UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

João Otávio Rodrigues Ferreira Frediani

Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada

Bauru
2024

João Otávio Rodrigues Ferreira Frediani

Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Estadual Paulista "Júlio de Mesquita Filho".

Área de concentração: Computação Aplicada

Financiadora: CAPES

Orientador: Prof. Associado Aparecido Nilceu Marana

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Bauru
2024

UNIVERSIDADE ESTADUAL PAULISTA

Campus de Bauru

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE JOÃO OTÁVIO RODRIGUES FERREIRA FREDIANI, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 26 dias do mês de agosto do ano de 2024, às 09:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de JOÃO OTÁVIO RODRIGUES FERREIRA FREDIANI, intitulada **Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. APARECIDO NILCEU MARANA (Orientador - Participação Virtual) da UNESP/Bauru SP, Profa. Dra. ROBERTA SPOLON (Participação Virtual) do Departamento de Computação/UNESP/Campus de Bauru, Dr. GUSTAVO HENRIQUE DE ROSA (Participação Virtual) da Microsoft. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma virtual, o discente recebeu o conceito final: **APROVADO**. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo Presidente da Comissão Examinadora.

gov.br
Documento assinado digitalmente
APARECIDO NILCEU MARANA
Data: 26/08/2024 21:56:17-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. APARECIDO NILCEU MARANA

Frediani, João Otávio Rodrigues Ferreira.

Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada / João Otávio Rodrigues Ferreira Frediani. – Bauru, 2024

59 f. : il., tabs.

Orientador: Prof. Associado Aparecido Nilceu Marana

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Bauru, 2024

1. Detecção de Discurso de Ódio. 2. Processamento de Linguagem Natural. 3. Transferência de Aprendizagem. I. Marana, Aparecido Nilceu II. Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências. III. Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem.

CDU – 518.72:76

João Otávio Rodrigues Ferreira Frediani

Detecção de Discurso de Ódio na Língua Portuguesa Utilizando Transferência de Aprendizagem Supervisionada

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Estadual Paulista "Júlio de Mesquita Filho".

Área de concentração: Computação Aplicada

Financiadora: CAPES

Comissão Examinadora

Prof. Associado Aparecido Nilceu Marana
UNESP - Câmpus de Bauru - SP
Orientador

Dr. Gustavo Henrique de Rosa
MICROSOFT

Profa. Associada Roberta Spolon
UNESP - Câmpus de Bauru - SP

Bauru
26 de agosto de 2024

Agradecimentos

Agradeço à minha família, meus pais e meus irmãos e minha sobrinha, que me serviram de apoio e inspiração durante todo o mestrado.

Agradeço à meus amigos que tornaram meu caminho mais fácil e cheio de alegria.

Agradeço à meus professores, em especial à meu orientador Prof. Associado Aparecido Nilceu Marana que me acompanhou em mais uma etapa.

Agradeço à meus companheiros de laboratório, em especial Gabriel Lino e Leandro Passos, por seus esforços para me ajudar.

Agradeço à CAPES, pelo suporte financeiro concedido.

Agradeço à UNESP e ao PPGCC-UNESP pela infraestrutura e recursos disponibilizados para a realização deste trabalho.

Se você tentar abrir um gato para ver como ele funciona, a primeira coisa que você vai ter em suas mãos é um gato que não funciona.

Douglas Adams

Resumo

Discurso de ódio refere-se ao discurso ofensivo direcionado a um grupo ou indivíduo com base em características inerentes, como, por exemplo, raça, religião ou gênero. Já é reconhecido que discurso de ódio pode causar danos a longo prazo e criar problemas severos para a sociedade. O uso massivo da Internet intensificou a propagação deste tipo de discurso, permitindo que este chegue a muitas pessoa rapidamente, por isso, governos e empresas começaram uma batalha para combater sua propagação. Este combate é desafiador devido a quantidade de dados publicados na Internet, que torna a análise humana impossível, levando a necessidade de automatizar a detecção de discurso de ódio. Apesar das dificuldades encontradas, como o caráter implícito de alguns discursos, muitos trabalhos foram realizados em anos recentes para a detecção de discurso de ódio na língua inglesa. Para a língua portuguesa, a ausência de grandes conjuntos de dados rotulados torna o desafio ainda maior. Visando mitigar este problema, este trabalho investigou três estratégias de aprendizado de máquina que supostamente permitem a transferência de aprendizado em modelos de processamento de linguagem natural (PLN) desenvolvidos para detectar discurso de ódio em textos escritos em português. Foram utilizados os modelos Bertimbau Base, Bertimbau Large e mBERT, e exploradas três estratégias de transferência de aprendizado entre os idiomas inglês-português e espanhol-português: (i) a transferência de aprendizado de uma tarefa fonte para uma tarefa alvo distinta; (ii) a estratégia *zero-shot learning* e (iii) a estratégia *few-shot learning*. Experimentos realizados sobre conjuntos de dados disponíveis na literatura mostraram que a tarefa fonte escolhida (detecção de linguagem ofensiva) não gerou conhecimento relevante suficiente para melhorar a performance dos modelos de PLN na tarefa alvo deste trabalho (detecção de discurso de ódio). Eles mostraram também que o conhecimento se generalizou de maneira mais eficiente com a estratégia de *few-shot learning* do que com *zero-shot learning*, em especial entre os idiomas inglês e português. Por fim, um experimento adicional mostrou que técnicas de reamostragem dos dados, podem levar a uma melhoria no desempenho dos modelos de PLN, em particular quanto às métricas precisão, revocação e pontuação F1, quando as classes dos conjuntos de dados são desbalanceadas, como ocorre com os conjuntos de dados utilizados neste trabalho.

Palavras-chave: Discurso de ódio, Processamento de Linguagem Natural, BERT, Transferência de Aprendizado

Abstract

Hate speech refers to offensive speech directed at a group or individual based on inherent characteristics, such as race, religion or gender. It is already recognized that hate speech can cause long-term damage and create severe problems for society. The massive use of the Internet has intensified the propagation of this type of speech, allowing it to reach many people quickly, so governments and companies have begun a battle to combat its spread. This fight is challenging due to the amount of data published on the Internet, which makes human analysis impossible, leading to the need to automate the detection of hate speech. Despite the difficulties encountered, such as the implicit nature of some speech, much work has been carried out in recent years to detect hate speech in the English language. For the Portuguese language, the lack of large labeled datasets makes the challenge even greater. Aiming to mitigate this problem, this work investigated three machine learning strategies that supposedly allow transfer learning in natural language processing (NLP) models developed to detect hate speech in texts written in Portuguese. The Bertimbau Base, Bertimbau Large and mBERT models were used, and three transfer learning strategies were explored between the English-Portuguese and Spanish-Portuguese languages: (i) transfer learning from a source task to a distinct target task; (ii) the zero-shot learning strategy; and (iii) the few-shot learning strategy. Experiments performed on datasets available in the literature showed that the chosen source task (offensive language detection) did not generate enough relevant knowledge to improve the performance of the NLP models in the target task of this work (hate speech detection). They also showed that knowledge generalized more efficiently with the few-shot learning strategy than with zero-shot learning, especially between the English and Portuguese languages. Finally, an additional experiment showed that data resampling techniques can lead to an improvement in the performance of NLP models, particularly regarding precision, recall and F1 score metrics, when the classes of the datasets are imbalanced, as is the case with the datasets used in this work.

Keywords: Hate Speech, Natural Language Processing, BERT, Transfer Learning

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	Application Programing Interface
BERT	Bidirectional Encoding Representations from Transformers
BPE	Byte Pair Encoding
GPU	Graphics Processing Unit
LASER	Language-agnostic Sentence Representations
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long Short Term Memory
mBERT	Multilingual Bidirectional Encoding Representations from Transformers
MLP	Multilayer Perceptron
NB	Naive Bayes
NER	Named Entity Recognition
PLN	Processamento de Linguagem Natural
SVM	Support Vector Machine
UE	União Europeia
XLM	Cross-lingual Language Model

Sumário

1 Introdução

As redes sociais foram criadas com o intuito de permitir o compartilhamento de informações, exposição de opiniões e incentivar comunicações construtivas entre seus usuários, e foram extremamente bem sucedidas em juntar uma enorme quantidade de usuários. Segundo o *DataReportal* (KEMP, 2022), existem atualmente mais de 4 bilhões e meio de usuários ativos em redes sociais no mundo. Entretanto, as redes sociais têm sido exploradas para compartilhar linguagem abusiva e incentivar o ódio contra indivíduos ou grupos (MOZAFARI et al., 2019).

Fortuna and Nunes (2018) realizam uma análise sobre o discurso de ódio sob o ponto de vista da ciência da computação e trazem a seguinte definição:

> Discurso de ódio é a linguagem que ataca ou humilha, que incita a violência ou ódio contra grupos baseado em características específicas como aparência física, religião, descendência, nacionalidade ou origem étnica, orientação sexual, identidade de gênero ou outra, e pode ocorrer com diferentes estilos linguísticos, até de formas sutis ou quando humor é utilizado. (tradução realizada pelo autor)

As redes sociais têm sido meio de proliferação deste tipo de discurso devido à velocidade com que informações se propagam e à possibilidade de anonimato, causando uma escalada na violência no mundo (MOZAFARI et al., 2019). As vítimas de discurso de ódio podem sofrer consequências duradouras que se assemelham às consequências de crimes físicos, como assalto e violência doméstica (WILLIAMS, 2019).

A intolerância, como uma potencial experiência traumática, pode causar nas vítimas ansiedade, estresse e pensamentos depressivos, ao mesmo tempo em que incentiva a polarização, reforça esteriótipos sobre minorias e agrava a relação entre os grupos (OBERMAIER et al., 2021). Em casos extremos, a violência realizada pela internet pode levar ao suicídio, uma vez que vítimas e agressores envolvidos em casos de *bullying online* têm uma maior probabilidade em atentar contra a própria vida (HINDUJA; PATCHIN, 2010).

O problema dos discursos de ódio já tomou grandes proporções na vida real. Em 2016, uma grande escalada na disseminação de discurso de ódio na Europa relacionada a crise dos refugiados levou a União Europeia a criar um código sobre discurso de ódio (ROSS et al., 2017). Posteriormente, a Comissão da União Europeia pressionou o *Facebook*, *Youtube*, *Twitter* e a *Microsoft* a assinarem este código se comprometendo a analisar a maioria das denúncias sobre discurso de ódio em menos de vinte e quatro horas (FORTUNA; NUNES, 2018).

Devido ao grande volume de informações compartilhadas em todas as redes sociais, é impossível que seres humanos consigam analisar todo o conteúdo publicado nestas redes para verificar a presença de discurso de ódio. Passou-se, então, a pesquisar métodos automatizados

de identificação de discurso de ódio (SILVA; ROMAN, 2020). A busca de solução para este problema tem recebido muita atenção tanto do meio acadêmico quanto industrial, os quais encontraram no aprendizado de máquina, em particular no aprendizado profundo, meios para promover a automatização do processo de identificação de discursos de ódio (MOZAFARI et al., 2020).

Este problema, assim como diversos outros problemas de processamento de linguagem natural (PLN), apesar de muito investigado, tem sido focado na língua inglesa devido à grande quantidade de dados disponibilizados e à sua posição como língua mais comum na maioria das redes sociais (FORTUNA; NUNES, 2018). Em anos recentes, línguas como árabe, turco, holandês e alemão têm sido mais exploradas, porém idiomas como o português, coreano, francês e chinês seguem com poucas publicações na área (JAHAN; OUSSALAH, 2023).

Uma das razões que ocasionam essa dificuldade em explorar diferentes idiomas é a quantidade reduzida de conjuntos de dados rotulados, que dificulta a implementação de técnicas de aprendizado de máquina supervisionado. Atualmente, um dos conceitos mais utilizados para diminuir os efeitos que a pequena quantidade de dados rotulados tem no desempenho de métodos de aprendizado de máquina (*Machine Learning*, ou ML) é a transferência de aprendizado (*Transfer Learning*).

Transferência de aprendizado objetiva aplicar generalização de conhecimento para modelos de aprendizado de máquina, aproveitando modelos treinados em uma tarefa para solução de outra. A ideia é aproveitar o conhecimento e os recursos aprendidos com a tarefa de origem para melhorar o desempenho do modelo na tarefa de destino, mesmo que a tarefa de destino tenha uma distribuição de dados ou características diferentes.

Outra dificuldade associada aos dados da tarefa de detecção de discurso de ódio é que os dados são coletados diretamente de redes sociais, o que leva a uma distribuição desbalanceada das classes, podendo levar modelos a criar viés para as classes mais comuns (SILVA; ROSA, 2023).

1.1 Hipótese

Este trabalho parte da hipótese de que a utilização de técnicas de transferência de aprendizado pode melhorar o desempenho de detectores automáticos de discurso de ódio para a língua portuguesa, uma vez que esta sofre de baixa disponibilidade de recursos.

1.2 Objetivo

Este trabalho visou implementar redes neurais de processamento de linguagem natural que utilizam transferência de aprendizagem e investigar seus desempenhos na tarefa de detecção automática de discurso de ódio em textos escritos em português.

O campo de transferência de aprendizado é extenso e existem diversos métodos que buscam implementá-la. Este trabalho deu ênfase às técnicas de *fine-tuning*, *zero-shot learning* e *few-shot learning*.

Tendo em vista o desbalanceamento das classes nos conjuntos de dados utilizados neste trabalho, foi também investigada uma técnica de reamostragem baseada em *undersampling* para observar se o desequilíbrio das classes pode ser uma barreira para o desempenho das redes neurais implementadas.

1.3 Contribuição do Trabalho

A tarefa de detecção automática de discurso de ódio é ainda pouco explorada para a língua portuguesa e a maior parte dos trabalhos encontrados foca na utilização de métodos clássicos de aprendizagem de máquina, sendo que os que utilizam métodos mais modernos usualmente estão focados na solução do problema para múltiplas línguas simultaneamente. Esse trabalho traz os resultados de experimentos realizados com métodos atuais de processamento de linguagem natural, baseados em mecanismos de atenção, que utilizam transferência de aprendizado, e apresenta uma comparação dos desempenhos dos diferentes modelos e técnicas.

O trabalho também aborda o desbalanceamento das classes nos conjuntos de dados, realizando experimentos utilizando a técnica de reamostragem *undersampling*.

1.4 Organização da Dissertação

Esta dissertação está organizado em sete capítulos:

Capítulo 1: Este capítulo traz a introdução do trabalho;

Capítulo 2: Este capítulo trata dos conceitos teóricos relacionados a detecção automática de discurso de ódio e transferência de aprendizado;

Capítulo 3: Este capítulo trata do problema de detecção automática de discurso de ódio;

Capítulo 4: Este capítulo apresenta os trabalhos correlatos separados segundo o idioma de foco;

Capítulo 5: Este capítulo descreve as bases de dados utilizadas e detalha a metodologia utilizada neste trabalho;

Capítulo 6: Este capítulo apresenta as configurações dos experimentos realizados neste trabalho e discute os resultados obtidos com os experimentos;

Capítulo 7: Este capítulo apresenta as conclusões do trabalho realizado e aponta algumas possibilidades de trabalhos futuros.

