INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS FLORIANO
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

MATEUS MIRANDA TORRES

ANÁLISE DE COLISÃO DE MARCAS NOMINATIVAS: UM ESTUDO SOBRE O DESEMPENHO DE MODELOS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

FLORIANO-PI
2023

MATEUS MIRANDA TORRES

ANÁLISE DE COLISÃO DE MARCAS NOMINATIVAS: UM ESTUDO SOBRE O DESEMPENHO DE MODELOS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Floriano.

Orientadora: Prof. Drª. Rafael Angelo Santos Leite.

FLORIANO-PI

2023

ANÁLISE DE COLISÃO DE MARCAS NOMINATIVAS: UM ESTUDO SOBRE O DESEMPENHO DE MODELOS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Mateus Miranda Torres[1]
Rafael Angelo Santos Leite[2]

RESUMO

Este estudo analisou a eficácia de diferentes modelos de Processamento de Linguagem Natural (PLN) na avaliação do risco de colisões entre marcas nominativas. Três modelos comuns de PNL para classificação de textos foram analisados - *Word2Vec*, *BERT* e *FastText* - e aplicados a 304 casos de marcas nominativas do banco de dados do Instituto Nacional da Propriedade Industrial (INPI) que tiveram registro negado devido a colisões. As métricas de Acurácia, Precisão, *Recall* e *F1-score* dos modelos foram calculadas para avaliar seu desempenho na identificação de riscos de colisão. Os resultados mostraram que, embora todos os modelos tenham alcançado 100% de *Recall* ao sinalizar todos os casos de colisão, o *BERT* teve melhor desempenho com 86,51% de Precisão e o *F1-score* de 92,77%, classificando corretamente a maioria dos casos. *Word2Vec* apresentou menor Acurácia e Precisão em comparação aos demais modelos. *FastText* também forneceu resultados relativamente precisos. Este estudo avança na compreensão da aplicação de abordagens de PNL como *Word2Vec*, *BERT* e *FastText* para analisar semanticamente a similaridade de marcas registradas. As descobertas oferecem orientação prática aos profissionais sobre como verificar colisões de forma eficiente e rápida usando técnicas de PNL.

Palavras-chave: Semelhança de marcas registradas; Processamento de Linguagem Natural; Colidência de Marcas Normativas.

ABSTRACT

This study analyzed the effectiveness of different Natural Language Processing (NLP) models in assessing the risk of collisions between word marks. Three common NLP models for text classification were analyzed - Word2Vec, BERT and FastText - and applied to 304 word mark cases from the National Institute of Industrial Property (INPI) database that were denied registration due to collisions. The models' Accuracy, Precision, Recall and F1-score metrics were calculated to assess their performance in identifying collision risks. The results showed that although all the models achieved 100% Recall when flagging all collision cases, BERT performed best with 86.51% Accuracy and an F1-score of 92.77%, correctly classifying the majority of cases. Word2Vec showed lower Accuracy and Precision compared to the other models. FastText also provided relatively accurate results. This study advances understanding of the application of NLP approaches such as Word2Vec, BERT and FastText to semantically analyze trademark similarity. The findings offer practical guidance to practitioners on how to check collisions efficiently and quickly using NLP techniques.

Keywords: Similarity of Trademarks; Natural Language Processing; Collision of Normative Trademarks.

Data de aprovação: 19/12/2023 (data de apresentação do TCC)

---

[1] Acadêmico do curso de TADS. IFPI Floriano. caflo2021114tads16@aluno.ifpi.edu.br
[2] Professor Eixo Gestão e Negócios. IFPI Floriano. rafaelangelo@ifpi.edu.br

1 INTRODUÇÃO

Marca é um sinal visualmente perceptível capaz de distinguir os produtos ou serviços de uma empresa, principalmente, em relação a outros concorrentes (INPI, 2023). O registro de marcas semelhantes ou idênticas é suscetível de causar confusão ou associação com marca alheia, conforme art 124, inciso XIX , da Lei de Propriedade Industrial (LPI) (BRASIL, 1996). Esse conflito pode gerar disputas legais entre as empresas envolvidas e atrasar ou impedir o registro da marca desejada (INPI, 2023).

De modo geral, as marcas registradas devem ser distintivas, pois, além das regras que impedem que as marcas sejam recusadas por motivos absolutos – como se pudessem enganar o público sobre a origem dos bens – elas podem ser recusadas por motivos relativos se entrarem em conflito com marcas existentes (FALL; GIRAUD-CARRIER, 2005). Assim, a verificação da similaridade entre marcas registradas é um processo fundamental para garantir a exclusividade da mesma, evitando conflitos de Propriedade Intelectual (PI).

De acordo com dados do boletim mensal de propriedade intelectual, o número de depósitos de marcas acumuladas em setembro de 2023 subiu 5.8%, totalizando um total de 416.541 pedidos nos últimos 12 meses (INPI, 2023). Esse grande volume de registros nos últimos anos representa uma forte procura por proteger marcas e identidades de negócios. Como resultado, o volume excessivo de pedidos adiciona pressão ao sistema de registro, que pode acabar tornando o processo mais demorado e burocrático do que o ideal. Felizmente, técnicas modernas de automatização do sistema podem ajudar a lidar com essa demanda acumulada de forma eficiente, agilizando o trâmite sem comprometer a análise minuciosa de cada solicitação. Isso traria benefícios tanto para quem procura proteger sua marca quanto para o órgão responsável pelo registro.

Os casos de infração estão relacionados à probabilidade de confusão do consumidor que, indevidamente, faz uma associação com outra marca. Isso pode ser causado por diversos fatores interdependentes, como a semelhança entre as embalagens, esquemas de cores, slogan e o nome da marca (SHARMA; SHARMA; AGGARWAL, 2016). Sendo evidente o fato de que é indispensável a verificação de similaridade entre marcas antes de solicitar o registro, a fim de evitar possíveis disputas judiciais ou administrativas.

