Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Detecção de Discurso de Ódio em Comentários Relacionados à Política

Aillkeen Bezerra de Oliveira

Campina Grande, Paraíba, Brasil

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Detecção de Discurso de Ódio em Comentários Relacionados à Política

Aillkeen Bezerra de Oliveira

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Processamento de Linguagem Natural

Cláudio de Souza Baptista, Ph.D
(Orientador)

Campina Grande, Paraíba, Brasil

O48d Oliveira, Aillkeen Bezerra de.
Detecção de discurso de ódio em comentários relacionados à política / Aillkeen Bezerra de Oliveira. – Campina Grande, 2024.
142 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Cláudio de Souza Baptista".
Referências.

1. Processamento de Linguagem Natural. 2. Detecção de Discurso de Ódio. 3. Cross-Lingual Learning. 4. Redes Sociais. I. Baptista, Cláudio de Souza. II. Título.

CDU 004.438:81'322.2(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secretaria-copin@computacao.ufcg.edu.br / copin@copin.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

AILLKEEN BEZERRA DE OLIVEIRA

DETECÇÃO DE DISCURSO DE ÓDIO EM COMENTÁRIOS RELACIONADOS À POLÍTICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 10/05/2024

Prof. Dr. CLÁUDIO DE SOUZA BAPTISTA, UFCG, Orientador

Prof. Dr. HERMAN MARTINS GOMES, UFCG, Examinador Interno

Prof. Dr. LUCIANO DE ANDRADE BARBOSA, UFPE, Examinador Externo

Documento assinado eletronicamente por **CLAUDIO DE SOUZA BAPTISTA**, **PROFESSOR 3 GRAU**, em 17/05/2024, às 11:24, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **HERMAN MARTINS GOMES**, **PROFESSOR 3 GRAU**, em 17/05/2024, às 18:34, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Luciano de Andrade Barbosa**, **Usuário Externo**, em 19/05/2024, às 20:23, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4417696** e o código CRC **F7F2F35F**.

Resumo

Em uma era em que as pessoas estão cada vez mais conectadas, a dispersão de discurso de ódio em redes sociais tornou-se mais frequente. Para contornar esse problema, a tecnologia computacional emergiu como uma ferramenta valiosa para identificar e mitigar discurso de ódio em redes sociais. Diante do poder computacional disponível, este trabalho contempla o uso de Processamento de Linguagem Natural para detectar discurso de ódio em textos provenientes de redes sociais. Além de abordar a detecção, outro objetivo é investigar o impacto da distância léxica entre os idiomas dos corpora empregados no treinamento dos modelos, explorando codificadores e decodificadores baseados na arquitetura de *Transformers*. Portanto, realizou-se uma investigação sobre a inclusão de *Cross-lingual Learning* (CLL) para aprimorar a detecção de discurso de ódio em diferentes idiomas, empregando diversas técnicas de CLL, bem como a aplicação de múltiplos idiomas como fonte de treino para o modelo. Os resultados revelaram que a aplicação de CLL, especialmente com múltiplos idiomas fonte, melhorou significativamente a eficácia desses modelos na classificação de discurso de ódio. Os modelos baseados em codificadores mostraram-se mais eficientes quando a distância léxica entre os idiomas era mais próxima, alcançando 96,92% na métrica F1-score. Em contraste, os modelos decodificadores mostraram-se mais eficientes quando a distância léxica entre os idiomas era mais distante, alcançando 96,58% na métrica F1-score. Sendo assim, esta dissertação destaca que a diversidade linguística e a consideração da distância léxica em modelos fundamentados em *Transformers* são cruciais para o desenvolvimento de sistemas eficazes para detectar discurso de ódio. Por fim, os achados desta pesquisa reforçam a viabilidade de utilizar CLL e múltiplos idiomas para aprimorar a detecção de discurso de ódio, oferecendo novas direções e percepções para pesquisas futuras nessa área.

Abstract

In an era where people are increasingly connected, the spread of hate speech on social networks has become more frequent. Consequently, computational technology has emerged as a valuable tool to identify and mitigate hate speech on these platforms. Given the available computational power, we used Natural Language Processing to detect hate speech in texts from social networks. Besides addressing detection, another goal was to investigate the impact of lexical distance between the languages of the corpora used in model training, exploring encoders and decoders based on Transformer architecture. Therefore, we investigated the inclusion of Cross-Lingual Learning (CLL) to enhance hate speech detection in different languages, employing various CLL techniques and the application of multiple languages as training sources for the model. The results revealed that applying CLL, especially with multiple source languages, significantly improved the effectiveness of the models in classifying hate speech. Moreover, encoder-based models were more efficient when the lexical distance between languages was closer, achieving 96.92% in the F1-score metric. In contrast, decoder models were more efficient when the lexical distance between languages was farther, achieving 96.58% in the F1-score metric. Thus, this work highlights that linguistic diversity and the lexical distance used in Transformer-based models are crucial for developing effective systems to detect hate speech. Finally, the findings of this research reinforce the feasibility of using CLL and multiple languages to improve hate speech detection, offering new directions and insights for future research in this area.

Agradecimentos

Em primeiro lugar, gostaria de expressar minha gratidão a Deus por me proporcionar esta oportunidade, assim como por me conceder forças e sabedoria para enfrentar os desafios.

Também desejo expressar minha profunda gratidão à minha família, que esteve ao meu lado em todos os momentos, oferecendo apoio, incentivo e força.