2 Fundamentação Teórica

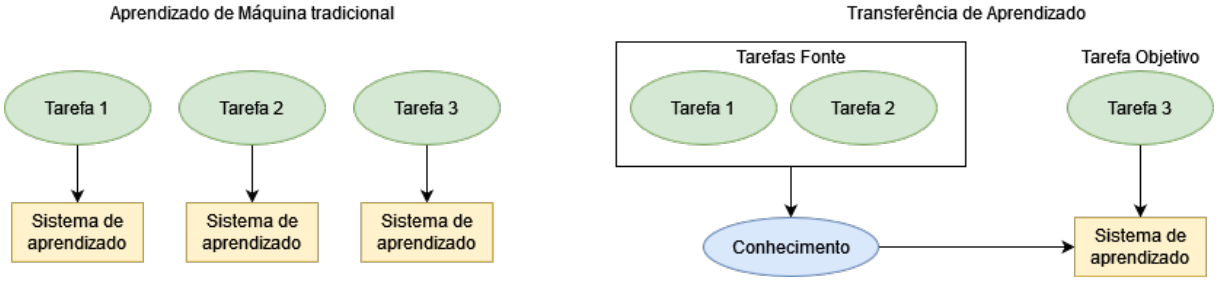
Neste capítulo apresenta-se a fundamentação teórica referente ao estudo realizado, que inclui os conceitos de transferência de aprendizado e processamento de linguagem natural.

2.1 Transferência de Aprendizado

Transferência de aprendizado é um conceito que surgiu inicialmente na Psicologia Educacional, sendo definida como resultado da generalização do conhecimento, como por exemplo uma pessoa que sabe tocar violino terá mais facilidade ao aprender a tocar piano, pois ambos são instrumentos musicais e compartilham um conhecimento semelhante para dominá-los, ou alguém que sabe andar de bicicleta terá mais facilidade em aprender a pilotar uma motocicleta. Entretanto, a transferência de aprendizado nem sempre é proveitosa, pois é necessário que haja semelhanças entre os domínios. Uma pessoa que sabe andar de bicicleta, por exemplo, não terá maior facilidade para aprender a tocar piano. Além disso, mesmo em situações em que os domínios são semelhantes é possível que as semelhanças sejam enganosas, como é o caso de pessoas que falam português e ao tentarem falar inglês cometem erros com falsos cognatos (pares de palavras que soam semelhante porém tem sentidos desconexos) (ZHUANG et al., 2020).

A transferência de aprendizado, para o aprendizado de máquina, busca a generalização do conhecimento em um domínio para ser aplicado em outro, sendo especialmente útil em casos em que o domínio objetivo possui poucos dados disponibilizados e em aprendizado multitarefa, em que um modelo de aprendizado de máquina é treinado para solucionar múltiplas tarefas simultaneamente. A Figura 1 ilustra o processo de aprendizagem no aprendizado de máquina tradicional supervisionado, onde o conjunto de treinamento e teste vem do mesmo domínio, e da transferência de aprendizado, onde o conjunto de treinamento vem, pelo menos parcialmente, de um domínio diferente do conjunto de teste (WEISS et al., 2016).

Figura 1 – Processo de aprendizagem de máquina tradicional supervisionado e por transferência de aprendizado (PAN; YANG, 2009).



Fonte: Pan and Yang (2009).

Como exemplo de transferência de aprendizado computacional, considere a existência de uma grande quantidade de dados rotulados para análise de sentimento de avaliações de câmeras digitais, entretanto há o interesse em preparar um algoritmo de análise de sentimento para avaliações de comida, problema que possui poucos dados rotulados. Tendo em vista que ambos os problemas possuem um número de características semelhante, se não igual, sendo textos escritos em uma mesma língua, a transferência de aprendizado pode ser empregada, considerando as avaliações de câmeras digitais como domínio fonte e as avaliações de comida como domínio objetivo (PAN; YANG, 2009).

Para formalizar o conceito de transferência de conhecimento é necessário primeiro definir os conceitos de domínio e tarefa (PAN; YANG, 2009).

Definição 1: Um domínio D é composto de duas partes, um conjunto de características X e a distribuição marginal $P(X)$, desta maneira tem-se:

$$D = \{X, P(X)\} \tag{1}$$

sendo X um conjunto de amostras definido como:

$$X = \{x | x_i \in X, i = 1, ..., n\} \tag{2}$$

Definição 2: Uma tarefa T consiste em um conjunto de rótulos Y e uma função implícita de decisão f a ser aprendida com o conjunto de dados, assim tem-se:

$$T = \{Y, f\} \tag{3}$$

Podemos, então, observar um domínio como um conjunto de dados com ou sem rótulos, como por exemplo o domínio D_s, correspondente a tarefa T_s, pode ser descrito como:

$$D_s = \{(x, y) | x_i \in X, y_i \in Y, i = 1, ..., n_s\} \tag{4}$$

Definição 3: Dados um domínio fonte D_s e uma tarefa T_s e um domínio objetivo D_t e uma tarefa T_t, a transferência de aprendizado busca facilitar o aprendizado de f_t treinando na tarefa T_s, onde $D_s \neq D_t$ e $T_s \neq T_t$.

É interessante notar que a condição $D_s \neq D_t$ implica que $X_t \neq X_s$ ou $P(X_t) \neq P(X_s)$, isto significa que as características dos conjuntos de dados são diferentes ou suas distribuições marginais são diferentes. No caso da análise de sentimento, por exemplo, a primeira opção pode significar que os textos estão escritos em línguas diferentes e a segunda que os textos tratam de contextos diferentes.

Quando temos $T_s \neq T_t$ significa que $Y_s \neq Y_t$ ou $f_s \neq f_t$. Ainda utilizando o exemplo de análise de sentimento, o primeiro caso pode indicar que um domínio trata o problema como

binário (positivo ou negativo) enquanto o outro possui 10 rótulos de diferentes sentimentos, e o segundo caso pode indicar que as classes estão muito desbalanceadas entre os domínios.

2.1.1 Fine-Tuning

Um dos métodos mais comuns de transferência de aprendizado é o processo de *fine-tuning* (ajuste-fino, em tradução livre). Este método se baseia em inicialmente treinar uma rede base em uma ou mais tarefas T_s do domínio fonte D_s, copiar as n primeiras camadas na rede final, adicionar uma ou mais camadas com seus valores iniciados aleatoriamente e então treinar a rede final para solucionar a tarefa objetivo. É possível escolher propagar o erro e ajustar os valores copiados da rede base ou congelar esses valores e treinar somente as camadas adicionadas para evitar *overfitting* (YOSINSKI et al., 2014).

2.1.2 Zero-Shot Learning

Zero-shot learning é um caso extremo de transferência de aprendizado em que um modelo é treinado para classificar dados não encontrados no conjunto de treinamento, sendo especialmente comum na área de processamento de imagens na qual modelos precisam detectar classes com poucos ou nenhum exemplo (POURPANAH et al., 2022).

Na área de processamento de linguagem natural, *Zero-shot learning* é empregado como uma técnica para aprendizado de línguas cruzadas, onde tenta-se transferir conhecimento obtidos em um idioma (normalmente com grandes quantidades de dados rotulados) para outro (com menor quantidade de dados disponíveis) (PAMUNGKAS et al., 2021). Portanto, neste contexto, *zero-shot learning* ocorre quando um modelo é treinado em uma língua porém testado em outra com que não teve qualquer contato.

2.1.3 Few-Shot Learning

Few-shot learning é uma estratégia menos extrema do que *zero-shot learning* uma vez que uma pequena quantidade de dados do domínio objetivo é inserida no conjunto de treinamento para permitir ao modelo ter contato com o conjunto em que será testado (PAMUNGKAS et al., 2021).

2.2 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é um subcampo interdisciplinar da Linguística, Ciência da Computação e Inteligência Artificial preocupado com as interações entre computadores e linguagem humana, em particular como programar computadores para processar e analisar grandes quantidades de dados de linguagem natural. O objetivo é um computador

capaz de "entender" o conteúdo dos documentos, incluindo as nuances contextuais da linguagem dentro deles (MANNING; SCHUTZE, 1999). A tecnologia pode, então, extrair com precisão informações e percepções contidas nos documentos, bem como categorizar e organizar os próprios documentos.

As tarefas de processamento de linguagem natural frequentemente envolvem reconhecimento de fala, compreensão de linguagem natural e geração de linguagem natural.

Para realizar tais tarefas é necessário converter os textos escritos em linguagem natural para um formato mais facilmente processável pelo computador, este processo é chamado de tokenização. Diferentes algoritmos exigem diferentes tipos de tokenização, porém uma etapa muito comum é a transformação de frases em sequência de *tokens*, que consistem na menor parte da frase a ser processada individualmente, podendo ser palavras, expressões, sequências de letras, entre outras (JURAFSKY, 2000).

2.2.1 Word Embedding

A representação de *tokens* em muitas técnicas sofre com generalizações que afetam o desempenho final em tarefas de processamento de linguagem natural, por exemplo, é normal que *tokens* como "futebol" e "vôlei" tenham representações completamente desconexas mesmo que possuam contextos significativamente semelhantes (LEVY; GOLDBERG, 2014).

A técnica de *word embedding* busca solucionar essa questão, trazendo representações para *tokens* que incluam contextos sintáticos e semânticos, os termos são então representados por longos vetores em que cada posição apresenta um valor numérico que relaciona o *token* com um contexto específico. Desta maneira, além de carregar grande quantidade de informações, esta técnica torna as operações computacionais fáceis uma vez que estas passam a ser operações de matrizes (LEVY; GOLDBERG, 2014).

Para exemplificação, segue um *embedding* fictício do *token* "rei":

$$\text{Rei} = [0{,}32 \quad 0{,}12 \quad 0{,}98 \quad 0{,}56 \quad 0{,}68]$$

Para simplificar o exemplo, o *embedding* foi representado como um vetor de dimensão 5, porém, o comum é que este valor seja maior. É possível imaginar que o valor 0.98 esteja relacionado a gênero, desta maneira valores próximos de 1 significam que o *token* é masculino enquanto valores próximos de zero que o *token* é feminino, entretanto estes valores são calculados pela máquina, sendo assim muito difícil realmente compreender qual característica está sendo quantificada.

Em alguns modelos, como é o caso do BERT (*Bidirectional Encoder Representations from Transformers*), que será abordado posteriormente neste trabalho, é possível utilizar um *token* especial para cada sequência a ser codificada, desta maneira a saída correspondente a

este *token* traz reunidas informações correspondentes a sequência inteira e pode ser utilizada para a classificação do texto (DEVLIN et al., 2018).

2.3 Mecanismo de Atenção

A atenção é um mecanismo cognitivo fundamental para a percepção humana. Quando se depara com algo desconhecido, o ser humano não realiza um processamento global para o decifrar, ele escolhe ignorar partes da informação e seletivamente se concentrar em outras partes. Por exemplo, ao encontrar uma imagem, o ser humano não busca absorvê-la por completo mas sim focar nos seus diferentes componentes a fim de compreendê-la. Desta maneira, os seres humanos processam, com alta eficiência, quantidades massivas de informação com uma quantidade limitada de recurso de processamento (NIU et al., 2021).

Para o ser humano existem dois tipos diferentes de atenção. O primeiro, mais subconsciente, é chamado de atenção de saliência, em que um estímulo exterior toma o foco da atenção, como, por exemplo, o ser humano tende a escutar a pessoa que fala mais alto em uma conversa, um exemplo computacional deste tipo de atenção é a operação de *max pooling* (NIU et al., 2021). O segundo, mais consciente, é chamada de atenção focada, em que o ser humano escolhe em que partes da informação quer focar para solucionar uma tarefa específica, o mecanismo de atenção busca simular o funcionamento deste tipo de atenção.

Para computar a atenção, uma rede precisa, inicialmente, codificar as informações do conjunto de dados em chaves k, que em processamento de linguagem natural podem ser os *embeddings* dos *tokens* de uma sequência, enquanto que em visão computacional podem ser partes de uma imagem. Um vetor q específico da tarefa é introduzido e representa a informação buscada.

A rede então calcula a relação de q e k utilizando uma função f para obter a pontuação de energia e:

$$e = f(q, k) \tag{5}$$

A função f determina como q se relaciona a k, podendo ser o produto escalar, a distância dos vetores, entre diversas outras opções.

Com as pontuações de energia calculadas, os pesos de atenção α são calculados:

$$\alpha = g(e) \tag{6}$$

em que g é uma função de distribuição, normalmente uma softmax.

Para calcular o vetor de contexto c é necessário introduzir v, uma representação vetorial que corresponde a exatamente uma chave k, sendo que k e v podem assumir o mesmo valor, ou podem ser diferentes representações do mesmo dado. Então, o vetor c pode ser calculado

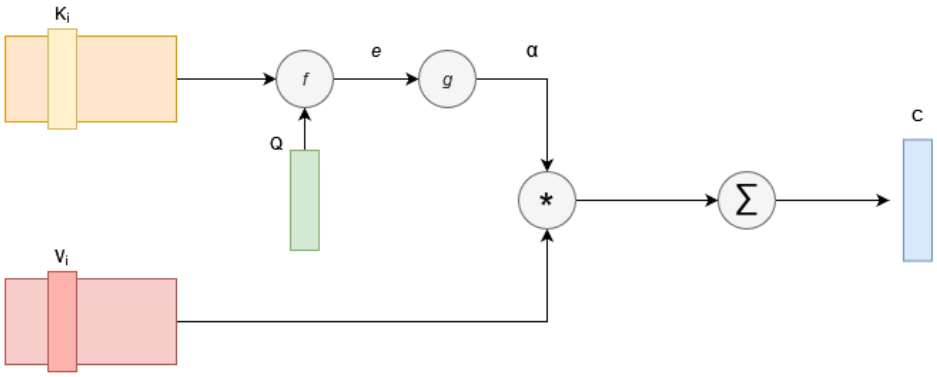
como:

$$z_i = \alpha_i . v_i \tag{7}$$

$$c = \sum_{i=1}^{N} z_i \tag{8}$$

A Figura 2 ilustra o conceito do mecanismo de atenção.

Figura 2 – Representação do mecanismo de atenção (NIU et al., 2021).



Fonte: (NIU et al., 2021).

2.3.1 Self-Attention

Self-attention (auto-atenção, em tradução literal), é uma variação do mecanismo de atenção proposta por Wang et al. (2016) que se baseia somente nos dados de entrada, ou seja os valores de k, q e v são diferentes representações da mesma sequência de entrada. Este modelo de atenção tem sido utilizado extensamente e é o método utilizado pelo modelo *transformer*. (DEVLIN et al., 2018).

2.3.2 Multi-Head Attention

Vaswani et al. (2017) propuseram a computação de múltiplas funções de atenção simultaneamente, projetando linearmente k, q e v, h vezes. Neste modelo, uma função de atenção é executada em cada projeção e as saídas são concatenadas e projetadas novamente para obter o vetor final. Desta maneira o cálculo da atenção se torna:

$$MultiHead(k, q, v) = Concatenação(head_1, head_2, ..., head_h) \tag{9}$$

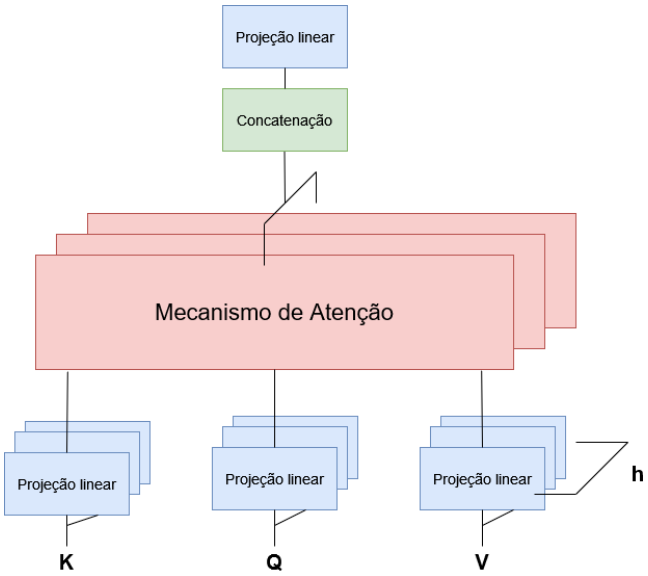
onde

$$head_i = (kW^k{}_i, qW^q{}_i, vW^v{}_i) \tag{10}$$

sendo W as matrizes de projeção.