A análise de colidência é feita quanto o aspecto gráfico (formas geométricas, imagens e sinais nominativos), fonético (sequência silábica, entonação das palavras e ritmo das frases) e ideológico (reprodução ou imitação de um conceito) quando existem semelhanças que geram risco de confusão ou associação indevida (INPI, 2023). Esse estudo se concentra no problema de colidência quanto aos aspectos gráficos, especificamente, sinais nominativos.

Recentemente, um aumento nos recursos computacionais, rápidos avanços no desenvolvimento de incorporações de palavras e Redes Neurais Profundas (RNPs) revolucionaram o campo do Processamento de Linguagem Natural (PLN) (SHOWKATRAMANI et al., 2019). O processamento em linguagem natural ensina aos computadores o significado semântico do texto em linguagem natural (STONE, 2007). A comparação conceitual de palavras e frases de marcas registradas é um problema no domínio da recuperação de marcas, que requer uma abordagem interdisciplinar envolvendo o PLN (SETCHI; ANUAR, 2016).

O problema em questão já havia sido abordado em estudos anteriores, nos quais se empregava PLN em conjunto com os modelos mais amplamente utilizados, como o Word2Vec. Sendo um método de medição de similaridade semântica proposto pelo Google para geração de vetores de palavras (TRAPPEY; TRAPPEY; LIN, 2020). Este modelo é uma rede neural superficial de duas camadas para medir a semelhança semântica entre palavras de texto. O BERT permite a codificação das propriedades semânticas de palavras por meio de representações vetoriais para organizar textos semelhantes próximos no espaço vetorial

(ARSLAN; CRUZ, 2022). O *FastText* aprimora os vetores de palavras com informações de subpalavras usando N-grama de caracteres adicional de comprimento variável, permitindo que o algoritmo identifique prefixos, sufixos, nuances morfológicas e estrutura sintática e tornando-a útil para idiomas morfologicamente ricos (SHOWKATRAMANI et al., 2019). Assim, acredita-se que o PLN seja uma alternativa promissora para auxiliar na verificação da colidência de marcas nominativas (REZENDE et al., 2022).

Nesse contexto, esta pesquisa tem como objetivo principal investigar o desempenho de modelos de PLN na verificação de risco de colidência de marcas, especificamente quanto aos sinais nominativos. A questão de pesquisa que orienta este estudo é: Qual o modelo de PLN mais eficiente para verificar riscos de colidência de marcas nominativas?

Para responder a essa questão de pesquisa, foram analisadas marcas nominativas com indeferimento de registro disponíveis no banco de dados do INPI. Foram aplicados e comparados três modelos: (I) *Word2vec* (REZENDE et al., 2022), (II) BERT (DEVLIN et al., 2019) e (III) *FastText* (JOULIN et al., 2016) para verificar o risco de colidência entre as marcas nominativas.

Os resultados deste estudo contribuem para o desenvolvimento de métodos mais eficazes na verificação de similaridade de marcas, bem como fornecem *insights* úteis para pesquisadores, profissionais de escritórios jurídicos no monitoramento e defesa de marcas dos seus clientes.

2 REFERENCIAL TEÓRICO

2.1 MARCAS

A LPI é um conjunto de direitos que engloba diversas formas de criação destinadas à atividade industrial. Entre essas criações, encontram-se as Marcas, Patentes de Invenção e Modelo de Utilidade, Registros de Desenho Industrial, e Indicações Geográficas, além de contemplar a repressão à concorrência desleal (BRASIL, 1996).

Marca é um sinal distintivo cujas funções principais são identificar a origem e distinguir produtos ou serviços de outros idênticos, semelhantes ou afins de origem diversa (ALSHOWAISH; AL-OHALI; AL-NAFJAN, 2022; INPI, 2023). Entretanto, marca na forma de apresentação nominativa, ou verbal, é o sinal constituído por uma ou mais palavras no sentido amplo do alfabeto romano, compreendendo, também, os neologismos e as condições e as combinações de letras e/ou algarismos romanos e/ou arábicos, desde que esses elementos não se apresentem sob forma fantasiosa ou figurativa (INPI, 2023).

Sendo assim, a proteção das marcas registradas é fundamental para garantir o direito à PI das empresas e fortalecer sua competitividade (HIEKATA; MOSER; INOUE, 2019; TRAPPEY; TRAPPEY; LIN, 2019). Isso garante que seus produtos e serviços sejam claramente distinguíveis dos produtos e serviços de outras empresas no mercado evitando a colidência entre marcas (NAKANO et al., 2016).

2.2 A COLIDÊNCIA DE MARCAS

A colidência ocorre quando uma semelhança gráfica (formas geométricas, imagens e sinais nominativos), fonética (sequência silábica, entonação das palavras e ritmo das frases), visual e/ou ideológica (reprodução ou imitação de um conceito) em relação a uma marca

anterior de terceiro suscetível de causar confusão ou associação com aquela marca alheia (INPI, 2023).

A semelhança gráfica do tipo sinais nominativos, nos quais a repetição de sequência de letras, número de palavras e estrutura das frases e expressões podem contribuir para gerar confusão ou associação indevida com outra marca já registrada na mesma classe. A Figura 1 demonstra um exemplo de colidência de sinais nominativos em que a presença da sequência de letras iniciais "MAND" em ambas as palavras, demonstra uma colisão.

Figura 1 - Exemplo de Colidência Nominativa

MEU MANDACARU para assinalar vestuário	X	**MANDAKKARU** para assinalar calçados

Fonte: (INPI, 2023)

Além disso, é importante considerar a Classificação Internacional de Produtos e Serviços de Nice (NLC) dividida em produtos (listados nas classes de 1 a 34) e serviços (listados nos números 35 a 45), adotado pelo INPI. Essa classificação define onde a marca será enquadrada, separando cada produto e serviço em seus respectivos mercados.