Agradeço aos meus amigos pelo apoio, paciência e conselhos.

Agradeço ao meu orientador, Cláudio de Souza Baptista, pelas oportunidades, assim como por sua orientação, sabedoria, confiança e incentivo. Suas orientações e conselhos foram fundamentais para o meu crescimento.

Agradeço aos professores Anderson Firmino e Anselmo Cardoso de Paiva. Embora não tenham atuado como co-orientadores, os conselhos prestados foram de muita valia.

Agradeço aos integrantes do Laboratório de Sistemas de Informação (LSI) que compartilharam seus conhecimentos e amizade comigo ao longo dos anos.

Enfim, agradeço a todos que me acompanharam e apoiaram. Cada gesto de amizade, cada palavra de encorajamento e cada ato de apoio foram inestimáveis para mim.

Conteúdo

Lista de Símbolos

PLN - *Processamento de Linguagem Natural*

AM - *Aprendizado de Máquina*

CLL - *Cross-Lingual Learning*

IA - *Inteligência Artificial*

CIF - *Corpora do Idioma Fonte*

CID - *Corpus do Idioma de Destino*

BERT - *Bidirectional Encoder Representations from Transformers*

ZST - *Zero-shot Transfer*

CL - *Cascade Learning*

JL - *Joint Learning*

ASO - *Almost Stochastic Order*

GPT - *Generative Pre-trained Transformer*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Ao longo dos anos, a humanidade alcançou significativos avanços na tecnologia de comunicação, percorrendo desde a era do rádio e televisão até as últimas décadas, que foram marcadas pela ascensão da Internet. A integração da Internet com dispositivos móveis, como tablets, celulares e smartphones, desempenhou um papel fundamental ao possibilitar a transmissão instantânea de informações e, por conseguinte, tornar a comunicação entre as pessoas mais rápida.

Boa parte da utilização desses dispositivos, atualmente, está direcionada a atividades sociais [19]. O interesse das pessoas nessas atividades, juntamente com a disponibilidade de comunicação instantânea, motivaram as empresas a estabelecerem extensas redes sociais, facilitando a troca de opiniões entre os usuários. Redes sociais representam uma estrutura composta por indivíduos conectados por um ou mais interesses, compartilhando valores e objetivos em comum. Sendo assim, o interesse das pessoas em expressar suas opiniões em plataformas sociais tem aumentado consideravelmente com o tempo [59].

Por meio dessas plataformas de rede social, como o Facebook[1], Instagram[2], rede social X[3], entre outras, a população usufrui da oportunidade de se expressar, podendo compartilhar suas opiniões, insatisfações, momentos de felicidade, etc. Esses compartilhamentos frequentemente ocorrem por meio de textos abertos ao público ou direcionados a indivíduos específicos, possibilitando que qualquer pessoa visualize e participe das discussões.

[1] https://www.facebook.com/

[2] https://www.instagram.com/

[3] https://twitter.com/

Contudo, essa prerrogativa de liberdade de expressão também é explorada para a disseminação de hostilidade em redes sociais. Dessa forma, os indivíduos perpetram ataques sob a forma de Agressão Cibernética [62; 103]. Esse comportamento dá origem ao que é conhecido como Discurso de Ódio, uma expressão que promove o aumento da violência e incita ataques direcionados a grupos específicos de pessoas [34]. Em sua maioria, tais ataques visam indivíduos que se enquadram em determinados critérios, tais como ascendência, nacionalidade, origem étnica, gênero, orientação política, entre outros. A Tabela 1.1 mostra alguns exemplos de frases com e sem discurso de ódio.

Tabela 1.1: Exemplos de frases com e sem discurso de ódio extraídos do corpus elaborado nesta dissertação (Capítulo 4).

Sem Discurso de Ódio	Com Discurso de Ódio
"Respeitem a opinião politica de outra pessoa, ninguém é obrigado a votar no seu candidato."	"Quantas vidas se perderam por culpa desse incompetente, genocida."
"Ser cristão é sobre amor e respeito ao próximo."	"Presidente burro! Foi na globo e levou enrabada."
"Contra fatos não há argumentos, já sabemos que o PT que irá vencer!"	"O presidente não sabe falar. Quem votou nisso, vai pra puta que pariu, cara."
"Vamos defender a democracia! Vamos nos unir!"	"Esse é o meu presidente e vão todos para a puta que pariu."

Com o passar do tempo, a Agressão Cibernética emergiu como um tema de considerável interesse e investigação no campo da Ciência da Computação [62]. A disseminação de discursos de ódio presentes nas plataformas online pode ter impactos psicológicos adversos nas pessoas que se tornam alvos desses ataques, potencialmente resultando em condições mais graves, como depressão e transtornos de ansiedade [68; 100]. Adicionalmente, os comportamentos agressivos praticados por indivíduos têm o potencial de incitar outros a replicarem tais comportamentos, propagando negativamente essas ações e afetando mais membros da sociedade.

Além disso, esses comportamentos podem estar correlacionados às práticas criminosas

[64]. Portanto, esse tema é de grande interesse e objeto de investigação para entidades jurídicas e organizações, visando a punição dos transgressores. O tratamento efetivo desse tipo de problema é de suma importância, pois a redução de casos de agressão cibernética implica também na diminuição dos índices de violência e criminalidade, contribuindo para a melhoria do bem-estar das pessoas que são alvo desses ataques.