A Figura 3 traz uma representação do mecanismo de *multi-head attention*.

Figura 3 – Representação do mecanismo multi-head attention (VASWANI et al., 2017).



Fonte: (VASWANI et al., 2017).

2.4 *Resampling*

Um problema relacionado a conjuntos de dados é o desbalanceamento de classes, em que uma ou mais classes compreendem um número muito maior de exemplos do que as outras (HAIXIANG et al., 2017). Treinar um classificador em um conjunto de dados que sofre de desbalanceamento de classes tende a levar a classificações enviesados que favorecem a classe majoritária enquanto negligencia a minoritária, resultando em uma performance insatisfatória (SILVA; ROSA, 2023).

Esse tipo de viés é normalmente abordado utilizando técnicas de *resampling* (reamostragem, em português). *Resampling* é portanto a técnica que busca equilibrar o número de exemplos para cada classe em um conjunto de dados, o que pode ser realizado de diferentes maneiras, por exemplo, *oversampling*, gerando amostras sintéticas da classe minoritária, e *undersampling*, que remove instâncias da classe maioritária (KAUR et al., 2019).

Neste trabalho a abordagem escolhida foi a *random undersampling*, que remove instâncias aleatórias da classe majoritária até uma proporção previamente definida. Apesar de seu funcionamento ser simples, *random undersampling* pode resultar em melhorias impactantes na performance final do classificador (SILVA; ROSA, 2023).

2.5 Modelos de Representação de Linguagens

2.5.1 O Modelo BERT

Bidirectional Encoder Representations from Transformers, ou simplesmente BERT, é um modelo de representação de linguagem introduzido por Devlin et al. (2018).

Este modelo foi criado com o objetivo de obter melhores representações de *tokens* ao coletar contexto bidirecionalmente, assim a representação é afetada tanto por *tokens* que surgiram antes na sequência quanto depois, adicionando dessa forma uma melhor compreensão de contexto para o algoritmo. O modelo é pré-treinado com uma grande quantidade de dados e depois pode ser especializado para uma outra tarefa de processamento de linguagem natural.

A arquitetura do modelo BERT se baseia fortemente na estrutura proposta por Vaswani et al. (2017) para o modelo chamado de *Transformer*[1]. Entretanto, o BERT se utiliza somente do componente referenciado na literatura como *Encoder*, sendo que na sua versão básica são empilhados doze destes componentes.

O *Encoder* é essencialmente formado por duas camadas: a camada de *multi-headed self-attention* e a camada de *feedforward*. A camada *feedforward* é uma camada completamente conectada simples. Entre cada camada é realizado um processo de normalização onde o valor passado para a próxima camada é o resultado da normalização entre o resultado da camada atual somado à entrada da camada. A Figura 4 mostra as camadas do *Encoder* do modelo BERT.

O modelo BERT aceita como entrada uma ou duas sequências, para problemas como o de resposta a perguntas, transformadas em uma sequência de *tokens*. O sistema de *tokens* utilizado é o proposto por Wu et al. (2016) com o algoritmo *Wordpiece* que conta com um vocabulário de 30.000 *tokens*. Toda sequência de *tokens* é iniciada com um *token* especial, [CLS], que ao final do processamento agregará informações da sequência inteira podendo ser utilizado em problemas de classificação, e um *token* [SEP] que demarca o final de uma frase. Como o modelo soluciona tarefas com múltiplas frases, este *token* também é utilizado para separar o início de uma e o final de outra.

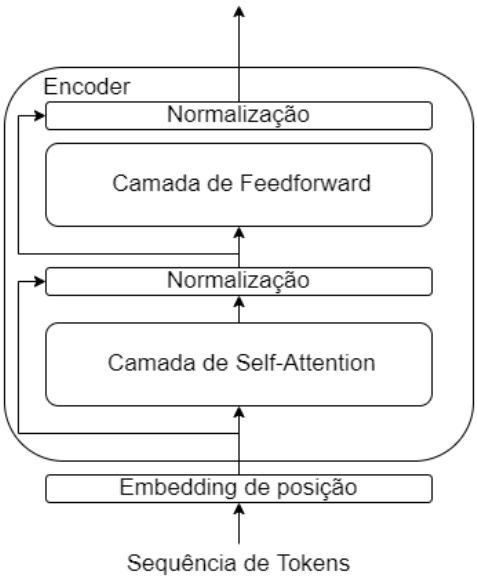
A entrada é então formada com a soma dos *embeddings* de cada palavra, aos de segmento e posição, como representados na Figura 5.

O algoritmo BERT é pré-treinado para conseguir uma melhor compreenção do funcionamento de linguagens naturais, ele é treinado em duas tarefas.

A primeira tarefa, objetivando um treinamento para representações bilaterais, é chamada de *masked language modeling* e consiste na substituição aleatória de quinze por cento dos

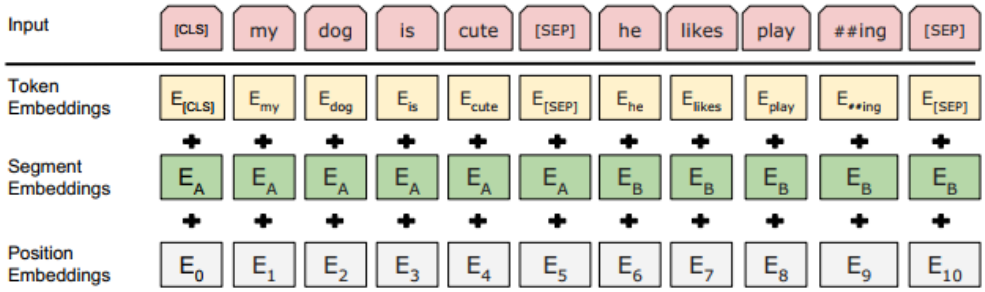
[1] O modelo Transformer é uma arquitetura para solução de tarefas de transdução de sequências que tem como principal estrutura o mecanismo de atenção, dispensando o uso de camadas convolucionais e recorrentes, e tem sido utilizado como base para diversos algoritmos de processamento de linguagem natural (ZHANG et al., 2021)

Figura 4 – Camadas do *Encoder* do modelo BERT.



Fonte:Elaborada pelo autor.

Figura 5 – Representação de entrada para o modelo BERT (DEVLIN et al., 2018).



Fonte: (DEVLIN et al., 2018).

tokens de todas as frases por um *token* de máscara, com o objetivo de que esta palavra seja descoberta unicamente por contexto. Entretanto, para evitar que o *token* de máscara influencie o procedimento de *fine-tuning*, em dez por cento dos casos a palavra escolhida a ser mascarada será substituida por algum outro *token* aleatório e em dez porcento mantém o *token* original, assim usa-se o *token* de máscara em oitenta por cento (DEVLIN et al., 2018).

Como exemplo segue a seguinte frase:

"[CLS] Deus ajuda quem [MÁSCARA] madruga [SEP]"

Então, utilizando apenas o contexto dado pelas palavras em volta o algoritmo deve buscar o *embedding* para o *token* máscara e alcançar valores semelhantes ao do *embedding* da palavra ocultada.

Como o modelo busca solucionar problemas como o de resposta a perguntas, a segunda tarefa foca na compreensão da relação entre duas frases. São, então, fornecidas duas frases,

sendo que em 50% dos casos a segunda frase é sequência da primeira e nos outros 50% dos casos a segunda frase é uma frase aleatória do conjunto de dados. Neste caso, a tarefa consiste em classificar a frase como sendo sequência ou não.

As duas frases a seguir representam esta tarefa:

Caso 1:"[CLS] Deus ajuda quem [SEP] cedo madruga"
Caso 2:"[CLS] Deus ajuda quem [SEP] não se olha os dentes"

O modelo deve então, utilizando seu conhecimento sobre linguagem natural, categorizar se as frases formam uma sequência, respondendo neste exemplo que no caso 1 as frases são uma sequência e que no caso 2 não.

2.5.1.1 Fine-Tuning no BERT

O pré-treinamento do modelo BERT é um processo longo e extremamento custoso, pois o modelo busca acumular conhecimento extenso sobre linguagem para então ser especializado em uma tarefa de processamento de linguagem natural. O modelo então passa por um processo de *fine-tuning*, como definido na Seção 2.1.1, e é treinado na tarefa objetivo em um processo mais curto e menos custoso que o pré-treinamento (DEVLIN et al., 2018).

O processo de *fine-tuning* do modelo BERT não congela os valores obtidos no pré-treinamento, pois as camadas iniciais do modelo acumulam conhecimento genérico sobre linguagem que são úteis para diversas tarefas de processamento de linguagem natural, enquanto as camadas finais armazenam conhecimento específico e são as mais afetadas pelo processo de *fine-tuning*. Este fato foi comprovado por Merchant et al. (2020) em um experimento baseado em congelar camadas do modelo durante o *fine-tuning* e avaliar as mudanças no desempenho, e após congelar as 8 primeiras camadas, ou seja modificando apenas as 4 últimas, a performance se mantém.

2.5.1.2 O Modelo mBERT

Inicialmente, o modelo BERT foi treinado como um modelo mono-linguístico em um grande conjunto de dados não rotulados em inglês, porém, como o algoritmo utiliza de transferência de aprendizado para gerar *embeddings* para as palavras, houve a intuição que este poderia ser treinado para lidar com mais de uma língua por vez.

O modelo foi então treinado com um conjunto enorme de dados composto de páginas do *Wikipedia* nas 104 línguas mais populares do site sem qualquer marcador denotando a língua, e foi chamado então de BERT multi-linguístico ou simplesmente mBERT. O mBERT foi testado em diversas linguagens e obteve um desempenho surpreendente mesmo sendo aplicado em casos de língua cruzada, ou seja quando o processo de *fine-tuning* é realizado em um idioma mas o problema solucionado está em outro (PIRES et al., 2019).

2.5.1.3 O Modelo BERTimbau

A transferência de aprendizado, definida na Seção 2.1, foi proposta como uma maneira de melhorar o desempenho de algoritmos, aproveitando o aprendizado que uma máquina pode obter de uma tarefa para solucionar uma outra tarefa relacionada, sendo, assim, especialmente útil em tarefas com poucos dados rotulados. Motivados por esta característica, Souza et al. (2020) realizaram o pré-treinamento do BERT em um *corpus* de páginas da internet contendo textos na língua portuguesa do Brasil.

O modelo foi chamado de BERTimbau e foi testado para as tarefas de reconhecimento de entidade nomeada (Name Entity Recognition, ou NER), similaridade de sentenças e reconhecimento de implicação textual, e em todas as tarefas o modelo obteve um desempenho superior ao modelos multilinguístico mBERT e à métodos de aprendizado de máquina tradicional, como LSTM e biLSTM.

3 Detecção de Discurso de Ódio

Este capítulo caracteriza o problema de detecção de discurso de ódio e discorre sobre algumas soluções que têm sido propostas.

3.1 Caracterização do Problema

Discurso de ódio tornou-se nos últimos anos um tópico muito popular, tanto pela cobertura midiática quanto pela atenção política. Porém, o tópico causa grandes discussões internacionais desde 1965 quando, em resposta ao *apartheid* e ao anti-semitismo pós guerra, a Assembleia Geral das Nações Unidas (AGNU) adotou a Convenção Internacional sobre a Eliminação de Todas as Formas de Discriminação Racial (ICERD). Em 2019, a AGNU, fortalecendo seu compromisso com o combate ao discurso de ódio, lançou a Estratégia e Plano de Ação sobre Discurso de Ódio (WILLIAMS, 2019).

A União Europeia (UE) também possui um histórico de combate ao discurso de ódio. Em 2003, adicionou um protocolo a Convenção de Cibercrime tratando da criminalização de atos racistas ou xenofóbicos cometidos por meio de sistemas de computadores. Como citado no Capitulo 1, a UE também criou, em 2016, um código de conduta visando diminuir o tempo de resposta a crimes de ódio em redes sociais (WILLIAMS, 2019). Inicialmente, este código de conduta produziu bons resultados, atingindo 90.4% de notificações avaliadas em menos de 24 horas em 2020. Porém, na avaliação publicada em novembro de 2022 este número abaixou para 64.4%, e também já havia abaixado em 2021 para 81%. A taxa de remoção também piorou na ultima avaliação com 63.3%, consideravelmente abaixo do ápice de 71% atingido em 2020 (UNIDAS, 2022).

Recentemente, o jornal *The New York Times* (FRENKEL; CONGER, 2022) publicou uma matéria sobre o aumento surpreendente de discurso de ódio no *Twitter* após ser comprado por Elon Musk, registrando uma aumento significativo no uso de palavras ofensivas a grupos como homossexuais, pessoas pretas e judeus. Esse aumento foi descrito como jamais visto em tão curto tempo em uma rede social, dentre as principais.

Apesar do problema do discurso de ódio ter um histórico longo e ter recebido atenção de grandes empresas e órgãos, ainda existem poucos estudos e trabalhos publicados acerca da detecção automatizada de discurso de ódio aplicando técnicas de processamento de linguagem natural (ARCO et al., 2021). Somado a esta baixa disponibilidade de trabalhos, existem diversos desafios relacionados à tarefa, começando pela sua definição, uma vez que diferentes fontes definem discurso de ódio da sua própria maneira. Neste trabalho, utiliza-se a definição dada por Fortuna and Nunes (2018) apresentada no Capitulo 1.

Outro ponto importante a se notar é que a detecção de discurso de ódio é muitas vezes confundida com linguagem ofensiva, que inclui insultos, palavrões ou ameaças. Linguagem ofensiva pode ser dividida em três subclasses separadas de menor a maior impacto social. A primeira é o "palavrão", o uso de palavras consideradas obscenas mas sem a intenção de ofender alguém, como "porra" ou "caralho". A segunda subclasse é o "insulto", quando a linguagem tem claramente a intenção de ofender um individuo associando a ele características desagradáveis. A última e mais impactante é o "abuso", em que para ofender um indivíduo associa-o a um grupo do qual ele seria um representante e, assim, ofende-se o grupo por completo, associando-o a características negativais universais e imutáveis. Assim, o discurso de ódio é parte do terceiro subconjunto, "abuso", em que se ofende um grupo completo relacionado a religião, origem étnica, orientação sexual, identidade de gênero e outros (ARCO et al., 2021).

Ainda ligado à dificuldade em definição do problema, mesmo agentes humanos discordam ao classificar textos como contendo ou não discurso de ódio. Para esta tarefa, é necessário possuir uma grande compreensão sobre a cultura e as estruturas sociais. Além disso, a constante evolução da linguagem associada às mudanças sociais dificultam ainda mais acompanhar todas as minorias (FORTUNA; NUNES, 2018).

Os conjuntos de dados para detecção de discurso de ódio são em grande parte retirados de redes sociais, ambiente em que a comunidade pesquisadora mais se atenta para a propagação deste tipo de discurso, sendo o *Twitter* a fonte mais explorada, seguido pelo *Facebook* e *Youtube* (JAHAN; OUSSALAH, 2023).

Somando então a dificuldade em definir conceitos, questionabilidade dos agentes que rotulam os dados e a necessidade de coletar dados de plataformas reais garantindo a privacidade dos autores tornam a criação de conjuntos de dados uma tarefa tão complexa quanto a detecção automática (JAHAN; OUSSALAH, 2023).