Contudo, PLN tem sido importante para análise de semelhança ortográfica, fonética e ideológica entre as marcas, com algoritmos treinados para reconhecer padrões de entoação, ritmo de frases e sequência silábicas que podem causar confusão ou associação indevida entre as marcas (TRAPPEY; TRAPPEY; LIN, 2020).

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O PLN é um subcampo da ciência da computação que se concentra em permitir que os computadores entendam a linguagem de forma "natural", como os humanos fazem (BEYSOLOW II, 2018). Nos últimos anos, o PLN experimentou avanços significativos com o surgimento da Inteligência Artificial (IA) (COHEN; FEIGENBAUM, 2014). Avanços em análise de sentimento (LIU, 2012), reconhecimento de voz (WEBER, 2003) e relação de padrões (CALIFF; MOONEY, 2003).

O PLN é um ramo importante da IA que adota algoritmos convencionais de aprendizado de máquina e métodos exclusivos de aprendizado profundo para construir modelos baseados em uma variedade de antecedentes de pesquisa para extrair informações e compreender documentos (KANG et al., 2020).

À medida que o PLN avança, alguns modelos emergem, como alguns mais especializados, *Word2Vec* (MIKOLOV *et al.*, 2013), Bidirectional Encoder Representations from Transformers (BERT) (DEVLIN *et al.*, 2018) e *FastText* (MIKOLOV, 2012)**.**

2.3.1 *WORD2VEC*

O *Word2Vec* é um algoritmo de aprendizado de máquina amplamente utilizado para criar representações vetoriais de palavras, também conhecidas como *word embeddings* (Incorporações de Palavras). Ele mapeia palavras de um vocabulário para vetores numéricos densos de tamanho fixo, nos quais a proximidade e a similaridade semântica entre as palavras são refletidas pela proximidade e orientação dos vetores no espaço vetorial (GOLDBERG; LEVY, 2014).

O *Word2Vec* é dividido em dois modelos chamados de Bolsa de Palavras Contínua (CBOW) e Skip-gram. A arquitetura do modelo CBOW usa o contexto da palavra para prever a palavra atual, como por exemplo: Olha que ____ mais linda / Olha que **coisa** mais linda. O Skip-gram usa a palavra alvo para prever o contexto da palavra, como por exemplo: Olha _?_ coisa _?_ _?_ / Olha que coisa mais linda (LEE et al., 2020). Para esse estudo o modelo a ser utilizado é o Skip-gram, pois estudos anteriores mostram que é mais preciso versus o CBOW (GOLDBERG; LEVY, 2014).

2.3.2 *BERT* (*BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS*)

O *BERT* foi introduzido em 2018 por Jacob Devlin e sua equipe do Google AI Language. O modelo é baseado em uma arquitetura de codificador transformer de várias camadas e é altamente eficaz em uma ampla variedade de tarefas de PLN, incluindo classificação de texto, resposta a perguntas e preenchimento de lacunas (DEVLIN et al., 2019).

BERT têm uma vantagem em estimar a similaridade de pares de sentenças devido a entradas duplas, atenção de múltiplas cabeças e captura de dependência de longa distância, como por exemplo: Eu amo __ cachorro / Eu amo **meu** cachorro (LI et al., 2020). Permitindo resultados de precisão sem precedentes em muitas tarefas automatizadas de processamento de texto (KOROTEEV, 2021)**.**

2.3.3 *FASTTEXT*

O *FastText* foi criado pelo laboratório de pesquisa de IA do Facebook. É um modelo simples, porém eficaz, que obtém resultados de ponta em tarefas de classificação de texto, com tempos de treinamento e inferência significativamente reduzidos em comparação com modelos de redes neurais complexas (JOULIN et al., 2016). Sendo uma abordagem poderosa para o PLN, oferecendo resultados precisos e rápidos na comparação de conjuntos de textos (PAUZI; CAPILUPPI, 2020).

Ele usa uma abordagem de aprendizado de máquina de várias camadas para aprender representações de palavras eficientes e compactas a partir de grandes quantidades de texto (QORINA et al., 2020). Podendo ser capaz de gerar melhores incorporações para palavras raras. Isso ocorre porque, mesmo que as palavras sejam raras, seus n-gramas de caracteres ainda são compartilhados com outras palavras, como por exemplo: Eu **amo** aprender / <am mo>, portanto as incorporações ainda podem ser boas (LIU et al., 2019).

2.4 TRABALHOS RELACIONADOS ENTRE PNL E ANÁLISE DE COLISÕES DE MARCAS NOMINATIVAS

A seleção dos modelos de PLN adotados neste estudo (*Word2Vec*, BERT e *FastText*) foi de uma análise prévia dos avanços reconhecidos na área de PLN com problema de classificação de textos. A Tabela 1 mostra a ocorrência desses modelos nos estudos mais recentes.

Tabela 1 - A presença de modelos de PLN em estudos recentes

Modelo de PLN	Presença nos estudos analisados	Referência
Word2Vec	4	(SHMATKOV; GOROKHOVATSKYI; VNUKOVA, 2023; TRAPPEY; TRAPPEY; LIN, 2020; JATNIKA; BIJAKSANA; SURYANI, 2019; GAO et al., 2017)
BERT	3	(DEVLIN et al., 2019; ARSLAN; CRUZ, 2022; YUAN; LEI; GUO, 2022)
FastText	3	(PAUZI; CAPILUPPI, 2020; ALAM; AFZAL; MALIK, 2020; LIU et al., 2019)

Fonte: Elaborada pelos autores (2023)

O modelo *Word2Vec* têm sido aplicado para problemas de análise de similaridade de texto, demonstrando avanço com alto desempenho em classificação de texto (SHMATKOV; GOROKHOVATSKYI; VNUKOVA, 2023; TRAPPEY; TRAPPEY; LIN, 2020; JATNIKA; BIJAKSANA; SURYANI, 2019; GAO et al., 2017).

