, 21 e 24.

GOSANGI, R.; ARORA, R.; GHEISARIEHA, M.; MAHATA, D.; ZHANG, H. On the use of context for predicting citation worthiness of sentences in scholarly articles. In: TOUTANOVA, K.; RUMSHISKY, A.; ZETTLEMOYER, L.; HAKKANI-TÜR, D.; BELTAGY, I.; BETHARD, S.; COTTERELL, R.; CHAKRABORTY, T.; ZHOU, Y. (Ed.). *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. [S.l.]: Association for Computational Linguistics, 2021. p. 4539–4545. Citado 4 vezes nas páginas 12, 19, 23 e 24.

JEBARI, C.; HERRERA-VIEDMA, E.; COBO, M. J. Context-aware citation recommendation of scientific papers: comparative study, gaps and trends. *Scientometrics*, v. 128, n. 8, p. 4243–4268, Aug 2023. ISSN 1588-2861. Citado na página 12.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. Citado na página 14.

MAGNA, A. A. R.; ALLENDE-CID, H.; TARAMASCO, C.; BECERRA, C.; FIGUEROA, R. L. Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *Ieee Access*, IEEE, v. 8, p. 106198–106213, 2020. Citado na página 18.

MAHESHWARI, H.; SINGH, B.; VARMA, V. Scibert sentence representation for citation context classification. In: *Proceedings of the Second Workshop on Scholarly Document Processing*. [S.l.: s.n.], 2021. p. 130–133. Citado 2 vezes nas páginas 13 e 18.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: IEEE. *2020 11th international conference on information and communication systems (ICICS)*. [S.l.], 2020. p. 243–248. Citado na página 22.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 14.

PHANG, J.; FÉVRY, T.; BOWMAN, S. R. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018. Citado na página 15.

SUGIYAMA, K.; KUMAR, T.; KAN, M.-Y.; TRIPATHI, R. C. Identifying citing sentences in research papers using supervised learning. In: *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*. [S.l.: s.n.], 2010. p. 67–72. Citado 2 vezes nas páginas 12 e 18.

TEUFEL, S.; SIDDHARTHAN, A.; TIDHAR, D. An annotation scheme for citation function. In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. USA: Association for Computational Linguistics, 2009. (SigDIAL '06), p. 80–87. ISBN 193243271X. Citado na página 12.

VARANASI, K. K.; GHOSAL, T.; KORDONI, V. Additional context helps! leveraging cited paper information to improve citation classification. In: *Proceedings of the 18th International Conference on Scientometrics and Informetrics, ISSI*. [S.l.: s.n.], 2021. p. 12–15. Citado na página 14.

VISSER, R.; DUNAISKI, M. Sentiment and intent classification of in-text citations using bert. In: *Proceedings of 43rd Conference of the South African Insti*. [S.l.: s.n.], 2022. v. 85, p. 129–145. Citado na página 18.

WRIGHT, D.; AUGENSTEIN, I. Citeworth: Cite-worthiness detection for improved scientific document understanding. *arXiv preprint arXiv:2105.10912*, 2021. Citado na página 18.

ZENG, T.; ACUNA, D. E. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. *Scientometrics*, Springer, v. 124, n. 1, p. 399–428, 2020. Citado 2 vezes nas páginas 12 e 18.

Apêndice A - Repositório contendo algoritmo desenvolvido utilizando a abordagem proposta

<https://drive.google.com/drive/folders/1s7Nfg-JH16dXjnWcCJRiG3wQ-AvOmVpK?usp=drive_link>

7 Apêndice B -Avaliação do desempenho do modelo pré-treinado proposto no texto desta monografia

O modelo desenvolvido nesse TCC foi aplicado ao texto desta monografia, considerando que o dataset possui similaridades com o tema tratado nela. Aplicando a nossa metodologia, identificamos 151 sentenças que foram rotuladas automaticamente. Devido ao tamanho reduzido do conjunto de dados, foi realizada uma confirmação da correção dos rótulos e da segmentação em sentenças. A Tabela 3 e 4 apresentam o desempenho do nosso método aplicado ao texto desse TCC.

Tabela 3 – Avaliação do Método

Precision	**Recall**	**F1**	**Accuracy**
0.2222	0.7200	0.3396	0.5364

Tabela 4 – Matriz de confusão

		Verdadeiro	
		True	**False**
Predito	**True**	18	63
	False	7	63

A Tabela 5 apresenta as matrizes de confusão por capítulo da monografia, mostrando o desempenho do método por capítulo.

Finalmente, a Tabela 6 apresenta a análise do autor sobre as recomendações dadas. A análise considerou a qualidade das citações realizadas no TCC. Em especial, o autor analisou 18 recomendações de citações referentes à sessão de Referencial Teórico e as considerou apropriadas, mas que o texto ainda precisa ser embasado com mais referências. A análise do autor sobre a recomendação dada revelou que o texto está bem fundamentado, mas que ainda precisa ser embasado com mais referências.

Tabela 5 – Matrizes de Confusão por Capítulos da Monografia.

Capítulo		**True**	**False**
Introdução	**True**	3	5
	False	4	8
Referencial Teórico	**True**	6	28
	False	1	14
Trabalhos Relacionados	**True**	5	7
	False	1	0
Método Proposto	**True**	3	13
	False	1	29
Resultados e Discussão	**True**	1	4
	False	0	8
Conclusão	**True**	0	6
	False	0	4

Tabela 6 – Avaliação do autor por capítulo.

Descrição	**I**	**RT**	**TR**	**M**	**RD**	**C**	**TOTAL**
Citação recomendada para sentença já citada	2	5	3	0	0	0	10
Citação recomendada já citada anteriormente	0	3	0	5	2	0	10
Citação recomendada, mas considerada não apropriada pelo autor	3	2	2	2	1	4	14
Citação recomendada apropriada, mas não citada pelo autor	0	18	2	6	1	2	29

I-Introdução, RT-Referencial Teórico, TR-Trabalhos Relacionados, M- Metodologia, RD-Resultados & Discussões, C-Conclusão.