3.2 Soluções Propostas

Avanços recentes na área de processamento de linguagem natural permitiram a criação de modelos poderosos, como a série GPT (*Generative Pre-trained Transformer*), treinados em quantidades enormes de dados que, além de desempenhar excepcionalmente em diversas tarefas, também são capazes de gerar textos em linguagem natural (LIU et al., 2023). Estes modelos atingiram a área de classificação de textos com a possibilidade de não só classificar os textos mas também justificar a sua decisão de maneira compreensível ao ser humano. Dada a sensibilidade do problema de detecção de discurso de ódio foi proposta a utilização do modelo ChatGPT como uma ferramenta auxiliar de anotação de dados (HUANG et al., 2023).

Diversos métodos de aprendizado de máquina foram propostos para solucionar o problema de detecção de discurso de ódio, e que utilizam de diferentes métodos de aprendizado, supervisionado, semi-supervisionado e não supervisionado. Todos estes métodos apresentaram

bons resultados, sem indicar uma clara vantagem para uma abordagem particular. Em anos recentes, o modelo BERT passou a aparecer como um dos mais utilizados, porém redes LSTM e CNN ainda são muito populares (JAHAN; OUSSALAH, 2023).

Uma estratégia comumente pensada é a busca por palavras-chave, isto é, palavras que automaticamente indicariam a presença de discurso de ódio. Porém, apesar do conteúdo agressivo não ser incomum, existem autores de discurso de ódio que utilizam de linguagem gramaticalmente perfeita e sem nenhuma palavra ofensiva. Por outro lado, é muito comum pessoas usarem palavras ofensivas para sarcasmo (FORTUNA; NUNES, 2018).

Considerando estas condições, acredita-se que arquiteturas como o BERT, que focam em compreender a linguagem e o contexto em que palavras estão inseridas, possam ter um melhor desempenho do que o aprendizado de máquina supervisionado tradicional.

4 Trabalhos Correlatos

O impacto na vida real de mensagens publicadas na Internet já foi alvo de estudos de muitos pesquisadores. Hinduja and Patchin (2010), por exemplo, realizaram um estudo estatístico sobre a influência do *cyberbullying* em adolescentes, focando especialmente na possível ligação à suicídio e comprovaram que pessoas envolvidas nestes casos possuíam mais pensamentos suicidas e tinham uma maior probabilidade de atentar contra as próprias vidas.

Em relação ao discurso de ódio, Williams (2019), motivado pelo aumento significativo de crimes de ódio *online* entre 2017 e 2018, preparou uma visão geral sobre o problema de acordo com uma visão legal, avaliando situações reais e soluções colocadas em prática por diversos países, mas, focando especialmente no Reino Unido, e concluiu que ambos governos e empresas não estão fazendo o suficiente para realmente acabar com o problema.

Este capítulo apresenta uma revisão da literatura sobre trabalhos importantes e recentes, que abordam o problema de detecção automatizada de discurso de ódio na língua inglesa e em línguas não inglesas.

4.1 Detecção de Discurso de Ódio na Língua Inglesa

Diversas estratégias já foram colocadas em prática para automatizar o processo de detecção de discurso de ódio em textos, como o proposto por Djuric et al. (2015), que busca gerar uma representação em vetores de baixa dimensão que podem ser utilizados como entradas para classificadores binários. No momento da publicação a técnica atingiu resultados no estado da arte.

Yuan et al. (2016) buscaram identificar discriminação em textos utilizando uma arquitetura de aprendizado profundo em duas fases, a primeira camada é uma LSTM[2] treinada para obter representações para os textos com um pequeno conjunto de dados rotulados, e a segunda fase é uma regressão logística que utiliza as representações obtidas na primeira fase para classificar os textos. O método proposto obteve bons resultados na época de sua publicação. Entretanto, seu desempenho foi restringido pela pouca quantidade de dados disponíveis.

Badjatiya et al. (2017) estudaram a aplicação de diversos modelos de aprendizado profundo para a detecção de discurso de ódio e compararam com o desempenho de modelos clássicos que utilizam de *n-grams*, e concluíram que, em geral, os modelos de aprendizado profundo desempenham consideravelmente melhor do que os modelos clássicos.

[2] Arquitetura de rede recorrente proposta por Hochreiter and Schmidhuber (1997) que buscava solucionar o problema da dissipação do grandiente, o modelo se tornou dominante na área de aprendizado de sequência até o surgimento do modelo *transfomers* (ZHANG et al., 2021).

Frenda et al. (2019) realizaram um estudo sobre como o sexismo e o machismo se apresentam nas redes sociais e propuseram a utilização de um *lexicon* como forma de automaticamente identificar conteúdo com este tipo especifico de discurso de ódio. Entretanto, observaram que este tipo de estratégia apresenta diversas falhas.

Alguns trabalhos buscaram utilizar algoritmos de transferência de aprendizado para detecção de discurso de ódio. Mozafari et al. (2019) propuseram a utilização do modelo BERT para detecção de discurso de ódio em conteúdo retirado de redes sociais, realizaram testes em dois conjuntos de dados contendo textos em inglês publicados no *Twitter* e concluíram que o modelo pode atingir desempenho melhor do que outros propostos na área. No mesmo trabalho, os autores propuseram uma estratégia de *fine-tuning* para examinar os efeitos das diferentes camadas do BERT na detecção de discurso de ódio e concluíram que o modelo tem melhor desempenho devido a informação sintática e contextual aprendida pelas camadas do *Transformer*.

Mozafari et al. (2020) propuseram novamente a utilização do modelo BERT para a detecção de discurso de ódio, porém, dessa vez, utilizando um mecanismo para aliviar o viés racial contido nos conjuntos de dados. Neste trabalho, os autores avaliaram dois conjuntos de dados referência para detecção de discurso de ódio em inglês e concluíram que estes continham viés que levava classificadores a classificarem textos que utilizavam vocabulário afro-americano como contendo conteúdo racista e concluíram que a adição do mecanismo realmente aliviou este tipo de viés.

Khan et al. (2022) apresentaram o BiCHAT, um modelo para detecção de discurso de ódio em inglês que combina as representações contextuais do modelo BERT, com uma rede convolucional profunda e uma BiLSTM[3]. O modelo foi testado em três grandes conjuntos de dados retirados do Twitter, seu desempenho foi melhor que todos os outros modelos base do trabalho em todas as métricas. Os autores realizaram também um estudo para avaliar o impacto no desempenho de cada componente da arquitetura e concluíram que a camada convolucional profunda tem o maior impacto na performance quando removida.

Rana and Jha (2022) apresentaram um *framework* de aprendizado multimodal para detecção de discurso de ódio em conteúdo multi-mídia, os autores propuseram utilizar o áudio para extrair um contexto emocional do orador, e, assim, concatenar com informações retiradas por um modelo BERT do texto falado. Os experimentos apresentados no trabalho demonstram que a adição dessas características emocionais retiradas da fala são benéficas para o desempenho de modelos. Os autores também teorizam que a inclusão de características visuais retiradas de vídeos também podem ser benéficas ao identificarem símbolos relacionados a grupos odiosos e expressões faciais.

Huang et al. (2023) analisaram a possibilidade de utilizar LLM (*Large Language*

[3] Variação bidirecional da arquitetura LSTM, onde duas redes redes LSTMs são utilizadas, com uma delas recebendo a entrada na direção contrária.

Models) para detectar discurso de ódio implícito em textos e justificar em linguagem natural a classificação. Os autores desenvolveram um modelo de frase para questionar se um texto continha discurso de ódio e pedindo para justificar a decisão, essa frase foi então inserida no modelo ChatGPT e experimentos com suas respostas foram feitos com humanos. Com estes experimentos, os autores concluíram que na maioria das vezes o modelo concorda com a classificação humana e tende a gerar explicações mais claras e convincentes sobre sua decisão do que as explicações dos humanos.

4.2 Detecção de Discurso de Ódio em Línguas não Inglesas

Outros trabalhos buscaram a detecção de discurso de ódio em línguas diferentes da inglesa, explorando a tarefa em casos em que é pouco explorada. Vigna et al. (2017) propuseram o primeiro conjunto de algoritmos de detecção de discurso de ódio em italiano. Os autores exploraram os algoritmos SVM (*Support Vector Machines*) e LSTM para a tarefa utilizando um conjunto de textos retirados do *Facebook* e atingiram precisão de 80.60% com o SVM no problema binário.

Sohn and Lee (2019) propuseram o MC-BERT4HATE, um modelo multicanal com três versões de BERT, uma multilinguística, uma para inglês e uma para mandarim, assim o texto era inserido em sua língua original para o mBERT e traduzido para as demais línguas. O modelo foi testado em três conjuntos de dados, um em espanhol, um em italiano e um em alemão, e obteve desempenho comparável ou melhor que o estado da arte, sendo o melhor dos modelos avaliados para o espanhol, a melhor acurácia e segunda melhor pontuação F1 para alemão e a segunda melhor acurácia e pontuação F1 para italiano, ficando atrás do modelo mBERT.

Aluru et al. (2020) fizeram uma análise de diferentes algoritmos de aprendizado de máquina para a detecção de discurso de ódio em múltiplas línguas. Os testes foram realizados em dezesseis diferentes conjuntos de dados em nove línguas, incluindo a língua portuguesa, e concluíram que em geral o modelo mBERT ou uma tradução do texto para inglês seguida do modelo BERT desempenhavam melhor para todas as línguas, com exceção do português, que teve seu melhor desempenho obtendo representações LASER (Language-agnostic Sentence Representations, ou em tradução livre representações de sentença independentes de idioma), método que gera representações baseadas em pré-treinamento de sentenças completas independente do idioma em que estão escritas, seguido de uma camada de regressão logística para classificação.

Arco et al. (2021) realizaram uma comparação de diferentes técnicas de aprendizado profundo com algoritmos pré-treinados baseados em transferência de aprendizado para a língua espanhola, sendo os algoritmos testados em dois conjuntos de dados recentemente publicados de textos em espanhol retirados do *Twitter*. O trabalho concluiu que os algoritmos baseados

em transferência de aprendizado possuem melhor desempenho que os outros propostos neste trabalho e em trabalhos prévios. Adicionalmente, o modelo BETO, um modelo BERT pré-treinado em espanhol, possui melhor desempenho que os modelos multilinguísticos, podendo indicar que algoritmos mono-linguísticos tendem a desempenhar melhor na tarefa.

Ali et al. (2022) estudaram a capacidade de modelos baseados em BERT, como distilBERT e XLM-Roberta, em detectar discurso de ódio em textos escritos em urdu, idioma oficial do Paquistão. Os modelos obtiveram um desempenho ligeiramente superior do que a maioria dos modelos utilizados como base, porém não conseguiu superar o estado da arte. Entretanto, os autores teorizaram que o desempenho pode ser melhorado com o uso de modelos pré-treinados na língua urdu.

Fortuna et al. (2019), considerando a situação precária de dados disponíveis para a detecção de discurso de ódio na língua portuguesa, publicaram um conjunto de dados formado por 5.668 textos retirados do *Twitter* em português para a tarefa. Este conjunto possui dois anotadores, um binário classificando textos como contendo ou não discurso de ódio, e um anotador que categoriza o texto em 81 classes hierárquicas.

Silva and Roman (2020) realizaram uma comparação de desempenho de quatro diferentes modelos clássicos de aprendizado de máquina com diferentes configurações para a tarefa de detecção de discurso de ódio utilizando o conjunto de dados proposto por Fortuna et al. (2019). Os autores concluíram que o desempenho destes modelos é drasticamente afetado pelo método de representação dos textos e que seu desempenho, com exceção do classificador de *Naive Bayes*), é melhor que o do modelo mais recente LSTM, indicando que possivelmente modelos clássicos podem ser competitivos em relação a modelos modernos de aprendizado de máquina.

Assis et al. (2024) realizaram um comparativo entre a performance de modelos baseados em BERT e os modelos generativos *ChatGPT* e *MariTalks*, aplicando diferentes técnicas de *prompting*, para detecção de discurso de ódio em português com o intuito de observar seu desempenho em solucionar tarefas em idiomas com poucos recursos. Os autores concluíram que, apesar da recente popularidade dos *chatbots*, os modelos baseados em BERT ainda atingem melhores resultados na maior parte dos casos.

Firmino et al. (2024) investigaram se o uso de conjuntos de dados em um idioma pode ser benéfico para solução de uma tarefa em outro. Foram utilizados um conjunto em inglês e um em italiano para treinar um modelo BERT e um modelo XLM-R para detecção de linguagem ofensiva. Após, os modelos foram testados no conjunto OffComBr-2 em português. Os experimentos indicaram que o uso dos idiomas com mais recursos pode ser proveitoso para solucionar a tarefa em português, atingindo uma pontuação F1 de $0,92$ com ambas línguas no treinamento.

5 Material e Metodologia

Neste capítulo são apresentados os conjuntos de dados utilizados neste trabalho e a metodologia empregada para deteção de discurso de ódio na língua Portuguesa, baseado em transferência de aprendizagem.

5.1 Conjuntos de Dados

Esta seção apresenta os conjuntos de dados contendo discurso de ódio (Subseção 5.1.1), nas línguas portuguesa, inglesa e espanhola, e os conjuntos de dados contendo linguagem ofensiva (Subseção 5.1.2), na língua portuguesa.

5.1.1 Conjuntos de Detecção de Discurso de Ódio em Português

5.1.1.1 Fortuna

O conjunto de dados em português utilizado neste trabalho foi proposto por Fortuna et al. (2019) e é formado por publicações redigidas na língua portuguesa na rede social *Twitter*, coletados utilizando a *Twitter* API entre janeiro e março de 2017. O corpus é formado por 5.668 textos escritos por 1.156 diferentes usuários, dos quais 1.788 são classificados como contendo discurso de ódio.

A coleta dos dados foi realizada de duas maneiras: i) utilizando-se um filtro de palavras-chave; e ii) utilizando-se um filtro de usuários que mirava em perfis conhecidos por produzir conteúdo de ódio. Posteriormente, cada texto foi avaliado por 18 pessoas falantes nativas da língua portuguesa que julgaram se o texto continha ou não continha discurso de ódio. Finalmente, foi utilizado um sistema de votação majoritária para decidir a classificação de cada texto.

Os autores também propuseram uma distribuição hierárquica para diferentes categorias de discurso de ódio, como por exemplo o ódio contra lésbicas é uma subcategoria de homofobia, e separaram então os dados que continham discurso de ódio nestas categorias. Desta maneira é possível compreender as relações entre diferentes tipos de discurso de ódio e categorias mais incomuns de discurso de ódio são mantidas, e como as classes são não disjuntas os dados destas categorias participam também de categorias mais genéricas.

5.1.1.2 HateBR

O conjunto de dados HateBr, proposto por Vargas et al. (2021), consiste de textos em português coletados da seção de comentários de publicações no *Instagram* de políticos. Seis

contas públicas foram selecionadas, três de partidos conservadores e três de partidos liberais. Quinhentos comentários foram extraídos de publicações feitas no segundo semestre de 2018, resultando em 15.000 textos.

Os dados passaram por um processo de remoção de *links*, caracteres sem valor semântico e comentários que continham somente *emoticons*, risadas ou menções a outros usuários, culminando em 7.000 comentários. Os textos foram então classificados em três níveis: (I) Classificação binária de linguagem ofensiva, (II) Classificação da linguagem como altamente, moderadamente ou pouco ofensiva, (III) Classificação binária de discurso de ódio.

5.1.2 Conjuntos de Detecção de Discurso de Ódio em Outras Línguas

5.1.2.1 HateXplain

Para a realização dos experimentos utilizando a técnica de *zero-shot learning* e *few-shot learning* foi necessário escolher um conjunto de dados em inglês, uma vez que grande parte dos modelos é pré-treinada para lidar com problemas na língua inglesa. O conjunto escolhido foi o HateXplain, proposto por Mathew et al. (2021) com o objetivo de trazer dados que possam ser utilizados para explicabilidade de classificadores.