Quanto ao modelo BERT, os estudos analisaram sua eficácia em várias tarefas de PLN, demonstrando resultados superiores em relação a outras abordagens existentes e concluíram que o BERT é um modelo poderoso e de fácil adaptação para diferentes tarefas (ARSLAN; CRUZ, 2022; YUAN; LEI; GUO, 2022; DEVLIN et al., 2019).

Quanto ao modelo *FastText*, os estudos analisaram seu desempenho em relação a outros modelos existentes, ele se destaca por sua velocidade de treinamento e concluíram que, apesar de ser simples, o *FastText* é competitivo com outros modelos disponíveis (PAUZI; CAPILUPPI, 2020; LIU et al., 2019).

2.5 AVALIAÇÃO DE DESEMPENHO DE MODELOS DE PLN

A avaliação empírica desempenha um papel central na estimativa do desempenho de modelos de PLN. Diversos indicadores unidimensionais são comumente utilizados nesse contexto, oferecendo uma visão abrangente sobre o desempenho do modelo.

Indicadores cruciais incluem Acurácia, *recall*, Precisão e *F1-score*, que são métricas amplamente reconhecidas na avaliação de modelos de classificação (QURASHI; HOLMES; JOHNSON, 2020). Conforme a Tabela 2, essas métricas são derivadas da análise da matriz de confusão, uma ferramenta fundamental na avaliação de modelos de classificação.

Tabela 2 - Matriz de Confusão

	Valor Previsto	
Valor Real	Verdadeiro Negativo (VN)	Falso Positivo (FP)
	Falso Negativo (FN)	Verdadeiro Positivo (VP)

Fonte: Elaborada pelos autores (2023)

Na Tabela 2, VN foram as instâncias corretamente classificadas como negativas pelo o modelo, FP as instâncias foram incorretamente classificadas como positivas pelo modelo, mas que na realidade são negativas. FN Instâncias que foram classificadas como negativas pelo o modelo, mas que na realidade são positivas. VP instâncias que foram corretamente classificadas como positivas pelo modelo.

O uso da matriz de confusão permite uma avaliação abrangente do desempenho de um classificador, fornecendo informações sobre a precisão, a taxa de acertos e os erros cometidos em diferentes classes. Essa ferramenta é fundamental para a análise e compreensão dos resultados de algoritmos de classificação em problemas de aprendizado de máquina (MARIA NAVIN; PANKAJA, 2016).

A partir dos resultados da matriz de confusão são fundamentais para o cálculo das métricas de avaliação.

Acurácia*:* $\frac{VN + VP}{VN + FP + FN + VP}$

Precisão*:* $\frac{VP}{VP + FP}$

Recall: $\frac{VP}{VP + FN}$

F1-score: $\frac{2 \times Precisão \times Recall}{Precisão + Recall}$

Assim, proporcionando uma análise abrangente e criteriosa do desempenho de cada modelo na identificação de riscos de colisões de marcas nominativas.

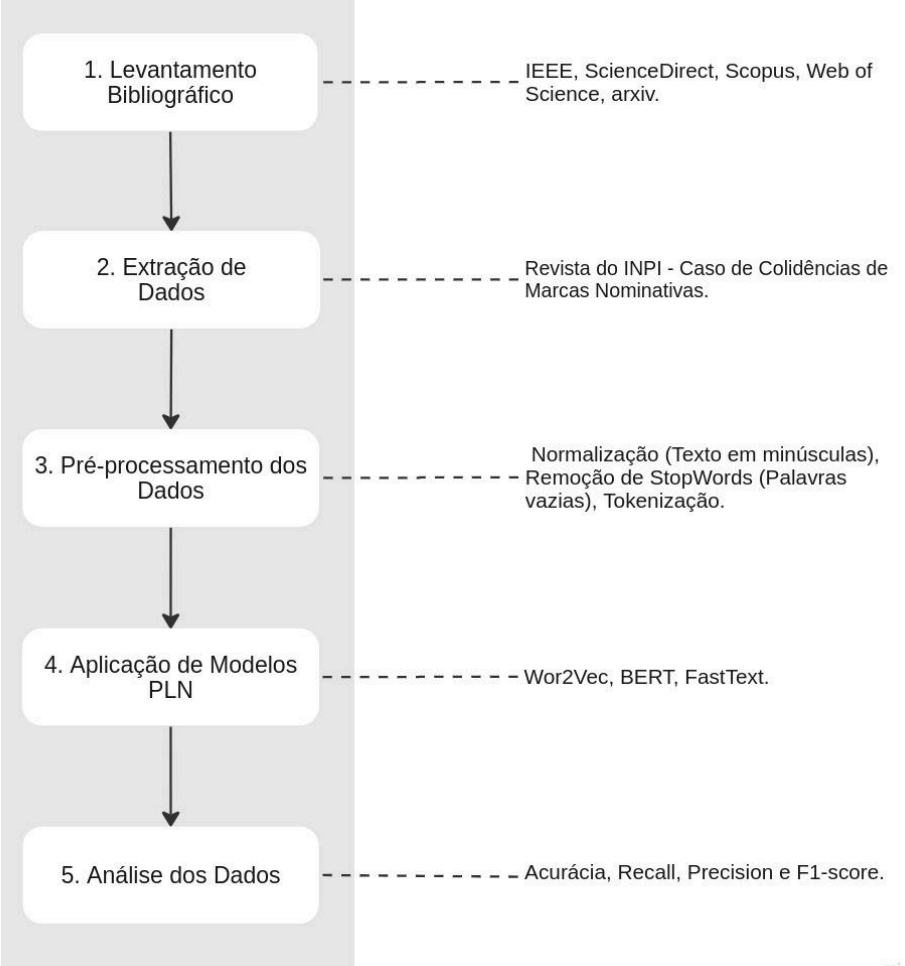
3 METODOLOGIA

Quanto à abordagem, este estudo é classificado como quantitativo, pois utiliza modelos de PLN para análise. Quanto ao objetivo, é explicativo, uma vez que envolve testes, comparações e interpretação dos resultados obtidos (YIN, 2009).