Entretanto, mitigar esse problema não é tão fácil, dada a vasta quantidade de mensagens enviadas diariamente em redes sociais. A rede social X, por exemplo, contempla uma alta incidência de mensagens relacionadas a discurso de ódio [46]. Entre os anos 2022 e 2023, aproximadamente 500 milhões de mensagens foram publicadas diariamente nessa rede social [6; 83; 2], o que equivale a aproximadamente 350 mil postagens por minuto. Apesar disso, a plataforma depende da colaboração dos seus usuários para identificar comentários agressivos [46], a fim de que tais comentários sejam removidos ou tratados. O monitoramento, a exclusão ou a restrição manual de tais mensagens em redes sociais se revelam tarefas extremamente exaustivas e onerosas para as empresas responsáveis por essas plataformas sociais.

Considerando a falta de controle e a inviabilidade de monitoramento por recursos humanos, técnicas computacionais apresentam-se como soluções para agilizar, reduzir custos e automatizar a identificação de tais problemas. O Processamento de Linguagem Natural [12] e o Aprendizado de Máquina [4] são abordagens computacionais aplicáveis para detectar e realizar o controle de discurso de ódio [1; 57; 66]. Além dessas dificuldades, o discurso de ódio apresenta outro desafio como a complexidade linguística, que apresenta ambiguidade com outras áreas (sarcasmo e ironia) e é altamente dependente do contexto [42; 80], dificultando a detecção de discurso de ódio por meio de modelos computacionais. Adicionalmente, outros desafios são a escassez de dados e as mudanças de expressões devido às transformações culturais na sociedade [42]. Isso dificulta a criação de modelos cada vez mais precisos, pois, à medida que novas gerações surgem, outras expressões de discurso de ódio aparecem devido à criação de novas formas de comunicação na sociedade (gírias e palavras), necessitando de coleta e rotulação de novos dados para o treinamento de novos modelos.

Sendo assim, diante do significativo potencial da computação para mitigar esse problema social, este trabalho propõe utilizar ferramentas computacionais estado-da-arte para detectar discurso de ódio presente em comentários nas redes sociais.

1.1 Objetivos

Esta seção contempla os objetivos gerais e específicos desta pesquisa.

1.1.1 Objetivos Gerais

Esta pesquisa tem como motivação o combate de discurso de ódio em redes sociais, bem como a mitigação dos problemas presentes na detecção de discurso de ódio, tais como: a inviabilidade de monitoramento por recursos humanos, mitigar o problema de treinamento de modelos devido à escassez de dados, bem como reduzir os impactos causados na detecção devido às transformações linguísticas por meio do uso de corpora com múltiplos idiomas distintos do idioma alvo. Portanto, esta pesquisa tem como objetivo principal utilizar meios computacionais para detectar discursos de ódio presentes em textos, empregando múltiplos idiomas no treino do modelo e um idioma de destino para testar a capacidade do modelo em detectar discurso de ódio em textos.

1.1.2 Objetivos Específicos

Alguns objetivos específicos foram elencados para alcançar o principal objetivo desta pesquisa, são eles:

- Avaliar o impacto ao utilizar múltiplos idiomas no treino do modelo e um idioma de destino para teste do modelo;
- Usar idiomas que têm uma distância léxica mais ampla. Por exemplo: usar os idiomas italiano e filipino no treino do modelo e o idioma inglês como idioma de destino para testar a capacidade do modelo em detectar discurso de ódio em textos;
- Usar idiomas que têm uma distância léxica menos ampla. Por exemplo: usar os idiomas italiano e espanhol no treino do modelo e o idioma português como idioma de destino para testar a capacidade do modelo em detectar discurso de ódio em textos;
- Utilizar modelos fundamentados em codificadores e decodificadores nos experimentos.

1.2 Questões de Pesquisa

Para alcançar os objetivos desta dissertação, algumas questões de pesquisa foram estabelecidas, as quais são:

- Questão 1: É possível aumentar a eficiência do modelo base ao utilizar CLL atrelado ao treinamento em múltiplos idiomas que sejam distintos do idioma de destino?
- Questão 2: Ao utilizar CLL em modelos codificadores, a eficiência do modelo é melhor quando a distância léxica dos idiomas utilizados no treino é mais próxima ou distante do idioma de destino?
- Questão 3: Ao utilizar CLL em modelos decodificadores, a eficiência do modelo é melhor quando a distância léxica dos idiomas utilizados no treino é mais próxima ou distante do idioma de destino?

1.3 Contribuições

Esta pesquisa tem como contribuição a análise da eficiência de modelos do tipo codificador e decodificador combinados com CLL para detectar discurso de ódio. Adicionalmente, trabalhos que empregam CLL para detectar discurso de ódio em textos utilizam apenas um idioma para o treino do modelo e apenas um idioma de destino para teste do modelo. Esta pesquisa tem como diferencial a inclusão de corpora com múltiplos idiomas para treino do modelo e um idioma de destino para teste do modelo. Ademais, foi avaliado o impacto dessa abordagem no resultado do modelo.

Destarte, o presente estudo se diferencia de outros trabalhos da mesma área de pesquisa, pelo uso e investigação de idiomas que apresentam diferenças léxicas próximas e distantes, visando examinar os efeitos causados nos dois casos. Como uma contribuição adicional, este trabalho também envolveu a criação e disponibilização de um corpus relacionado a discurso de ódio no idioma português. Esse corpus será disponibilizado publicamente para futuros pesquisadores na mesma área [22].