O conjunto HateXplain é formado por 20.148 textos retirados de duas redes sociais, sendo 9.055 do *Twitter* e 11.093 da *Gab*[4]. Os textos foram entregues a trabalhadores da *Amazon Mechanical Turk* (MTurk) que categorizaram o texto como normal, ofensivo ou contendo discurso de ódio, depois anotaram qual comunidade os textos visavam ofender se fossem da categoria ofensivo ou contendo discurso de ódio e, posteriormente, destacaram quais trechos do texto, frases ou palavras, os levaram a classificação.

5.1.2.2 HaterNet

HaterNet é um conjunto de dados de discurso de ódio em espanhol, proposto por Pereira-Kohatsu et al. (2019), formado de 6.000 textos retirados do *Twitter*. Os dados foram coletados, utilizando a *Twitter* API, em diferentes datas entre fevereiro e dezembro de 2017, para desta maneira evitar que eventos criem dados enviesados, com filtro para publicações realizadas na Espanha, resultando em um total de 2 milhões de *tweets*.

Os dados foram então limpos, removendo-se caracteres não alfanuméricos e substituindo risadas, menções, URLs e outros elementos sem informação semântica por *tokens*. Os dados foram então novamente filtrados para remoção de todos os textos que não continham qualquer palavra que indicasse insulto ou discurso de ódio, resultando em 8.710 textos.

[4] Rede social estadunidense lançada em 2016 como alternativa do *Twitter* que chegou a ser retirada da loja de aplicativos para celular da Google por permitir a proliferação de discurso de ódio (ZANNETTOU et al., 2018)

Finalmente, os dados foram encaminhados a quatro falantes nativos de espanhol para classificá-los segundo a definição de discurso de ódio fornecida pelos pesquisadores, o rótulo foi então definido pelo voto da maioria, com um quinto voluntário em caso de empate. Devido a restrições de tempo apenas 6.000 textos foram rotulados, dos quais 1.567 (26%) foram classificados como contendo discurso de ódio.

5.1.3 Conjuntos de Detecção de Linguagem Ofensiva

5.1.3.1 ToLDBR

ToLDBR (*Toxic Language Dataset for Brazilian Portuguese*, ou Conjunto de Dados de Linguagem Ofensiva para Português Brasileiro em tradução livre) é um conjunto de dados de linguagem ofensiva, criado por Leite et al. (2020), formado por 21.000 publicações coletadas do *Twitter*, categorizados em sete categorias de linguagem tóxica. Os dados foram coletados utilizando a ferramenta GATE *Cloud* [5] entre julho e agosto de 2019 utilizando duas estratégias. A primeira consistiu na busca de termos e *hashtags* usualmente associadas a publicações ofensivas. A segunda buscava evitar o viés que poderia ser introduzido buscando apenas por palavras chave, selecionando 50 usuários influentes da rede e coletando publicações que os citavam.

Ao todo foram coletados mais de 10 milhões de publicações e foram aleatoriamente selecionados 21.000, dos quais 60% são provenientes da primeira estratégias e o restante da segunda. Os textos foram encaminhados então para 42 voluntários selecionados para representar diferente demografias. Cada texto foi avaliado por três indivíduos e classificado em uma de sete categorias: LGBTQfobia, obscenidade, insulto, racismo, misoginia, xenofobia ou não ofensivo.

5.1.3.2 OlidBR

OlidBR (Offensive Language Identification Dataset for Brazilian Portuguese, ou Conjunto de Dados para Identificação de Linguagem Ofensiva em Português Brasileiro em tradução livre) é um conjunto de dados de linguagem ofensiva criado por Trajano et al. (2023), que contém $6,354$ textos coletados do *Twitter* e da seção de comentários do *Youtube*. Os dados foram categorizados por voluntários em três níveis: (I) classificação binária de linguagem tóxica, (II) classificação em 11 categorias de linguagem tóxica, (III) classificação de qual o alvo das ofensas.

5.2 Métricas

Para poder avaliar e comparar objetivamente o desempenho dos modelos nos diferentes experimentos realizados foi necessário selecionar métricas que quantificassem esse desempenho.

[5] Disponível em:<https://cloud.gate.ac.uk/info/about/twitter.html>

Foram selecionadas as métricas de acurácia, precisão, revocação e pontuação F1, que são comumente empregadas nos trabalhos correlatos.

Para facilitar a interpretação das equações para cada métrica é necessário definir alguns termos. Para o caso de problemas binários existem efetivamente 4 resultados para uma classificação, o *true positive* (TP), ou verdadeiro positivo em português, quando uma instância é classificada como positiva e ela é efetivamente positiva, o *true negative* (TN), ou verdadeiro negativo, quando uma instância é classificada como negativa e ela é efetivamente negativa, ambas compõe os acertos do modelo, o *false positive* (FP), ou falso positivo, quando uma instância é classificada como positiva e ela é na verdade negativa, e o *false negative* (FN), quando uma instância é classificada como negativa e ela é na verdade positiva, sendo que estes últimos dois compõem os erros do modelo.

5.2.1 Acurácia

A métrica mais comum para avaliação de classificadores binários é a acurácia, que calcula a proporção de acertos do modelo em relação a todas as classificações.

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

5.2.2 Precisão

A precisão tem como foco quantificar a capacidade do modelo em classificar corretamente as instância positivas.

$$Precisão = \frac{TP}{TP + FN} \tag{12}$$

5.2.3 Revocação

Revocação tem o foco oposto ao da precisão, buscando a capacidade do modelo em classificar corretamente as instâncias negativas.

$$Revocação = \frac{TN}{TN + FP} \tag{13}$$

5.2.4 Pontuação F1

Pontuação F1 é o resultado da média harmônica entre a precisão e a revocação, desta maneira um alto valor de F1 representa tanto um alto valor de precisão quanto de revocação, essa métrica é especialmente útil em casos em que o conjunto de teste é não balanceado, sendo portanto uma medida mais confiável do que a acurácia.

$$F1 = \frac{1}{\frac{1}{Precisão} + \frac{1}{Revocação}} \tag{14}$$

5.3 Metodologia

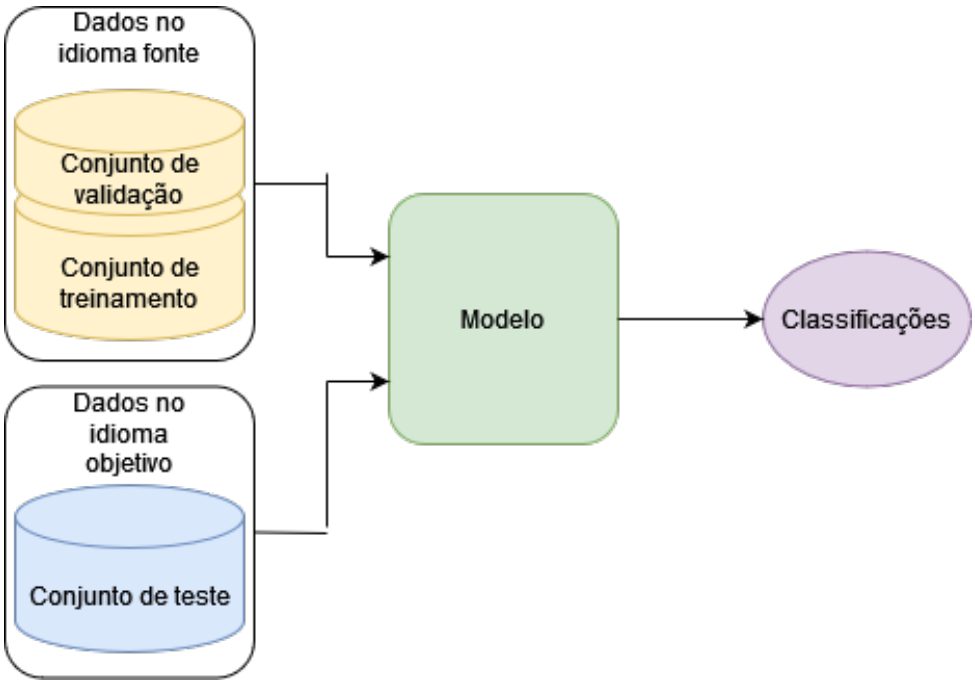
Diversos trabalhos encontrados na literatura buscam solucionar o desafio da detecção automática de discurso de ódio utilizando técnicas que implementam transferência de aprendizado e aqueles que lidam com a língua portuguesa usualmente focam em abordagens multilinguísticas.

Este trabalho propõe a exploração de três técnicas de transferência de aprendizado (*fine-tuning*, *zero-shot learning* e *few-shot learning*) aplicados a três modelos de PLN baseados no BERT, que por sua vez é baseado em um mecanismo de atenção, sendo dois modelos monolinguísticos (Bertimbau Base e Bertimbau Large) e um multilinguístico (mBERT).

Para analisar o impacto da aplicação das técnicas de transferência de aprendizado e de balanceamento das classes dos conjuntos de dados nos desempenhos dos modelos, os seguintes protocolos experimentais foram definidos:

- **Protocolo 1:** Para experimentar a técnica de *zero-shot learning* entre línguas cruzadas, deve-se realizar o processo de *fine-tuning* com os conjuntos de detecção de discurso de ódio em espanhol e inglês (HaterNet e HateXplain). Como conjunto de teste deve-se utilizar os conjuntos de detecção de discurso de ódio em português (Fortuna e HateBR), a fim de se observar se o conhecimento é transferido entre os idiomas. A Figura 6 traz um diagrama que representa este protocolo;
- **Protocolo 2:** Para experimentar a técnica de *few-shot learning* entre línguas cruzadas, deve-se realizar o processo de *fine-tuning* com os conjuntos de detecção de discurso de ódio em espanhol ou inglês (HaterNet e HateXplain) adicionando $k\%$ dos dados em português (Fortuna e HateBR) ao conjunto de treinamento. O conjunto de teste deve ser formado somente por dados em português (Fortuna e HateBR) para que seja possível observar se a adição de algumas amostras do idioma objetivo afeta positivamente o desempenho dos modelos. O valor de k deve variar para permitir avaliar como o número de amostras adicionadas ao conjunto de treinamento afeta o desempenho do modelo. A Figura 7 traz um diagrama que representa este protocolo;
- **Protocolo 3:** Para explorar a técnica de *fine-tuning* isolada os modelos devem ser submetidos ao processo de *fine-tuning* nos conjuntos de dados de detecção de discurso de ódio (Fortuna, HateBR e JoinedHateBR); este protocolo é utilizado para observar o desempenho das diferentes técnicas utilizadas neste trabalho em comparação ao desempenho do *fine-tuning* simples. A Figura 8 traz um diagrama que representa este protocolo;

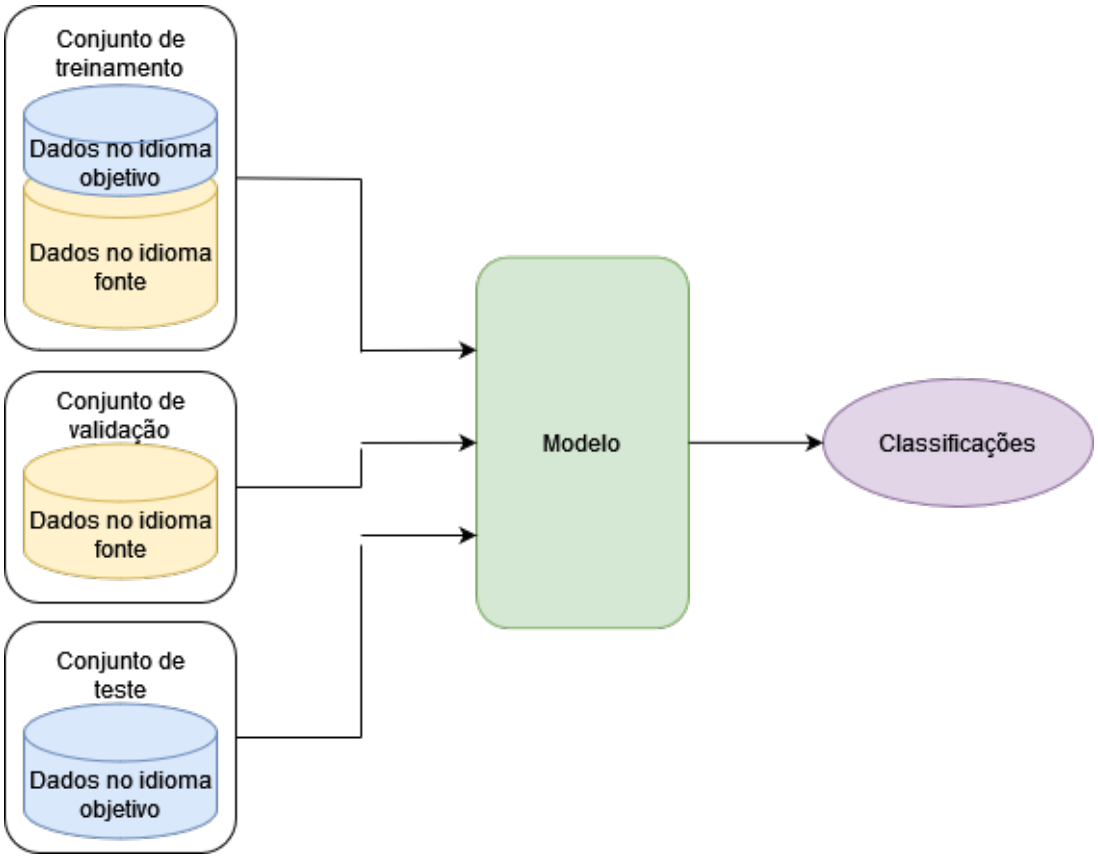
Figura 6 – Diagrama do Protocolo 1 (técnica de *zero-shot learning)*.



Fonte: Elaborada pelo autor.

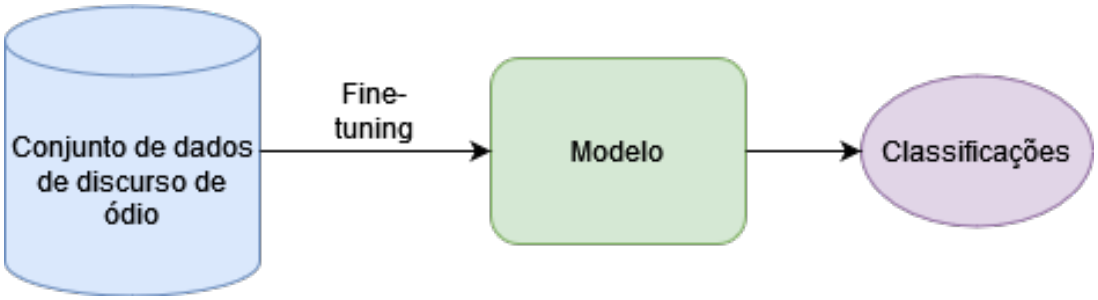
- **Protocolo 4:** Para aplicar a técnica de transferência de aprendizado entre tarefas, foi escolhida a tarefa de detecção de linguagem ofensiva como tarefa fonte devido à sua semelhança à tarefa objetivo de detecção de discurso de ódio. Segundo este protocolo, deve-se realizar, inicialmente, o processo de *fine-tuning* do modelo com os conjuntos de dados da tarefa fonte (ToLDBR e OlidBR) para que ele acumule conhecimento sobre a tarefa de detecção de linguagem ofensiva. Após, deve-se realizar o processo de *fine-tuning* do modelo com os conjuntos de dados da tarefa objetivo (Fortuna e HateBR) para se observar se o conhecimento adquirido é transferido para a tarefa de detecção de discurso de ódio. A Figura 9 traz um diagrama que representa este protocolo;
- **Protocolo 5:** Para evitar a ocorrência de viés, deve-se balancear as classes nos conjuntos de detecção de discurso de ódio em português (Fortuna e HateBR), uma vez que esses conjuntos de dados são bastante desbalanceados. Deve-se, então, aplicar uma técnica de *resampling*, como por exemplo a função de *random undersampling*, aos conjuntos de dados e, após, realizar o processo de *fine-tuning* nos modelos com os dados reamostrados. A Figura 10 traz um diagrama que representa este protocolo.