Quanto ao método de coleta de dados, trata-se de um estudo de monitoramento que acompanha casos de indeferimento de marcas por colidências com aspecto gráfico, mas especificamente sinais nominativos. Além disso, o estudo é classificado como *ex post facto*, pois o pesquisador observa o que já ocorreu, sem manipular as variáveis. Quanto ao tempo, é um estudo de natureza longitudinal, pois a extração dos dados das marcas nominativas foi realizada de 31 de agosto a 03 de novembro (COOPER; SCHINDLER, 2013).

A metodologia do estudo está ilustrada na Figura 2 e consiste em 5 etapas.

Figura 2 - Percurso metodológico



Fonte: Elaborada pelos autores (2023)

Conforme descrito na Figura 2, este estudo consistiu em um levantamento bibliográfico para identificar os modelos PLN existentes nas bases de pesquisa IEEE, ScienceDirect, Scopus, Web of Science e Arxiv.

A extração dos dados de marcas nominativas foi realizada através do *download* manual do arquivo em formato PDF (Formato de Documento Portátil), disponível na Revista da PI - http://revistas.inpi.gov.br/rpi/ - do INPI. Os dados foram coletados manualmente, sendo aqueles com motivo conforme o "Detalhes do despacho" destacado na Figura 3, resultando em um total de 304 marcas coletadas como colidentes.

Figura 3 - Exemplo de rejeição

829220267 Indeferimento do pedido
Titular: EPI - EMPREENDIMENTO PATRIMONIAL INDUSTRIAL S/A [BR/PR]
Procurador: Manoel Paixao do Nascimento
NCL(9): 21
Especificação: ESPONJAS DE LIMPEZA PARA USO DOMÉSTICO, ESCOVAS PARA LOUÇAS, ESCOVAS PARA CALÇADOS, ESCOVAS PARA ESFREGAR, VASSOURAS. (DA CLASSE 21)

Detalhes do despacho: A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregistrável de acordo com o inciso XIX do Art. 124 da LPI: Processo 828906564 (BRILLMAX).

Fonte: Revista INPI (2023)

Após a extração de dados estes foram salvos em uma nova base para análise e, posteriormente, foram pré-processados utilizando a biblioteca NLTK disponível na linguagem *Python*, assim realizando a remoção de *StopWord* (Palavras vazias), transformação em minúsculas e tokenização das palavras (divisão do texto em palavras individuais). Esse pré-processamento de dados desempenham um papel vital na limpeza de palavras, caracteres ou pontuações indesejadas que não são úteis para as máquinas interpretarem (TABASSUM; PATIL, 2020).

Após a conclusão do pré-processamento, a fase subsequente consistiu na aplicação dos modelos de PLN (*Word2Vec*, BERT e *FastText)*, previamente treinados skip-gram 50 (BOJANOWSKI *et al.*, 2017), bert-base-nli-mean-tokens (REIMERS; GUREVYCH, 2019) e wiki (BOJANOWSKI *et al.*, 2017) . A aplicação dos modelos produzem números em um espaço vetorial para serem calculados usando a semelhança do coseno. (LIMA; MAIA, 2018). A semelhança do cosseno é utilizada para avaliar a similaridade entre os vetores de palavras derivados dos modelos, permitindo a identificação de potenciais situações de confusão.

A semelhança do cosseno mede a similaridade entre dois vetores de um espaço de produto interno. Os intervalos do valor do cosseno são [0], [1], e quanto o grau mais próximo de 1, mais semelhante. E quanto mais próximo de 0, menor semelhante (CHENG; YU; WANG, 2017). Estabelece-se o limiar de 0,6 como critério para avaliação de similaridade.

Para análise dos resultados, são adotados como índices de avaliação neste artigo em destaque no tópico 2.5. Tais índices são úteis para analisar os resultados das classificações.

4 RESULTADOS E DISCUSSÃO

Nesta seção, os resultados dos modelos de PLN (*Word2Vec*, BERT e *FastText*) são apresentados e discutidos. A exploração desses resultados desempenham um papel crucial na compreensão do desempenho desses modelos.

Os resultados dos modelos foram avaliados usando métricas comuns de desempenho, como Acurácia, *Recall*, Precisão e *F1-score*. A Tabela 3 resume esses resultados.

Quadro 1 - Resultado dos modelos

Modelo	Acurácia *(%)*	*Recall (%)*	Precisão *(%)*	*F1-score (%)*
Word2Vec	73,36	100,0	71.48	83,37
BERT	86,51	100,0	86,51	92.77
FastText	84,21	100,0	83,89	91.24

Fonte: Elaborada pelos autores (2023)

Observa-se no Quadro 1, que todos os modelos atingiram um *Recall* de 100%, indicando que os 304 casos de marcas colidentes foram identificados pelos modelos, ou seja, nenhum caso teve uma similaridade igual a zero. No entanto, a Acurácia*,* Precisão *e o F1-score* variam entre os modelos.