1.4 Publicações

Esta pesquisa teve como primeira publicação um artigo submetido à International Conference on Enterprise Information Systems (ICEIS) de 2023, classificada como Qualis A3. Essa publicação contempla os resultados iniciais obtidos neste estudo [21]. Os resultados são provenientes dos experimentos apresentados no Capítulo 5 com o modelo BERT e os idiomas cuja distância léxica é mais ampla (italiano, inglês e filipino).

Uma outra publicação foi aceita na Lecture Notes in Business Information Processing (LNBIP) em 2024, categorizada como Qualis B1. Nessa submissão, o modelo BERT foi empregado e mais idiomas foram incluídos em relação a publicação anterior. Os idiomas considerados foram: italiano, inglês, filipino, alemão e turco. Os resultados dos experimentos encontram-se no Capítulo 5, na parte de experimentos com codificadores e idiomas cuja distância léxica é mais ampla.

Uma terceira publicação foi aceita no ACM Symposium on Applied Computing (ACM SAC) em 2024, classificado como Qualis A2 [23]. Nessa submissão, foram apresentados os resultados provenientes do GPT-3, abrangendo os idiomas italiano, inglês, filipino, alemão e turco. Os resultados desses experimentos encontram-se no Capítulo 5, na parte de experimentos com decodificadores e idiomas cuja distância léxica é mais ampla.

1.5 Organização da Dissertação

A estrutura restante desta dissertação está organizada da seguinte maneira. No Capítulo 2, destaca-se a fundamentação teórica, apresentando os principais conceitos que regem esta pesquisa. No Capítulo 3, são descritos os trabalhos relacionados. No Capítulo 4, é apresentada a metodologia abordada neste trabalho, incluindo os corpora utilizados nos experimentos. No Capítulo 5, são discorridos os experimentos deste trabalho. Por fim, no Capítulo 6, encontram-se as conclusões finais desta pesquisa.

Capítulo 2

Fundamentação Teórica

O objetivo deste capítulo é abordar os principais conceitos que fundamentam esta pesquisa. Este capítulo está organizado da seguinte maneira: a Seção 2.1 destaca os conceitos relacionados ao Processamento de Linguagem Natural; a Seção 2.2 discute sobre aprendizado de máquina e aprendizado de máquina supervisionado; a Seção 2.3 aborda os *transformers* e sua arquitetura; a Seção 2.4 apresenta a definição de *Cross-Lingual Learning* e sua utilidade; a Seção 2.5 discute as principais métricas utilizadas para avaliação dos modelos neste trabalho; a Seção 2.6 aborda a temática sobre discurso de ódio e a sua importância na investigação desta pesquisa; a Seção 2.7 descreve sobre distância léxica; a Seção 2.8 apresenta os conceitos sobre teste de significância. Por fim, a Seção 2.9 contempla as considerações finais deste capítulo.

2.1 Processamento de Linguagem Natural

O termo linguagem natural engloba qualquer meio de comunicação desenvolvido e assimilado pelos seres humanos em seu entorno, utilizado para a interação entre eles [79]. Dentro desse contexto, essa linguagem possibilita a expressão de informações e conhecimentos de maneira independente do formato de comunicação adotado. O Processamento de Linguagem Natural (PLN) emerge como uma disciplina originada da convergência entre Inteligência Artificial e linguística, buscando analisar e entender as interações entre a linguagem humana e os sistemas computacionais. Destarte, o PLN contempla múltiplas técnicas destinadas a viabilizar o entendimento da comunicação natural humana por sistemas computacionais [12].

Essencialmente, o PLN representa uma abordagem que busca transcender os desafios inerentes à comunicação entre humanos e computadores, contemplando elementos teóricos e práticos, fornecendo a aptidão de lidar com diversas formas de comunicação humana, como texto e imagens. Essa capacidade possibilita a execução de várias tarefas, permitindo detectar discurso de ódio, realizar traduções entre vários idiomas, análise de sentimentos e emoções presente em textos, automação na resposta de perguntas em sistemas inteligentes, e a facilitação da interação entre humanos e computadores [28].

2.2 Aprendizado de Máquina

A aquisição de novas formas de compreensão é caracterizada como um processo fundamental de aprendizado. Dentre os diversos tipos de aprendizado empregados pelos seres humanos, destaca-se o aprendizado coletivo. Nessa abordagem, o aprendizado é aplicado por meio da passagem de conhecimento entre os seres humanos, facilitando assim o ser humano a aprender novos conhecimentos.

A concepção de um computador capaz de aprender similarmente aos seres humanos, representa um desafio investigado por estudos da área de Inteligência Artificial (IA) [81]. Ao longo dos anos, pesquisadores dessa área desenvolveram algoritmos que atualmente contemplam a subárea específica da IA conhecida como Aprendizado de Máquina (AM) [61].