Figura 7 – Diagrama do Protocolo 2 (técnica de *few-shot learning*).



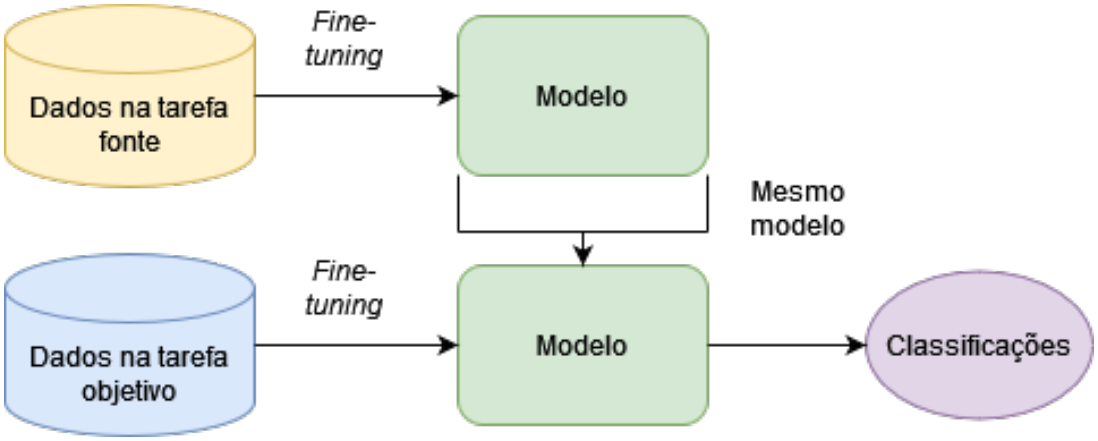
Fonte: Elaborada pelo autor.

Figura 8 – Diagrama do Protocolo 3 (*Fine-tuning* simples).



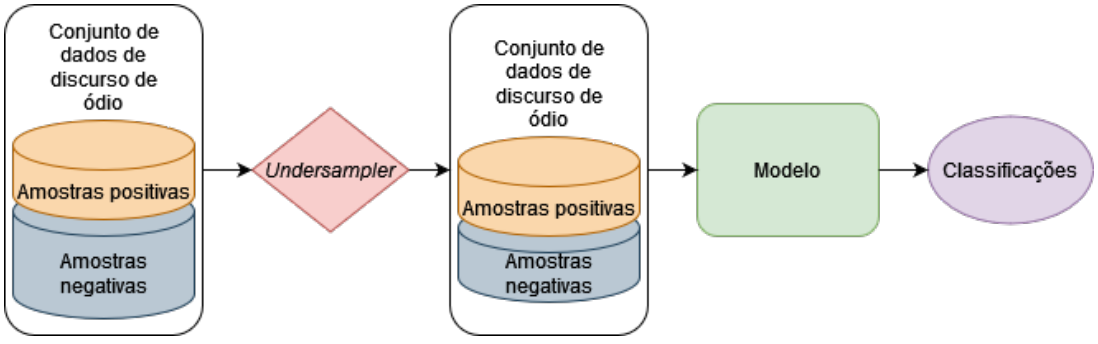
Fonte: Elaborada pelo autor.

Figura 9 – Diagrama do Protocolo 4 (transferência de aprendizado entre tarefas).



Fonte: Elaborada pelo autor.

Figura 10 – Diagrama do Protocolo 5 (reamostragem dos conjuntos de dados para balanceamento das classes).



Fonte: Elaborada pelo autor.

6 Resultados e Discussão

Neste capítulo são descritas as especificações do hardware utilizado nos experimentos, são definidos os valores dos hiperparâmetros dos algoritmos empregados, são detalhados os preprocessamentos realizados e, finalmente, são apresentados e discutidos os resultados obtidos, tanto aqueles referentes ao *benchmark* quanto os referentes aos protocolos definidos na Seção 5.3.

6.1 Especificações do Hardware

Todos os experimentos foram executados no *Google Colab*[6], um serviço que permite o acesso a recursos computacionais pela nuvem. As máquinas utilizadas possuíam 51 GB de RAM de sistema, 15 GB de RAM de GPU e 201 GB de espaço de disco.

6.2 Hiperparâmetros

As implementações utilizadas neste trabalho são aquelas disponibilizadas pela biblioteca *Huggingface*[7]. Os modelos utilizados foram: o BERTimbau, em suas versões básica[8] e grande[9], e o mBERT[10] em sua versão básica.

Os processos de *fine-tuning* foram realizados com uma taxa de aprendizado $\alpha = 2,5 \times 10^{-5}$ com uma queda de peso de $0,01$ em 5 épocas com lotes de tamanho 16.

6.3 Pré-Processamento

Como a tarefa objetivo é a classificação binária de discurso de ódio, em todos os conjuntos de dados todas as colunas, exceto o rótulo binário e texto, foram removidas.

Foi realizada uma etapa simples de pré-processamento dos dados, com o auxílio da biblioteca NLTK[11] e pontuações com pacote de pontuação do *Python*, removendo palavras vazias, ou *stop words* em inglês, que são palavras comuns e que trazem pouca informação além de sua função gramatical como artigos e preposições.

[6] <https://colab.google/>
[7] https://huggingface.co/
[8] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[9] https://huggingface.co/neuralmind/bert-large-portuguese-cased
[10] https://huggingface.co/bert-base-multilingual-cased
[11] https://www.nltk.org/

Para os experimentos que utilizam de *resampling*, foi aplicada a função de *random undersampling* da biblioteca *Imbalanced Learn*[12] nos conjunto de dados de discurso de ódio em português. A Tabela 1 apresenta detalhes sobre os conjuntos de dados antes e pós o *random undersampling*. Para facilitar a compreensão, o conjunto resultante da união dos conjuntos de dados Fortuna e HateBR é referenciado neste trabalho como JoinedHateBR.

Tabela 1 – Número de instâncias e proporções das classes para os conjuntos de dados antes e após o *random undersampling*.

Antes do *random undersampling*				
	Número de instâncias	Instâncias positivas	Instâncias negativas	Proporção (%)
Fortuna Dataset	5.668	1.788	3.880	31,55/68,45
HateBR	7.000	702	6.298	10,02/89,98
JoinedHateBR	12.668	2.490	10.178	19,66/80,34
Após o *random undersampling*				
Fortuna Dataset	3.576	1.788	1.788	50/50
HateBR	1.404	702	702	50/50
JoinedHateBR	4.980	2.490	2.490	50/50

6.4 *Benchmarks*

Para possibilitar a análise dos efeitos das técnicas de transferência de aprendizado no desempenho dos modelos utilizados neste trabalho (Bertimbau Base, Bertimbau Large e mBERT), as Tabelas 2 apresenta *benchmarks* utilizados como referência nos protocolos experimentais definidos.

Os *benchmarks* consistem na aplicação dos modelos sem qualquer processo de *fine-tuning* nos conjuntos de dados em português, tomando estes dados como base é possível avaliar como cada técnica melhora o desempenho dos modelos.

Ressalta-se que neste experimento os conjuntos de dados foram utilizados inteiramente como conjunto de teste (ou seja, não houve subdivisão em subconjuntos de treinamento, validação e teste).

6.5 Resultados do Protocolo 1

Este protocolo experimental foi proposto para permitir analisar a capacidade do modelo multilinguístico mBERT em solucionar a tarefa de detecção de discurso de ódio em português realizando o *fine-tuning* em outros idiomas, como os idiomas inglês e espanhol, ou seja, utilizando a estratégia de transferência de aprendizagem *zero-shot learning*. Para tanto, foram usados os conjuntos de dados HateXplain e HaterNet, para os idiomas inglês e espanhol, respectivamente.

[12] Disponível em: <https://imbalanced-learn.org>

Tabela 2 – Performances dos modelos Bertimbau Base, Bertimbau Large e mBERT sem qualquer *fine-tuning* nos conjuntos de discurso de ódio em português (conjuntos Fortuna, HateBR e JoinedHateBR).

Modelo	Acurácia	Precisão	Revocação	F1
Conjunto de dados Fortuna				
Bertimbau Base	33,61%	0,4673	0,9233	0,3128
Bertimabau Large	36,04%	0,4228	0,7427	0,2955
mBERT	65,80%	0,0909	0,0542	0,2811
Conjunto de dados HateBR				
Bertimbau Base	12,82%	0,1816	0,9644	1002
Bertimabau Large	10,10%	0,1824	0,9347	1003
mBERT	89,22%	0,0105	0,0056	0,0666
Conjunto de dados JoinedhateBR				
Bertimbau Base	57,37%	0,2988	0,4622	0,2208
Bertimabau Large	48,09%	0,2232	0,3796	0,1581
mBERT	30,74%	0,3239	0,8442	0,2004

Para garantir que o modelo mBERT estava aprendendo a solucionar a tarefas nos idiomas fonte foi separado um subconjunto de validação com $12,5\%$ dos dados. A Tabela 3 apresenta o desempenho desse modelo no conjunto de validação nos idiomas inglês e espanhol.

Tabela 3 – Performance do modelo multiliguístico mBERT na tarefa de detecção de discurso de ódio nos idiomas fonte, inglês e espanhol, nos conjuntos de dados HateXplain e HaterNet, respectivamente.

Idioma	Acurácia	F1	Revocação	Precisão
Inglês	84,59%	0,7392	0,7392	0,7393
Espanhol	77,79%	0,5454	0,5409	0,5500

O modelo mBERT, após o processo de *fine-tuning* nos idiomas fontes, inglês e espanhol, foi testado com os três conjuntos de dados inteiros em português: Fortuna, HateBR e JoinedHateBR. Os desempenhos obtidos nesses três conjuntos de dados estão apresentados nas Tabelas 4 e 5, para os idiomas inglês e espanhol, respectivamente.

Tabela 4 – Performance do modelo multiliguístico mBERT, *fine-tuned* em inglês, no conjunto de dados HateXplain, na tarefa de detecção de discurso de ódio nos conjuntos de dados Fortuna, HateBr e JoinedHateBR.

Conjunto	Acurácia	F1	Revocação	Precisão
Fortuna	68,50%	0,0044	0,0022	0,6666
HateBr	89,92%	0,0024	0,0032	0,1800
JoinedHateBR	80,33%	0,0032	0,0016	0,4444

Observa-se nas nas Tabelas 4 e 5 que as métricas apresentam valores baixos, indicando que o modelo multilinguístico mBERT não foi capaz de generalizar o conhecimento entre as línguas, mesmo em línguas semelhantes como o português e o espanhol. Nota-se também que

Tabela 5 – Performance do modelo multiliguístico mBERT, *fine-tuned* em espanhol, no conjunto de dados HaterNet, na tarefa de detecção de discurso de ódio nos conjuntos de dados Fortuna, HateBr e JoinedHateBR.

Conjunto	Acurácia	F1	Revocação	Precisão
Fortuna	69,41%	0,1135	0,0621	0,6607
HateBr	85,40%	0,1590	0,1282	0,2093
JoinedHateBR	78,80%	0,1301	0,0807	0,3361

os modelos em geral desempenham ainda pior do que os benchmarks apresentados na Tabela 2, principalmente o modelo pré-treinado em inglês, reforçando que o processo de *fine-tuning* em outro idioma não permitiu ao modelo acumular conhecimento útil para solução da tarefa.

6.6 Resultados do Protocolo 2

Este protocolo experimental foi proposto para permitir analisar a técnica de transferência de aprendizagem *few-shot learning*. Para isto, diferentemente do Protocolo 1, foi adicionado aos conjuntos de treinamento dos idiomas inglês e espanhol o equivalente a 5%, 10% e 20% dos dados dos conjuntos de dados no idioma português (Fortuna, HateBr e JoinedHateBr), desta maneira os modelos são expostos à língua objetivo durante o processo de *fine-tuning* e é possível avaliar se a quantidade de dados disponibilizados no idioma objetivo no treinamento influencia no seu desempenho.

As Tabelas 6 e 7 apresentam os desempenhos dos modelos. Para validação foi separado do conjunto de treinamento um subconjunto contendo $12,5\%$ dos dados.

Tabela 6 – Performance do modelo *fine-tuned* em inglês com 5%, 10% e 20% das amostras em português nos conjuntos de discurso de ódio em português.

	Validação				Teste			
5% dos dados do treinamento em português								
Conjunto	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Fortuna	84,57%	0,7123	0,6742	0,7548	71,35%	0,4992	0,4479	0,5639
HateBr	85,06%	0,6921	0,6107	0,7985	90,57%	0,1240	0,0674	0,7692
JoinedHateBR	84,83%	0,6913	0,6108	0,7960	80,59%	0,6058	0,3176	0,6518
10% dos dados de treinamento em português								
Fortuna	83,58%	0,7158	0,7234	0,7082	70,32%	0,5598	0,5862	0,5357
HateBr	85,56%	0,7229	0,7063	0,7404	91,13%	0,3241	0,2141	0,6666
JoinedHateBR	84,15%	0,7169	0,7287	0,7055	81,87%	0,3394	0,2379	0,5919
20% dos dados de treinamento em português								
Fortuna	82,96%	0,6990	0,6479	0,7588	73,25%	0,5364	0,4820	0,6048
HateBr	86,10%	0,7407	0,7281	0,7537	92,89%	0,6212	0,5884	0,6578
JoinedHateBR	85,24%	0,7209	0,6651	0,7869	83,43%	0,4800	0,3902	0,6236

Para facilitar a análise dos resultados, os dados apresentados nas Tabelas 6 e 7 foram dispostos nos gráficos apresentados nas Figuras 11, 12 e 13, para a língua inglesa, e nos gráficos apresentados nas Figuras 14, 15 e 16, para a língua espanhola.

Tabela 7 – Performance do modelo *fine-tuned* em espanhol com 5%, 10% e 20% das amostras em português nos conjuntos de discurso de ódio em português.

	Validação				Teste			
5% dos dados do treinamento em português								
Conjunto	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Fortuna	78,20%	0,5618	0,5070	0,6300	70,67%	0,4322	0,3533	0,5565
HateBr	79,49%	0,5628	0,5049	0,6358	89,36%	0,0531	0,0298	0,2439
JoinedHateBR	80,64%	0,5770	0,5024	0,6776	78,60%	0,3154	0,2510	0,4242
10% dos dados de treinamento em português								
Fortuna	78,94%	0,5028	0,4202	0,6259	71,53%	0,3308	0,2207	0,6544
HateBr	80,04%	0,5914	0,6113	0,5728	88,28%	0,3777	0,3573	0,4007
JoinedHateBR	76,37%	0,5480	0,5879	5131	79,32%	0,3030	0,2286	0,4495
20% dos dados de treinamento em português								
Fortuna	75,87%	0,5639	0,6017	0,5305	71,93%	0,4745	0,3969	0,5898
HateBr	81,14%	0,5435	0,5236	0,5650	89,38%	0,3909	0,3461	0,4490
JoinedHateBR	80,92%	0,5277	0,4567	0,6250	80,86%	0,3730	0,2893	0,5249

Figura 11 – Performance do modelo *fine-tuned* em inglês com 5%, 10% e 20% das amostras em português no conjunto Fortuna.