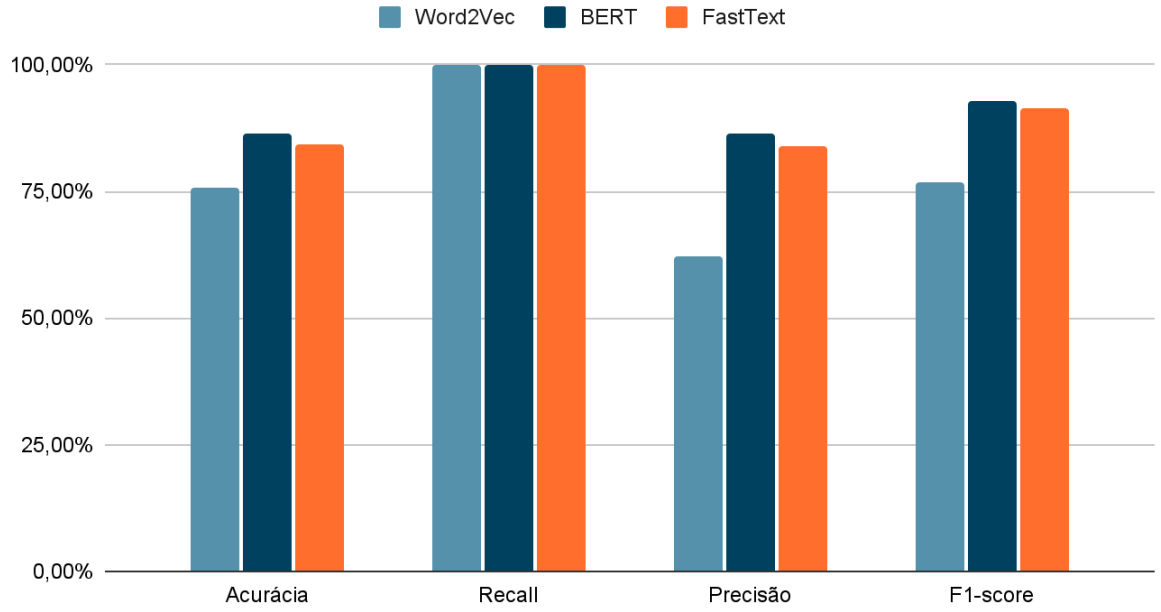
O *Word2Vec*, embora tenha alcançado uma Acurácia de 73.36% (223/304), apresentou a Precisão menor (71,48%), não classificando todos os casos com limiar igual a 0.6. Assim resultando em um *F1-score* menor em comparação aos outros modelos.

O modelo BERT demonstrou um desempenho sólido na verificação de riscos de colidência de marcas nominativas. Ele classificou corretamente 86,51% (263/304) das marcas,

com uma Precisão de 86,51% e um *Recall* de 100%. O *F1-score* de 92,77% indica que o modelo foi eficaz na identificação de marcas colidentes.

O *FastText* também obteve resultados promissores, obtendo 84,21% (255/304) de Acurácia, 83,89% de Precisão. Assim atingindo o *F1-score* de 91,24% indicando uma boa harmonia entre Precisão e *Recall*.

Gráfico 1 - Comparação de resultados experimentais



Fonte: Elaborada pelos autores (2023)

Conforme visto no Gráfico 1 os resultados do modelo *Word2Vec* não foram satisfatórios. Com base nos resultados, descobriu-se que os modelos BERT e *FastText* estão fornecendo resultados mais precisos do que a abordagem do modelo *Word2Vec*. No entanto, o modelo BERT mostrou-se como mais eficiente para a análise de risco de colidência de marcas nominativas.

Quanto ao modelo *Word2Vec*, nossos resultados estão de acordo com Trappey, Trappey e Lin (2019) que alcançaram uma precisão significativa na identificação de similaridade de marcas nominativas, atingindo uma acurácia de validação acima dos 75%, ou seja, um pouco acima da acurácia deste estudo.

Quanto ao modelo BERT, nossos resultados estão de acordo com Devlin et al. (2019) - que alcançaram resultados de 80,5% na acurácia - e com Yuan, Lei, Guo (2022) - que alcançou 87% na acurácia. Sendo este último, muito semelhante aos nossos resultados.

Quanto ao modelo *FastText*, nossos resultados estão de acordo com Pauzi, Capiluppi (2020) que alcançaram a acurácia de 80%. Sendo notavelmente próximos aos obtidos neste estudo.

5 CONCLUSÃO

Esse estudo buscou responder a seguinte questão de pesquisa: “Qual o modelo de de PLN mais eficiente para verificar riscos de colidência de marcas nominativas?”. Para isso, a abordagem metodológica adotada foi a utilização de modelos pré-treinados. Essa estratégia revelou-se eficaz dada a limitação de dados para o treinamento dos modelos.

Embora o *Word2Vec* tenha demonstrado a capacidade de identificar todas as correspondências arriscadas, sua exatidão e precisão foram deficientes em comparação com outras abordagens. BERT e *FastText* surgiram como principais concorrentes, classificando com precisão a maioria dos casos e alcançando pontuações muito altas na *F1-score*. Dos dois, o BERT destacou-se como a solução mais eficaz para a indagação deste estudo, identificando corretamente os riscos de colisão em quase 87% das vezes.

No âmbito teórico, este estudo avançou a compreensão da aplicação de modelos de PLN, como *Word2Vec*, BERT e *FastText*, na verificação de similaridade entre marcas registradas. A análise detalhada desses modelos proporcionou uma compreensão mais aprofundada sobre sua eficácia na identificação de riscos de colisão, contribuindo assim para o campo da Propriedade Intelectual e Direito de Marcas.

Do ponto de vista prático, os resultados indicaram que o modelo BERT se destacou como a abordagem mais eficaz, identificando corretamente os riscos de colisão em quase 87% das vezes. Essa descoberta tem implicações diretas para profissionais envolvidos no registro e proteção de marcas, sugerindo uma alternativa eficiente para agilizar o processo de verificação de colisões e, por conseguinte, otimizar o registro de marcas exclusivas.

A abordagem manual de coleta de dados, embora necessária para este trabalho exploratório, traz desafios inerentes à precisão e escalabilidade que métodos mais automatizados poderiam ajudar a reduzir as limitações na extração dos dados.

Considerando as limitações identificadas, sugere-se que estudos futuros explorem abordagens para aprimorar a precisão da coleta de dados, como o uso de técnicas de extração automática. A exploração de conjuntos de dados mais diversificados e a análise de casos práticos adicionais também podem enriquecer as descobertas e a aplicabilidade desses modelos.

REFERÊNCIAS

ALAM, F.; AFZAL, M.; MALIK, K. M. **Comparative analysis of semantic similarity techniques for medical text**. Em: 2020 International Conference on Information Networking (ICOIN), Anais...IEEE, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9016574/?casa_token=KPnwm5hZ9IMAAAAA:gcmPK2LyRa4VwMcqXjAUJFRNPYJVMz-5NN6So2D-rxLmfiFb6yQVbG2jtAkDT8X7Wnlii7DXM6W_BA. Acesso em: 12 dez. 2023.