Portanto, o AM desempenha o papel de examinar informações e extrair generalizações com o propósito de adquirir novos conhecimentos. Esse processo utiliza um algoritmo de computador para automatizar o aprendizado [63]. Dessa forma, um modelo computacional de AM pode ser personalizado em conformidade com parâmetros específicos, e o procedimento de aprendizado é realizado por meio da execução de um algoritmo computacional que otimiza as propriedades do modelo por meio de dados que são utilizados para treino e aprendizado da máquina [4]. O modelo pode ser do tipo preditivo, visando realizar previsões para eventos futuros ou pode ser do tipo descritivo, com o intuito de adquirir conhecimento a partir dos dados [4]. Os algoritmos de AM podem ser categorizados em [109]:

- Supervisionado: os modelos são treinados com dados anotados, fornecendo ao modelo conhecimento sobre cada dado dentro do conjunto, utilizando esse conhecimento adquirido para efetuar a classificação de dados futuramente apresentados ao modelo;

- Não supervisionado: os modelos funcionam de maneira inversa aos modelos supervisionados. Sendo assim, não são fornecidos dados anotados ao modelo, deixando a máquina analisar e classificar os dados por conta própria;

- Semi-supervisionado: contempla a junção das duas categorias anteriores. Portanto, o modelo é treinado usando um pequeno conjunto de dados rotulados. Esse conjunto servirá de base para o modelo. Em seguida, uma maior quantidade de dados será apresentada ao modelo. Esses novos dados não possuem anotação, deixando o modelo realizar a análise e classificação dos dados por conta própria.

2.2.1 Aprendizado de Máquina Supervisionada

O Aprendizado de Máquina Supervisionado (AMS) é uma técnica de aprendizado de máquina na qual modelos, constituídos por algoritmos computacionais, são treinados para identificar padrões em conjuntos de dados rotulados. Por meio desse processo de identificação de padrões, esses modelos são capazes de realizar a classificação de dados ou antecipar resultados. De maneira mais formal, é possível definir o aprendizado desses modelos como uma função F cujo propósito é mapear um dado X em uma classe Y associada ao dado X [109]. Ademais, é possível categorizar problemas relativos ao AMS nas seguintes categorias: em classificação e regressão [95]. Esses conceitos podem ser definidos da seguinte maneira:

- Classificação: nesta abordagem, o conteúdo analisado pode estar associado a uma categoria binária, por exemplo: “possui conteúdo ofensivo” ou “não possui conteúdo ofensivo”. Também é possível ter dados relacionados a múltiplas categorias, como em uma análise de emoção com categorias como: “alegre”, “triste”, “raiva”, entre outras. O propósito do modelo, nessas duas variantes de problemas, é classificar ou prever a qual categoria a informação apresentada ao modelo pertence.

- Regressão: nessa perspectiva, as informações não estão vinculadas a uma categoria, mas sim a um valor numérico real. Sob essa condição, quando os dados são fornecidos ao modelo, a finalidade é prever um valor numérico como saída que esteja associado a esses dados.

2.3 Transformers

O *transformer* é uma arquitetura composta por componentes como: redes neurais, codificadores e decodificadores [98]. O mecanismo de *Self-Attention* (SA) caracteriza-se como uma parte primordial dos *transformers*. Portanto, o SA permite capturar relações complexas entre diferentes partes textuais, tornando modelos baseados nessa arquitetura eficazes em lidar com dados sequenciais de comprimento variável. A Figura 2.1 ilustra a estrutura completa do *transformer* e seus principais componentes. Na parte esquerda temos os codificadores e na parte da direita temos os decodificadores. A arquitetura original é composta por seis codificadores e seis decodificadores [98].

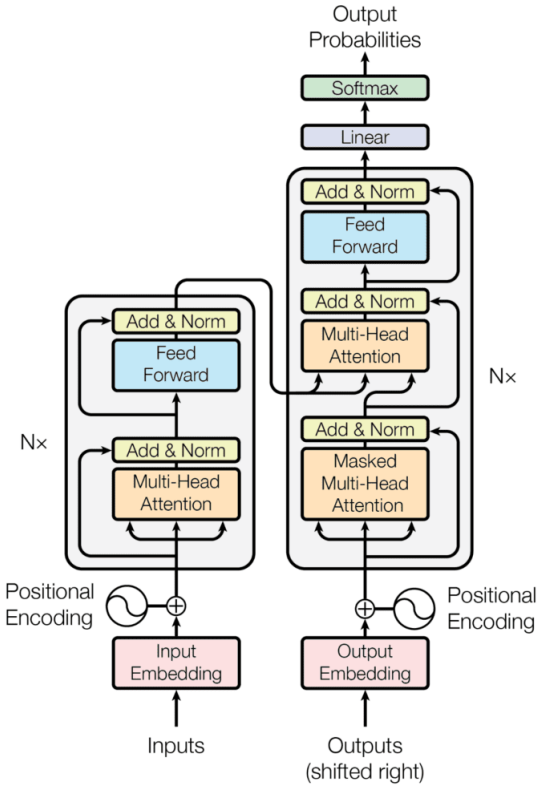

Figura 2.1: Representação da estrutura do *transformer*. Fonte: Vaswani et al. (2017).

Na Figura 2.1, os *inputs* denotam os textos recebidos na entrada do *transformer* e que posteriormente serão processados. Após a aquisição dos dados, eles são encaminhados à etapa de *input embedding*. Nesta etapa, ocorre a transformação dos textos em representações vetoriais. Essas representações são obtidas através de técnicas de AM que mapeiam

palavras ou frases em vetores [101; 60; 48]. Como resultado, esses vetores acoplam as propriedades semânticas e sintáticas das palavras, fazendo com que palavras mais semelhantes fiquem mais próximas e as menos semelhantes fiquem mais distantes. A Figura 2.2 exibe um exemplo de representação vetorial usando embedding.