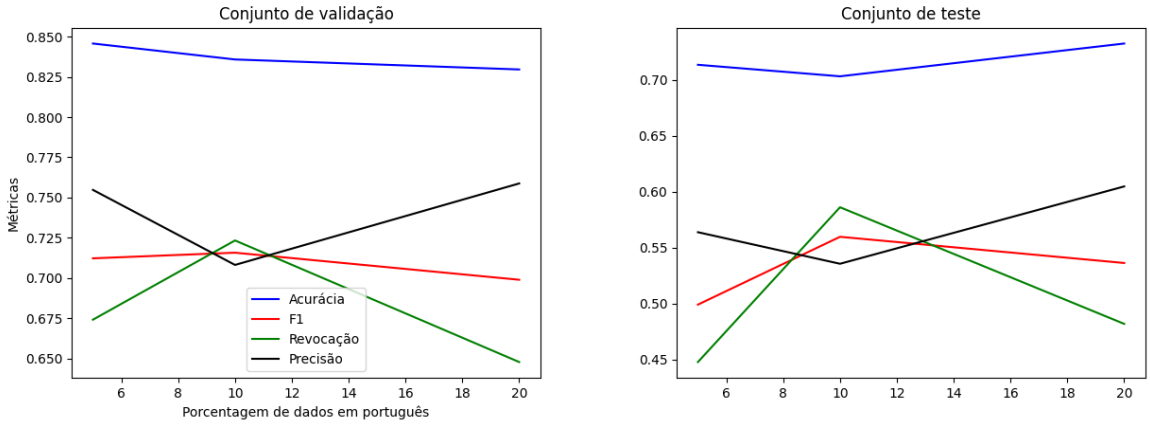


Figura 12 – Performance do modelo *fine-tuned* em inglês com 5%, 10% e 20% das amostras em português no conjunto HateBR.

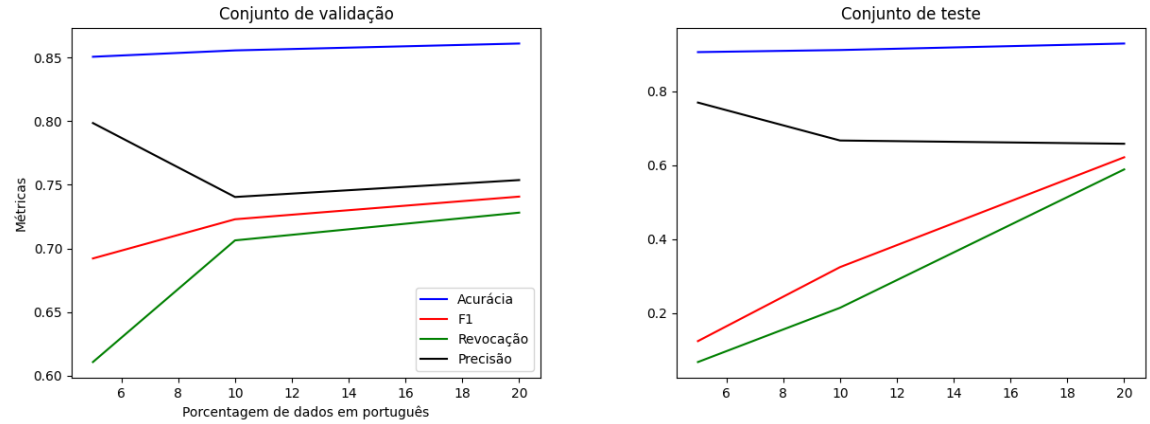


Figura 13 – Performance do modelo *fine-tuned* em inglês com 5%, 10% e 20% das amostras em português no conjunto JoinedHateBr.

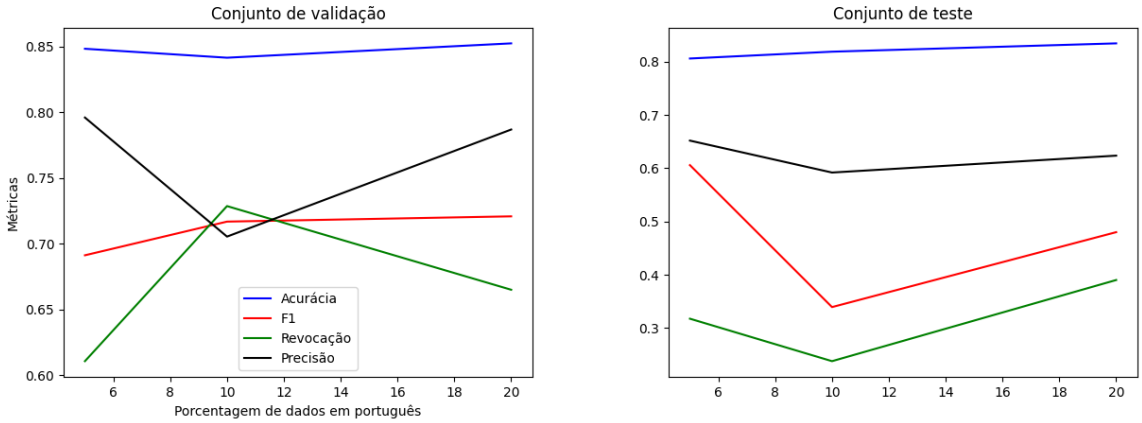
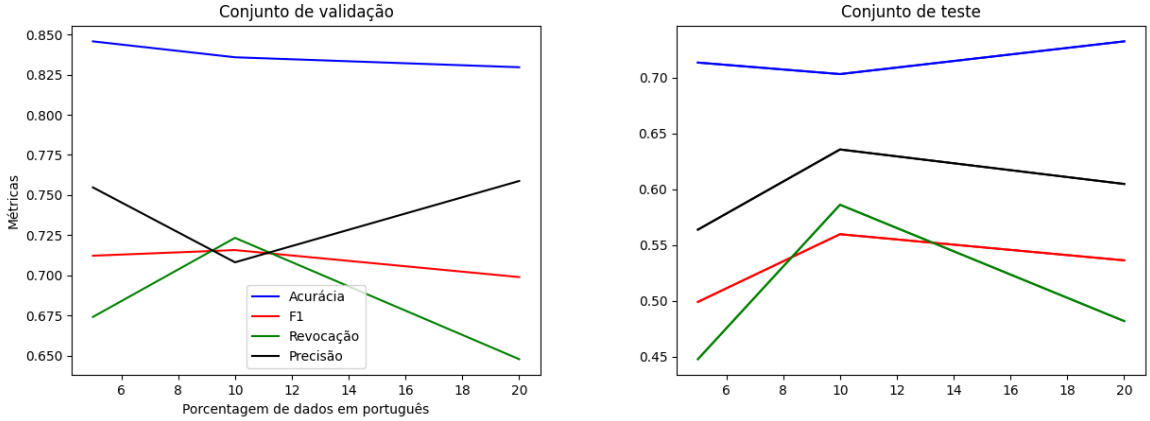


Figura 14 – Performance do modelo *fine-tuned* em espanhol com 5%, 10% e 20% das amostras em português no conjunto Fortuna.



Em sua maioria os modelos treinados com 20% dos dados em português desempenham melhor que os modelos treinados com uma quantidade menor, entretanto essa melhoria é, na maior parte dos casos, pequena. É notável que ainda que o melhor desempenho para o espanhol seja atingido com 20% dos dados em português, para a maior parte das métricas o modelo desempenha melhor com 5% dos dados em português do que 10%, indicando que o aumento no número de dados em português pode não ser proveitoso.

A generalização se dá de maneira mais eficiente entre o inglês e o português, atingindo valores superiores na maior parte das métricas comparado ao espanhol, para ambos os idiomas os resultados foram significativamente melhores do que os obtidos no Protocolo 1 e aos *benchmarks*, porém ainda significativamente baixos, indicando um sucesso na transferência de conhecimento entre os idiomas ainda que não suficientemente bom para atingir o estado da arte.

Figura 15 – Performance do modelo *fine-tuned* em espanhol com 5%, 10% e 20% das amostras em português no conjunto HateBR.

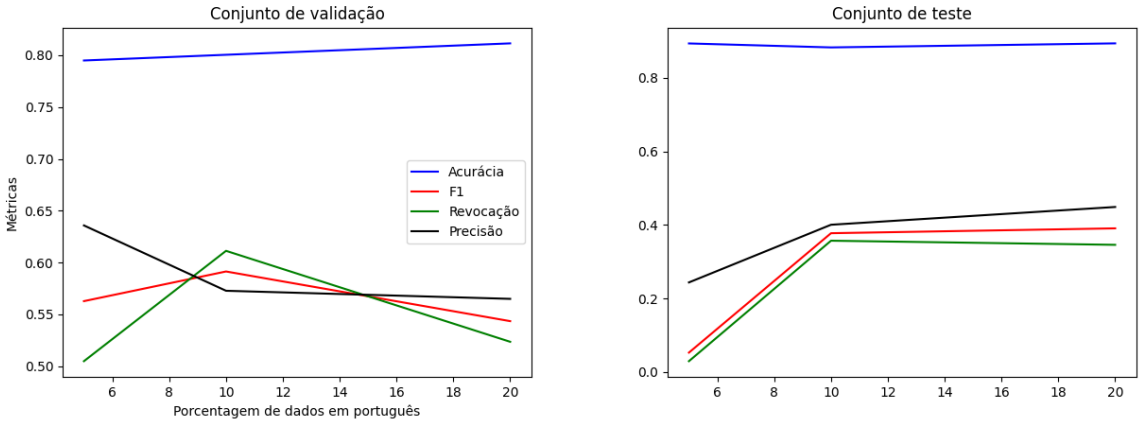
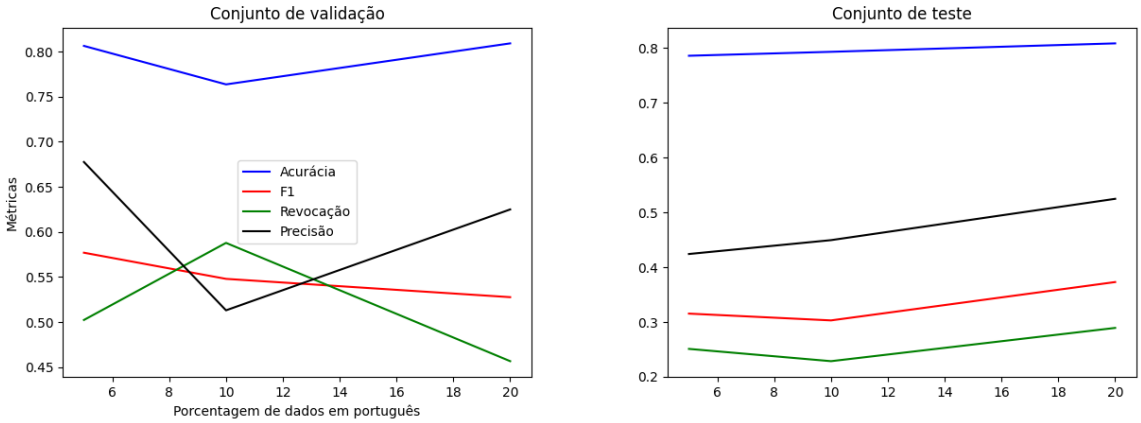


Figura 16 – Performance do modelo *fine-tuned* em espanhol com 5%, 10% e 20% das amostras em português no conjunto JoinedHateBR.



6.7 Resultados do Protocolo 3

A Tabela 8 apresenta os resultados obtidos pelos três modelos após o processo de *fine-tuning* nos três conjuntos de dados de detecção de discurso de ódio em português (Fortuna, HateBR e JoinedHateBR). Os conjuntos de dados foram divididos em 3 subconjuntos, sendo 75% dos dados para os subconjuntos de treinamento, $12,5\%$ para os subconjuntos de validação e $12,5\%$ dos dados para o subconjunto de teste.

Como pode ser observado, os três modelos obtêm performances semelhantes para os três conjuntos de dados. Os modelos monolinguísticos, Bertimbau Base e Bertimbau Large, tendem a obter melhores resultados para as métricas F1, revocação e precisão, que são mais significativas do que a acurácia nos casos em que há desbalanceamento entre as classes, como ocorre nos três conjuntos de dados utilizados neste trabalho.

Tabela 8 – Performances dos modelos Bertimbau Base, Bertimbau Large e mBERT após realização de *fine-tuning* nos subconjuntos de validação e teste de discurso de ódio em português dos conjuntos Fortuna, HateBR e JoinedHateBR.

	Validação				Teste			
Modelo	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Conjunto de dados Fortuna								
Bertimbau Base	74,72%	0,6065	0,6026	0,6106	75,14%	0,5889	0,6160	0,5641
Bertimabau Large	76,27%	0,6199	0,6313	0,6088	77,25%	0,6036	0,5928	0,6147
mBERT	77,11%	0,5668	0,4930	0,6666	75,49%	0,5714	0,5035	0,6603
Conjunto de dados HateBR								
Bertimbau Base	93,94%	0,6624	0,7123	0,6190	93,93%	0,6923	0,6990	0,6857
Bertimabau Large	94,17%	0,6277	0,5584	0,7166	94,02%	0,6960	0,6228	0,7888
mBERT	93,94%	0,7103	0,6701	0,7558	92,41%	0,5846	0,5229	0,6628
Conjunto de dados JoinedHateBR								
Bertimbau Base	85,72%	0,6412	0,6495	0,6332	86,23%	0,6325	0,6580	0,6090
Bertimabau Large	83,19%	0,5476	0,5062	0,5962	86,05%	0,6142	0,6063	0,6224
mBERT	83,82%	0,5328	0,4740	0,6083	84,70%	0,5641	0,5080	0,6342

Os três modelos atingiram os melhores resultados quando aplicados ao conjunto HateBR. Porém, como pode ser observado na Tabela 1, este é o conjunto de dados mais desbalanceado utilizado nos experimentos, sendo que apenas 10% das instâncias referem-se à discurso de ódio, tornando os resultados menos confiáveis.

É possível também observar que os modelos atingem resultados superiores aos apresentados nos protocolos 1 e 2, reforçando a observação de que os modelos têm dificuldade em generalizar o conhecimento entre os idiomas ao atingir resultados superiores realizando *fine-tuning* somente em português.

6.8 Resultados do Protocolo 4

Este protocolo experimental foi proposto para permitir analisar se o uso de uma tarefa semelhante à tarefa de detecção de discurso de ódio pode auxiliar na sua solução. Inicialmente, os modelos Bertimbau Base, Bertimbau Large e mBERT foram submetidos ao *fine-tuning* com os conjuntos da tarefa fonte (detecção de linguagem ofensiva), ToLDBR e OlidBR. Estes dois conjuntos de dados foram concatenados e o conjunto resultante desta concatenação foi posteriormente dividido em subconjuntos de treinamento, com 75% dos dados, validação, com $12,5\%$, e teste, também com $12,5\%$. A Tabela 9 apresenta as performances obtidas pelos três modelos na tarefa fonte (detecção de linguagem ofensiva).

Após, foi realizado o *fine-tuning* nos modelos para a tarefa objetivo (detecção de discurso de ódio) com os dados sendo separados na mesma proporção da tarefa fonte. A Tabela 10 apresenta os desempenhos obtidos pelos três modelos. É possivel observar que há pouca variação nos valores das métricas ao comparar com a Tabela 8, indicando que o conhecimento obtido solucionando a tarefa fonte não contribuiu significativamente para a solução da tarefa objetivo, ou que os modelos foram incapazes de generalizar suficientemente o conhecimento.

Tabela 9 – Performance dos modelos Bertimbau Base, Bertimbau Large e mBERT após realização de *fine-tuning* na tarefa fonte de detecção de linguagem ofensiva em português, no conjunto de dados resultante da concatenação dos conjuntos ToLDBR e OlidBR.