ALSHOWAISH, H.; AL-OHALI, Y.; AL-NAFJAN, A. **Trademark image similarity detection using convolutional neural network**. Applied Sciences, v. 12, n. 3, p. 1752, 2022.

ARSLAN, M.; CRUZ, C. **Semantic taxonomy enrichment to improve business text classification for dynamic environments**. Em: 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Anais...IEEE, 2022. Disponível em: https://ieeexplore.ieee.org/abstract/document/9894173/?casa_token=B6yXfhgoxVgAAAAA:kwC2tgAE5N15G0LGmxLJkG9r_wLjB1e4wofzAfAZmCybbuetNXswmfRAUC7ofHOf0Ghht0-hYF7Y. Acesso em: 11 dez. 2023.

BEYSOLOW II, T. **Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing**. Berkeley, CA: Apress, 2018.

BOJANOWSKI, Piotr et al. **Enriching word vectors with subword information. Transactions of the association for computational linguistics**, v. 5, p. 135-146, 2017.

BRASIL. **Lei nº 9279, de 14 de maio de 1996**. Lei nº 9.279 de 14/05/1996. Diário Oficial da União, 15 maio 1996. Disponível em: https://legis.senado.leg.br/norma/551155. Acesso em: 29 out. 2023.

CALIFF, M. E.; MOONEY, R. J. **Bottom-up relational learning of pattern matching rules for information extraction**. Journal of Machine Learning Research, v. 4, n. Jun, p. 177–210, 2003.

CHENG, N.; YU, Z.; WANG, K. **String similarity computing based on position and cosine**. Em: 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), Anais...IEEE, 2017. Disponível em: https://ieeexplore.ieee.org/abstract/document/8076557/?casa_token=j7Due98fiW8AAAAA:8NleGjzzsd4Jg6ks-LMzqT0nAW-gWep7j5roCT9TYhyIs0CP2YU37e7VLyJsgBunanHmXuhxmHP8hg. Acesso em: 12 dez. 2023.

COHEN, P. R.; FEIGENBAUM, E. A. **The handbook of artificial intelligence: Volume 3**. [s.l.] Butterworth-Heinemann, 2014. v. 3

COOPER, D.; SCHINDLER, P. **Business Research Methods: 12th Edition**. [s.l.] Mcgraw-hill Us Higher Ed, 2013.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.**arXiv, , 24 maio 2019. . Disponível em: http://arxiv.org/abs/1810.04805. Acesso em: 11 dez. 2023.

FALL, C. J.; GIRAUD-CARRIER, C. **Searching trademark databases for verbal similarities**. World Patent Information, v. 27, n. 2, p. 135–143, 2005.

GAO, J. et al. **Duplicate short text detection based on Word2vec**. Em: 2017 8th ieee international conference on software engineering and service science (icsess), Anais...IEEE, 2017. Disponível em: https://ieeexplore.ieee.org/abstract/document/8342858/?casa_token=EfwyGrb1_CsAAAAA:r9ALBiNl5nwxXGqcD_X0hXpN8hfy_Smi60Aitjd2FqiPc6VVWOM7xzK_qpPQs1MO6Ct-WRSsYZ1S. Acesso em: 11 dez. 2023.

GOLDBERG, Y.; LEVY, O. **word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.** arXiv, , 15 fev. 2014. . Disponível em: http://arxiv.org/abs/1402.3722. Acesso em: 11 dez. 2023.

HIEKATA, K.; MOSER, B.; INOUE, M. **Transdisciplinary Engineering for Complex Socio-technical Systems**: Proceedings of the 26th ISTE International Conference on Transdisciplinary Engineering, July 30–August 1, 2019. [s.l.] IOS Press, 2019. v. 10

INPI. **Manual de Marcas**. Disponível em: http://manualdemarcas.inpi.gov.br/. Acesso em: 11 dez. 2023.

JATNIKA, D.; BIJAKSANA, M. A.; SURYANI, A. A. **Word2vec model analysis for semantic similarities in english words**. Procedia Computer Science, v. 157, p. 160–167, 2019.

JIN, X.; ZHANG, S.; LIU, J. **Word semantic similarity calculation based on word2vec**. Em: 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), Anais...IEEE, 2018. Disponível em: https://ieeexplore.ieee.org/abstract/document/8570612/?casa_token=e7H70up0kvwAAAAA:ph0CvGa7-OeqHQSHdFNkRdbQ7zhFmAHCUXLOZWfg8XdmHXe3_P9rAM3_uYV_Z4hxQpHAOsFchcq2. Acesso em: 18 dez. 2023.

JOULIN, A. et al. **Bag of Tricks for Efficient Text Classification**. arXiv, 9 ago. 2016. . Disponível em: http://arxiv.org/abs/1607.01759. Acesso em: 11 dez. 2023.

KANG, Y. et al. **Natural Language Processing (NLP) in Management Research: A Literature Review**. Journal of Management Analytics, v. 7, n. 2, p. 139–172, 2 abr. 2020.

KOROTEEV, M. V. **BERT: A Review of Applications in Natural Language Processing and Understanding**. arXiv, , 22 mar. 2021. . Disponível em: http://arxiv.org/abs/2103.11943. Acesso em:

11 dez. 2023.

LEE, C. et al. **Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database**. Technovation, v. 96, p. 102140, 2020.

LI, Z. et al. **Cross2Self-attentive bidirectional recurrent neural network with BERT for biomedical semantic text similarit**y. Em: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Anais...IEEE, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9313452/?casa_token=0wJ3U5JXDiYAAAAA:T36GgZqL60S4pYkKxUJziWziKPaByH6vr-hbGf7aEDnvp5j_a6SxPXFJNnP8BU5l-dXEWm49gk1r. Acesso em: 11 dez. 2023.