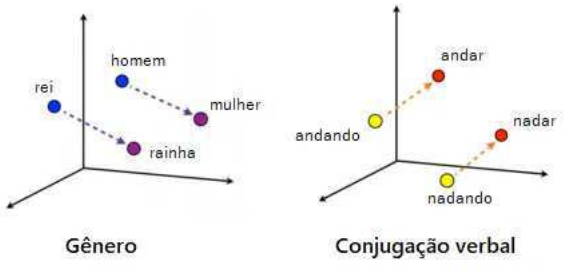

Figura 2.2: Representação vetorial de algumas palavras usando *embedding*. Fonte: Fonseca (2021).

Para a elaboração de bons *embeddings*, estes devem ter dimensionalidade adequada, capturando informações semânticas relevantes. Portanto, é necessário um corpus denso e com boa representação textual, contemplando a interação de diversas palavras para uma melhor captura semântica entre elas. Adicionalmente, bons *embeddings* devem também representar palavras e conceitos de forma que aqueles semanticamente semelhantes estejam próximos no espaço vetorial. Ademais, a capacidade de generalização e a adaptabilidade a novas palavras são importantes para lidar com dados novos e variados. Por fim, outro fator importante é a eficiência computacional para garantir a praticidade no uso de recursos.

Uma vez obtida a representação vetorial das palavras, esses vetores são encaminhados à etapa de *positional encoding*. Esta etapa é essencial para se obter a posição de *tokens* na sentença recebida. A necessidade do *positional encoding* se fundamenta na importância de representar a ordem das palavras em um texto. Por exemplo, consideremos as seguintes frases: “o gato viu o rato” e “o rato viu o gato”. É perceptível que, ao alterar a posição das palavras “gato” e “rato” na mesma frase, o significado é completamente modificado. Portanto, na fase de *positional encoding*, um vetor indicando a posição é adicionado a cada vetor de *embedding* (texto vetorizado da etapa anterior). Isso permite ao modelo considerar a ordem das palavras durante o processamento subsequente, garantindo uma representação mais completa e contextual dos dados.

Posteriormente, os dados são encaminhados para o codificador. Após o recebimento desses dados, estes são direcionados para uma camada denominada *Multi-Head Attention*. Nessa camada, é realizado o mecanismo de SA, o qual examina as relações entre os termos nos textos vetorizados provenientes da etapa anterior. Em outras palavras, para cada texto na sequência, são calculados pesos que indicam a relevância das demais palavras na sentença.

Essa avaliação é realizada por meio de projeções lineares, representadas por *Query (Q)*, *Keys (K)* e *Values (V)*. A Figura 2.3 oferece uma representação visual dessas projeções. Esse processo é crucial para capturar as conexões semânticas e contextuais envolvendo as palavras, contribuindo para uma representação mais rica e informacional na fase de codificação do modelo.

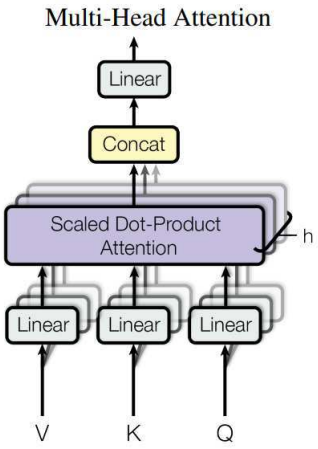

Figura 2.3: Representação da camada *Multi-Head Attention*. Fonte: Vaswani et al. (2017).

Os vetores Q, K, e V são concebidos para cada termo na sentença recebida como entrada. Esses vetores são requeridos no cálculo do peso relativo de um termo X em relação a outro termo Y dentro da sentença. Essa ponderação é necessária para a avaliação da relação que envolve os termos X e Y, bem como com as demais palavras presentes no texto. Essa abordagem permite a codificação das inter-relações semânticas entre os termos, contribuindo para uma representação mais rica e informacional durante a fase de processamento do modelo. O procedimento é repetido para cada elemento na sentença recebida como entrada, e a computação é realizada por meio das seguintes etapas:

- Efetuar a operação $Q_i \times K_j$, em que i e j representam a posição das palavras a serem

analisadas na sentença;

- Dividir o resultado pela raiz quadrada levando em consideração a dimensão dos vetores *keys* ($\sqrt{d_k}$);
- Utilizar o valor obtido como parâmetro. Este parâmetro será recebido como entrada na camada *softmax*. Esta camada normaliza os valores de maneira que sejam todos positivos. Ademais, a somatória total deve ser igual a 1. A finalidade disso é avaliar a relevância da palavra em análise relacionada às demais;
- Proceder à multiplicação dos valores do vetor *values* pelo resultado do *softmax*;
- Por último, realizar a soma dos vetores resultantes, gerando um vetor final. Esse vetor indica o peso da palavra atual comparada à(s) outra(s) palavra(s).

Por exemplo, considere a seguinte sentença: "Aprendendo transformer com ilustração". Para calcular a relação da palavra "Aprendendo" com a palavra "transformer", o algoritmo realiza inicialmente o cálculo do peso da palavra "Aprendendo" consigo mesma ($q1 \times k1$) e, logo após, calcula o peso relacionado à palavra "transformer" ($q1 \times k2$). Posteriormente, os valores de cada operação são normalizados pela divisão por $\sqrt{d_k}$, onde $d_k = 64$. Portanto, $\sqrt{d_k} = \sqrt{64} = 8$. O valor 64 foi adotado porque ele representa o valor utilizado na estrutura original do *transformer* [98]. Após essa etapa, os valores são encaminhados para a camada de *softmax*, e o valor resultante é multiplicado pelo vetor *values*. Por fim, os valores são somados, gerando o vetor final (Z) para cada termo. A Figura 2.4 ilustra o exemplo mencionado.