	Validação				Teste			
Modelo	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Bertimbau Base	78,30%	0,7744	0,8126	0,7396	73,14%	0,7462	0,8237	0,6820
Bertimabau Large	80,65%	0,7894	0,8211	0,7600	78,72%	0,7774	0,8203	0,7388
mBERT	79,02%	0,7760	0,7972	0,7541	79,58%	0,7685	0,8062	0,7342

Tabela 10 – Performance dos modelos Bertimbau Base, Bertimbau Large e mBERT após realização de *fine-tuning* na tarefa objetivo de detecção de discurso de ódio em português, nos conjuntos de dados Fortuna, HateBR e JoinedHateBR.

	Validação				Teste			
Modelo	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Conjunto de dados Fortuna								
Bertimbau Base	76,00%	0,6352	0,6218	0,6491	74,20%	0,5964	0,5922	0,6007
Bertimabau Large	75,85%	0,6174	0,6188	0,6161	76,10%	0,6424	0,6802	0,6085
mBERT	73,58%	0,6128	0,6245	0,6016	72,87%	0,5677	0,5719	0,5636
Conjunto de dados HateBR								
Bertimbau Base	93,25%	0,6508	0,6962	0,6111	92,94%	0,6781	0,6695	0,6869
Bertimabau Large	93,14%	0,6552	0,7037	0,6129	94,60%	0,7135	0,7802	0,6574
mBERT	93,71%	0,6153	0,5500	0,6984	93,17%	0,5714	0,5000	0,6666
Conjunto de dados JoinedHateBR								
Bertimbau Base	87,93%	0,6232	0,5583	0,7053	85,51%	0,5959	0,5580	0,6392
Bertimabau Large	86,03%	0,6018	0,5566	0,6549	84,68%	0,5559	0,5014	0,6549
mBERT	83,07%	0,5889	0,5783	0,6000	82,28%	0,5795	0,5670	0,5925

6.9 Resultados do Protocolo 5

Este protocolo experimental busca investigar se o grande desequilíbrio entre as classes nos conjuntos de detecção de discurso de ódio em português podem ser uma barreira no desempenho dos modelos.

Neste protocolo, os modelos realizam *fine-tuning* com os dados após o processo de *random undersampling* com os dados divididos na mesma proporção de 75% para treinamento, $12,5\%$ para validação e $12,5\%$ para teste.

Comparando-se a Tabela 11 com a Tabela 8 é possível constatar um aumento em todas as métricas, exceto na acurácia, que considerando o desequilíbrio de classes é uma métrica ineficaz.

Os modelos ainda atingem a melhor performance no conjunto de dados HateBR, entretanto este conjunto conta com apenas 1.404 instâncias, sendo portanto consideravelmente pequeno, levando novamente a resultados menos confiáveis. Considerando este fator é possível observar que o desempenho dos modelos utilizando o conjunto JoinedHateBR, tanto na Tabela 8 quanto na Tabela 11, é superior para o conjunto de dados Fortuna, podendo indicar que um aumento no número de dados pode ser positivo.

Tabela 11 – Performance dos modelos após realização de *fine-tuning* nos conjuntos de discurso de ódio em português após *undersampling*.

	Precisão				Teste			
Modelo	Acurácia	F1	Revocação	Precisão	Acurácia	F1	Revocação	Precisão
Conjunto de dados Fortuna								
Bertimbau Base	70,91%	0,6962	0,6803	0,7129	72,74%	0,7132	0,6814	0,7479
Bertimabau Large	72,93%	0,7126	0,7177	0,7075	74,90%	0,7517	0,7359	0,7683
mBERT	73,60%	0,7561	0,7625	0,7500	70,87%	0,7060	0,7432	0,6725
Conjunto de dados HateBR								
Bertimbau Base	87,42%	0,8450	9090	0,7895	80,37%	0,8055	0,8446	0,7699
Bertimabau Large	85,14%	0,8522	8426	0,8620	89,90%	0,9059	0,9060	0,9060
mBERT	87,43%	0,8854	8947	0,8762	81,81%	0,8257	0,8911	0,7692
Conjunto de dados JoinedHateBR								
Bertimbau Base	75,08%	0,7543	7301	0,7803	76,11%	0,7592	0,7904	0,7304
Bertimabau Large	77,65%	0,7831	7943	0,7723	79,20%	0,7958	0,7968	0,7948
mBERT	72,50%	0,7323	7597	0,7069	74,32%	0,7343	0,7521	0,7174

7 Conclusão

Este trabalho teve como principal objetivo a exploração de técnicas de transferência de aprendizado para solução do problema de detecção automática de discurso de ódio em português, um idioma que sofre com a escassez de dados rotulados e trabalhos, e, adicionalmente, após observar a existência de desbalanceamento das classes nos conjuntos de dados utilizados nesta área, foi decidido explorar uma técnica de reamostragem.

Foram exploradas técnicas de *fine-tuning*, como transferência de conhecimento entre tarefas, *zero-shot* e *few-shot learning* como transferência de aprendizado entre línguas cruzadas. Explorando a transferência de conhecimento entre tarefas observou-se que os modelos explorados não foram capazes de generalizar o conhecimento de maneira favorável ao desempenho, atingindo uma performance muito semelhante a dos modelos não expostos à tarefa fonte, indicando que os modelos foram incapazes de utilizar o conhecimento acumulado na tarefa fonte para solicionar a tarefa objetivo.

Explorando *zero-shot learning* foi possível observar que o modelo multilinguístico foi incapaz de generalizar eficientemente o conhecimento entre os idiomas utilizados, levando os modelos a classificações aleatórias e atingindo em diversas métricas valores inferiores a $0,1$, porém o modelo que passou pelo processo de *fine-tuning* em espanhol atingiu melhores resultados se comparado ao que passou pelo processo em inglês.

A técnica de *few-shot learning* se mostrou mais eficiente que *zero-shot learning*, com um desempenho melhor do modelo treinado com textos em inglês, entretanto ainda desempenhando pior do que os *benchmarks apresentados*. Foi possível também observar que os modelos com uma maior exposição a português durante o treinamento obtiveram o melhor desempenho, porém não é possível garantir que o aumento na quantidade de dados no idioma objetivo no conjunto de treinamento será positiva uma vez que o modelo exposto a apenas 5% dos dados em português e o restante em espanhol desempenhou melhor que o exposto a 10%.

Finalmente, buscando reduzir os efeitos que o desbalanceamento das classes ocasionam no desempenho dos modelos, foi utilizada a técnica de *undersampling*, que se mostrou eficiente, obtendo resultados superiores para F1, revocação e precisão comparada aos *benchmarks* e com apenas uma pequena perda de acurácia, demonstrando que o desbalanceamento das classes pode ser um desafio para os modelos e os levando a criação de viés.

7.1 Trabalhos Futuros

Como trabalhos futuros, considerando a melhoria no desempenho dos modelos utilizando *undersampling*, podem ser explorados diferentes métodos de reamostragem no intuito de buscar

o que melhor se aplica ao problema. Podem ser também explorados diferentes modelos multilinguísticos para observar se algum apresenta melhor desempenho na generalização de conhecimento entre idiomas, assim como a exploração de diferentes combinações de idiomas, da mesma maneira que é possível explorar se outras tarefas agregam conhecimento mais útil a detecção de discurso de ódio do que a de detecção de linguagem ofensiva.

É possível também, considerando a popularização de modelos generativos, observar se estes se aproveitam melhor das estratégias de transferência de aprendizado e atingem performances superiores às apresentadas neste trabalho.

7.2 Trabalhos Publicados

O artigo intitulado *"Hate Speech Detection in Portuguese Using BERTimbau"* foi aceito para apresentação e publicação nos anais do 27th Iberoamerican Congress on Pattern Recognition, CIARP 2024 (CIARP 2024), Talca, Chile.

Referências

ALI, R.; FAROOQ, U.; ARSHAD, U.; SHAHZAD, W.; BEG, M. O. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, Elsevier, v. 74, p. 101365, 2022.

ALURU, S. S.; MATHEW, B.; SAHA, P.; MUKHERJEE, A. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

ARCO, F. M. Plaza-del; MOLINA-GONZÁLEZ, M. D.; URENA-LÓPEZ, L. A.; MARTÍN-VALDIVIA, M. T. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, Elsevier, v. 166, p. 114120, 2021.

ASSIS, G.; AMORIM, A.; CARVALHO, J.; OLIVEIRA, D. de; VIANNA, D.; PAES, A. Exploring portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. [S.l.: s.n.], 2024. p. 301–311.

BADJATIYA, P.; GUPTA, S.; GUPTA, M.; VARMA, V. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th international conference on World Wide Web companion*. [S.l.: s.n.], 2017. p. 759–760.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DJURIC, N.; ZHOU, J.; MORRIS, R.; GRBOVIC, M.; RADOSAVLJEVIC, V.; BHAMIDIPATI, N. Hate speech detection with comment embeddings. In: *Proceedings of the 24th international conference on world wide web*. [S.l.: s.n.], 2015. p. 29–30.

FIRMINO, A. A.; BAPTISTA, C. de S.; PAIVA, A. C. de. Improving hate speech detection using cross-lingual learning. *Expert Systems with Applications*, Elsevier, v. 235, p. 121115, 2024.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 4, p. 1–30, 2018.

FORTUNA, P.; SILVA, J. R. da; WANNER, L.; NUNES, S. et al. A hierarchically-labeled portuguese hate speech dataset. In: *Proceedings of the third workshop on abusive language online*. [S.l.: s.n.], 2019. p. 94–104.

FRENDA, S.; GHANEM, B.; GÓMEZ, M. Montes-y; ROSSO, P. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, IOS Press, v. 36, n. 5, p. 4743–4752, 2019.

FRENKEL, S.; CONGER, K. *Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find*. 2022. Available on: <https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html?fbclid=PAAabqeEixPRfe0gbCUXuQJG0dAOpZyQPmH7-mFDWxyVaHGnHcaMDowyTRSXk>.

HAIXIANG, G.; YIJING, L.; SHANG, J.; MINGYUN, G.; YUANYUE, H.; BING, G. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, Elsevier, v. 73, p. 220–239, 2017.

HINDUJA, S.; PATCHIN, J. W. Bullying, cyberbullying, and suicide. *Archives of suicide research*, Taylor & Francis, v. 14, n. 3, p. 206–221, 2010.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

HUANG, F.; KWAK, H.; AN, J. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*, 2023.

JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, Elsevier, p. 126232, 2023.

JURAFSKY, D. *Speech & language processing*. [S.l.]: Pearson Education India, 2000.

KAUR, H.; PANNU, H. S.; MALHI, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 4, p. 1–36, 2019.

KEMP, S. *Digital 2022: Global Overview Report*. 2022. Available on: <https://datareportal.com/reports/digital-2022-global-overview-report>.

KHAN, S.; FAZIL, M.; SEJWAL, V. K.; ALSHARA, M. A.; ALOTAIBI, R. M.; KAMAL, A.; BAIG, A. R. Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 34, n. 7, p. 4335–4344, 2022.

LEITE, J. A.; SILVA, D. F.; BONTCHEVA, K.; SCARTON, C. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*, 2020.

LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. [S.l.: s.n.], 2014. p. 302–308.

LIU, Y.; HAN, T.; MA, S.; ZHANG, J.; YANG, Y.; TIAN, J.; HE, H.; LI, A.; HE, M.; LIU, Z. et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.

MANNING, C.; SCHUTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999.

MATHEW, B.; SAHA, P.; YIMAM, S. M.; BIEMANN, C.; GOYAL, P.; MUKHERJEE, A. Hatexplain: A benchmark dataset for explainable hate speech detection. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2021. v. 35, n. 17, p. 14867–14875.

MERCHANT, A.; RAHIMTOROGHI, E.; PAVLICK, E.; TENNEY, I. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. A bert-based transfer learning approach for hate speech detection in online social media. In: SPRINGER. *International Conference on Complex Networks and Their Applications*. [S.l.], 2019. p. 928–940.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 8, p. e0237861, 2020.

NIU, Z.; ZHONG, G.; YU, H. A review on the attention mechanism of deep learning. *Neurocomputing*, Elsevier, v. 452, p. 48–62, 2021.

OBERMAIER, M.; SCHMUCK, D.; SALEEM, M. I'll be there for you? effects of islamophobic online hate speech and counter speech on muslim in-group bystanders' intention to intervene. *new media & society*, Sage Publications Sage UK: London, England, p. 14614448211017527, 2021.

PAMUNGKAS, E. W.; BASILE, V.; PATTI, V. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, Elsevier, v. 58, n. 4, p. 102544, 2021.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2009.

PEREIRA-KOHATSU, J. C.; QUIJANO-SÁNCHEZ, L.; LIBERATORE, F.; CAMACHO-COLLADOS, M. Detecting and monitoring hate speech in twitter. *Sensors*, MDPI, v. 19, n. 21, p. 4654, 2019.

PIRES, T.; SCHLINGER, E.; GARRETTE, D. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

POURPANAH, F.; ABDAR, M.; LUO, Y.; ZHOU, X.; WANG, R.; LIM, C. P.; WANG, X.-Z.; WU, Q. J. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, 2022.

RANA, A.; JHA, S. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*, 2022.

ROSS, B.; RIST, M.; CARBONELL, G.; CABRERA, B.; KUROWSKY, N.; WOJATZKI, M. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

SILVA, A.; ROMAN, N. Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression. In: SBC. *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2020. p. 1–12.

SILVA, R. C. C. da; ROSA, T. C. Combining data transformation and classification approaches for hate speech detection: A comparative study. *Thierson, Combining Data Transformation and Classification Approaches for Hate Speech Detection: A Comparative Study*, 2023.

SOHN, H.; LEE, H. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In: IEEE. *2019 International Conference on Data Mining Workshops (ICDMW)*. [S.l.], 2019. p. 551–559.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian conference on intelligent systems*. [S.l.], 2020. p. 403–417.

TRAJANO, D.; BORDINI, R. H.; VIEIRA, R. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, Springer, p. 1–27, 2023.

UNIDAS, O. das N. *EU Code of Conduct against online hate speech: latest evaluation shows slowdown in progress*. 2022. Available on: <https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7109>.

VARGAS, F. A.; CARVALHO, I.; GÓES, F. R. de; BENEVENUTO, F.; PARDO, T. A. S. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*, 2021.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

VIGNA, F. D.; CIMINO, A.; DELL'ORLETTA, F.; PETROCCHI, M.; TESCONI, M. Hate me, hate me not: Hate speech detection on facebook. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. [S.l.: s.n.], 2017. p. 86–95.

WANG, B.; LIU, K.; ZHAO, J. Inner attention based recurrent neural networks for answer selection. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2016. p. 1288–1297.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. *Journal of Big data*, SpringerOpen, v. 3, n. 1, p. 1–40, 2016.

WILLIAMS, M. Hatred behind the screens: A report on the rise of online hate speech. Mishcon de Reya, 2019.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? *Advances in neural information processing systems*, v. 27, 2014.

YUAN, S.; WU, X.; XIANG, Y. A two phase deep learning model for identifying discrimination from tweets. In: *EDBT*. [S.l.: s.n.], 2016. p. 696–697.

ZANNETTOU, S.; BRADLYN, B.; CRISTOFARO, E. D.; KWAK, H.; SIRIVIANOS, M.; STRINGINI, G.; BLACKBURN, J. What is gab: A bastion of free speech or an alt-right echo chamber. In: *Companion Proceedings of the The Web Conference 2018*. [S.l.: s.n.], 2018. p. 1007–1014.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H.; HE, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, IEEE, v. 109, n. 1, p. 43–76, 2020.