LIMA, J. M. C.; MAIA, J. E. B. **A topical word embeddings for text classification**. Em: Anais do XV Encontro Nacional de Inteligência Artificial e Computacional, Anais...SBC, 2018. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/4401. Acesso em: 12 dez. 2023.

LIU, X. et al. **Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding**. arXiv, , 20 abr. 2019. . Disponível em: http://arxiv.org/abs/1904.09482. Acesso em: 13 dez. 2023.

LIU, Bing. **Sentiment Analysis and Opinion Mining. Cham: Springer International Publishing**, 2012. E-book. ISBN 9783031010170. Disponível em: https://doi.org/10.1007/978-3-031-02145-9. Acesso em: 4 nov. 2023.

MARIA NAVIN, J. R.; PANKAJA, R. **Performance analysis of text classification algorithms using confusion matrix**. International Journal of Engineering and Technical Research (IJETR), v. 6, n. 4, p. 75–8, 2016.

NAKANO, S. et al. **Search for Similar Marks for Detecting Trademark Infringement and Dilution**. [s.d.]Disponível em: http://db-event.jpn.org/deim2016/papers/374.pdf. Acesso em: 11 dez. 2023.

PAUZI, Z.; CAPILUPPI, A. **Text Similarity Between Concepts Extracted from Source Code and Documentation**. Em: ANALIDE, C. et al. (Ed.). Intelligent Data Engineering and Automated Learning – IDEAL 2020. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. 12489p. 124–135.

QORINA, E. S. et al. **Comparative analysis of the performance of the fasttext and word2vec methods on the semantic similarity query of sirah nabawiyah information retrieval system: A systematic literature review**. Em: 2020 8th International Conference on Cyber and IT Service Management (CITSM), Anais...IEEE, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9268827/?casa_token=Gt75hLdnPP8AAAAA:u17kXZl97zMirex4Cm6rbJaAlBh-4iDWV1lwvmwj_qa06viQjATW6r_SzZjXHiIKX_rBmuJSsdHl. Acesso em: 11 dez. 2023.

QURASHI, A. W.; HOLMES, V.; JOHNSON, A. P. **Document processing: Methods for semantic text similarity analysis**. Em: 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Anais...IEEE, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9194665/?casa_token=0P7RszmBVuAAAAAA:zXAHgXe6nqUe6CElEZmbcKqVdayQHkHwb_7yxNNf1lwbHgmCHCDjMkuOCUDZDE94YK--Mwsg1sBauw. Acesso em: 12 dez. 2023.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, [*s. l.*], v. 5, p. 135–146, 2017.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep**

Bidirectional Transformers for Language Understanding. 11 out. 2018. Disponível em: http://arxiv.org/abs/1810.04805.

MIKOLOV, T. **Statistical language models based on neural networks**. [*S. l.: s. n.*], 2012. Disponível em: http://www.fit.vutbr.cz/~imikolov/rnnlm/google.pdf. Acesso em: 18 dez. 2023.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. **Efficient Estimation of Word Representations in Vector Space**. 16 jan. 2013. Disponível em: http://arxiv.org/abs/1301.3781.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 27 ago. 2019. Disponível em: http://arxiv.org/abs/1908.10084.

REZENDE, J. M. D. et al. **Combining Natural Language Processing Techniques and Algorithms LSA, Word2vec and WMD for Technological Forecasting and Similarity Analysis in Patent Documents**. Technology Analysis & Strategic Management, p. 1–22, 9 ago. 2022.

SETCHI, R.; ANUAR, F. M. **Multi-faceted assessment of trademark similarity**. Expert Systems with Applications, v. 65, p. 16–27, 2016.

SHARMA, G.; SHARMA, S.; AGGARWAL, A. **Trademark Infraction Prevention Policies and Challenges**. Indian Journal of Science and Technology, v. 9, p. 44, 2016.

SHMATKOV, D.; GOROKHOVATSKYI, O.; VNUKOVA, N. **Elaborative Trademark Similarity Evaluation Using Goods and Services Automated Comparison**. 2023. Disponível em: http://repository.hneu.edu.ua/handle/123456789/29710. Acesso em: 11 dez. 2023.

SHOWKATRAMANI, G. et al. **Deep learning approach to trademark international class identification**. Em: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Anais...IEEE, 2019. Disponível em: https://ieeexplore.ieee.org/abstract/document/8999317/?casa_token=IwwUaridhpMAAAAA:OQIqylijN8uey9iTJWfxgonleVkVtFGv_blBVTwPVbR3E7WzZkjkfte-CMR8BjMy5lfqZkoMCQkq. Acesso em: 11 dez. 2023.

STONE, A. **Natural-language processing for intrusion detection**. Computer, v. 40, n. 12, p. 103–105, 2007.

TABASSUM, A.; PATIL, R. R. **A survey on text pre-processing & feature extraction techniques in natural language processing**. International Research Journal of Engineering and Technology (IRJET), v. 7, n. 06, p. 4864–4867, 2020.

TRAPPEY, C. V.; TRAPPEY, A. J.; LIN, S. C.-C. **Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies**. Advanced Engineering Informatics, v. 45, p. 101120, 2020.

TRAPPEY, J. C.; TRAPPEY, C. V.; LIN, S. C. **Detecting trademark image infringement using convolutional neural networks**. Adv. Transdiscipl. Eng, v. 10, p. 477–486, 2019.

WEBER, D. **Object Interactive User Interface Using Speech Recognition and Natural Language Processing**. The Journal of the Acoustical Society of America, v. 114, n. 1, p. 32, 2003.

YIN, R. K. **Case study research: Design and methods**. [s.l.] sage, 2009. v. 5

YUAN, W.; LEI, Y.; GUO, X. **Research on text similarity calculation based on BERT and Word2Vec**. Em: ICETIS 2022; 7th International Conference on Electronic Technology and Information Science, Anais...VDE, 2022. Disponível em: https://ieeexplore.ieee.org/abstract/document/9788651/. Acesso em: 11 dez. 2023.