Uma das características vantajosas do *transformer* é o potencial de realizar os cálculos do *Self-Attention* em paralelo. Em outras palavras, é possível representar conjuntos de palavras vetorizadas em forma de matrizes e efetuar os cálculos simultaneamente. Na Figura 2.3, essa funcionalidade é representada pela camada denominada *Scaled Dot-Product Attention*. A Equação 2.1 ilustra como os valores podem ser calculados em formato de matriz nessa camada [98]. Os símbolos Q, K e V referem-se, respectivamente, às matrizes formadas pelos valores dos vetores das *queries*, *keys* e *values* das entradas. O processo envolve a aplicação da função *softmax* em QK^T, seguida pela divisão por $\sqrt{d_k}$ (dimensão das *keys*). O valor obtido é então multiplicado por V.

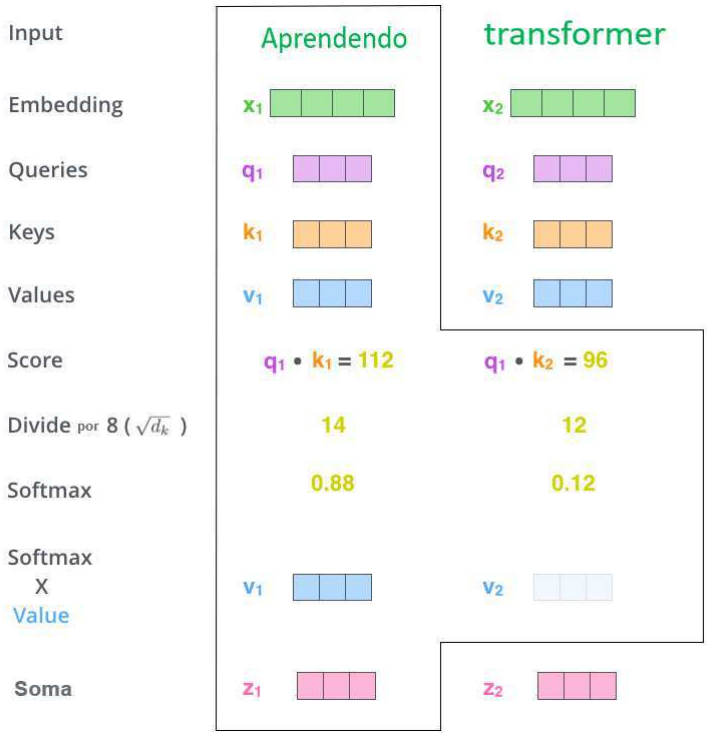

Figura 2.4: Exemplo de cálculo do *Self-Attention* (adaptado pelo autor). Fonte: Alammar (2018).

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

Após executar esse procedimento, o codificador encaminhará o valor resultante para a próxima camada denominada *Feed Forward*, a qual é composta por um modelo computacional fundamentado em redes neurais *feed-forward* [3]. Todo esse processo é iterado até o último codificador. Ao atingir o último codificador, o valor resultante é direcionado para os decodificadores.

Os decodificadores têm a mesma estrutura dos codificadores. No entanto, embora os decodificadores compartilhem as mesmas camadas e cálculos que o codificador, existem três distinções entre eles:

- Camada *Multi-Head Attention*: diferentemente do codificador, esta camada recebe os valores retornados pelo último codificador na estrutura do *transformer*;
- Camada *Masked Multi-Head Attention*: os decodificadores incorporam esta camada

adicional. Ela mascara a próxima palavra na sentença, de maneira que o modelo aprenda qual a próxima palavra mediante a observação das relações envolvendo as palavras na sentença. Isso possibilita que o modelo adquira representações mais ricas e complexas, capturando informações contextuais das sentenças recebidas como entrada;

- O resultado do último decodificador é reintroduzido no primeiro decodificador. Essa palavra assume um papel contextual para o modelo e é empregada para efetuar a predição da palavra subsequente.

Apesar de complexo, o *transformer* contribui significativamente para a área de PLN, sendo utilizado para criação de vários modelos, tais como: BERT [25], GPT [75], GPT-2 [76], GPT-3 [10], T5 [77], dentre outros. Neste trabalho foram utilizados três modelos: BERT, GPT-3 e o GPT 3.5.

2.4 *Cross-Lingual Learning*

Um dos desafios presentes na criação de modelos para classificação de dados é a insuficiência de dados abrangendo vários idiomas. Estudos indicam que uma parcela significativa dos experimentos relacionados ao PLN estão concentrados no idioma inglês, abrangendo aproximadamente 60% das publicações [34; 74]. Consequentemente, outros idiomas enfrentam uma carência de corpora e, por conseguinte, de investigação científica [73].

Para contornar a escassez de dados, ao longo do tempo, diversas abordagens foram desenvolvidas. Uma dessas abordagens é o *Cross-lingual Learning* (CLL). Essa abordagem emprega técnicas de AM, onde características derivadas de um ou mais idiomas-fonte são explorados para aprimorar a eficiência em uma língua-destino, frequentemente caracterizada por recursos linguísticos relativamente limitados [73]. O CLL tem como objetivo desenvolver modelos capazes de generalização em uma língua ou várias línguas, mesmo diante de disparidades léxicas e/ou sintáticas.

Além disso, o CLL demonstra-se eficaz em aprimorar modelos em diversas áreas do PLN, incluindo análise de sentimentos, tradução automática, reconhecimento de entidades nomeadas, entre outras [73]. Dessa forma, o CLL representa uma técnica que utiliza dados

de um determinado idioma fonte para treinar um modelo de IA a classificar dados em um idioma diferente do idioma fonte [73]. Isso permite que um modelo possa ser aplicado a diversas tarefas de classificação, mesmo que não haja grandes quantidades de dados de treinamento disponíveis para esses idiomas.

É crucial observar que, para maximizar a passagem de aprendizado, é necessário incorporar fontes linguísticas distintas provenientes de um ou mais idiomas que o modelo ainda não tenha visto. Dessa forma, a inclusão de vários corpora de múltiplos idiomas pode significativamente aprimorar a eficiência do modelo. Neste estudo, foram empregados até sete idiomas como fontes e um idioma como destino, sendo os idiomas de destino e fonte distintos.

2.5 Métricas para Avaliação do Desempenho do Modelo

Considerando que este trabalho emprega modelos de AM para detectar discurso de ódio em textos, torna-se essencial a aplicação de métricas para analisar a eficiência do modelo nas classificações, bem como para comparar a eficiência entre diferentes modelos. Portanto, para verificar a eficácia dos modelos, foram adotadas as métricas de Precisão (*Precision*), Revocação (*Recall*) e Medida-F1 (*F1-Score*) [108; 39]. O cálculo dessas métricas pode ser expresso da seguinte maneira:

- VP (verdadeiro positivo): é o total de instâncias classificadas de forma correta pelo modelo para a classe positiva;
- FP (falso positivo): é o total de instâncias classificadas de forma errada pelo modelo para a classe positiva;
- VN (verdadeiro negativo): é o total de instâncias classificadas de forma correta pelo modelo para a classe negativa;
- FN (falso negativo): é o total de instâncias classificadas de forma errada pelo modelo para a classe negativa.

Com base nos conceitos previamente apresentados, a métrica Precisão pode ser calculada conforme mostra a Equação 2.2. Essa métrica mede a precisão das previsões positivas feitas pelo modelo. Ela representa a proporção de classificações positivas corretas (VP) em relação

ao número total de classificações positivas (a soma dos VPs e FPs). Em outras palavras, mostra quantos dos itens que o modelo identificou como positivos são realmente positivos.

$$\text{Precisão} = \frac{VP}{FP + VP} \tag{2.2}$$

A Revocação, pode ser calculada conforme demonstrado na Equação 2.3. Essa métrica avalia dentre todas as ocorrências esperadas da classe rotulada como positiva, quantas foram corretamente identificadas pelo modelo. Em outras palavras, ela examina a proporção de exemplos da classe rotulada como positiva (VP) em relação ao total de FNs somados aos VPs.

$$\text{Revocação} = \frac{VP}{FN + VP} \tag{2.3}$$

Por fim, temos a Medida-F1. Essa métrica pode ser calculada conforme mostra a Equação 2.4. Essa medida representa uma média harmônica da precisão e revocação. Sendo útil em casos nos quais ambas as métricas são importantes e há a necessidade de equilibrar esses dois aspectos.

$$\text{Medida-F1} = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.4}$$

2.6 Discurso de Ódio

Nos últimos tempos, o volume de informações tem aumentado significativamente devido à internet. Atualmente, os cidadãos podem se expressar livremente por meio de aplicativos disponíveis em dispositivos eletrônicos, como celulares, tablets, computadores, entre outros. Tais aplicativos proporcionam meios para as pessoas expressarem suas opiniões por meio do envio de fotos, vídeos, áudios, textos e outros formatos de conteúdo.

A liberdade de expressão é um princípio fundamental nas sociedades, reconhecendo o direito do cidadão em expressar suas opiniões livremente, sem censura ou restrições indevidas [69]. Entretanto, embora nos meios digitais exista a possibilidade do público produzir conteúdos saudáveis, é importante destacar que esse mesmo espaço pode ser uma ferramenta para as pessoas disseminarem conteúdos com discurso de ódio, caracterizando-se por

expressões que têm como propósito insultar, intimidar ou assediar pessoas [88]. A disseminação desse tipo de conteúdo pode manifestar-se por meio de expressões que buscam atacar ou difamar grupos sociais, muitas vezes incitando à violência contra esses grupos, seja por características físicas, religiosas, origem étnica, entre outras [34].

A mitigação desse problema assume um papel crucial, dado que tais expressões têm o potencial de incitar violência social relacionada a questões como raça, religião, xenofobia, sexismo, entre outras. A identificação e antecipação da ocorrência desse problema desempenham um papel preventivo, contribuindo para evitar e reduzir a incidência de violência social [85]. No entanto, é impraticável realizar o controle manual das mensagens compartilhadas por usuários em plataformas digitais devido ao elevado volume de comentários.

Na rede social X, por exemplo, em torno de 500 milhões de mensagens são compartilhadas diariamente[6; 83; 2]. Diante desse volume expressivo, torna-se evidente que a mitigação de conteúdos com discurso de ódio em redes sociais representa um desafio complexo [46]. Dada a natureza desse problema e a impossibilidade prática de controle manual, tem-se observado um aumento na adoção e desenvolvimento de ferramentas automatizadas para mitigar esse problema ao longo dos últimos anos [85].

2.7 Distância Léxica

Entende-se como léxico o vocabulário presente em um determinado idioma disponibilizado para que o público possa se expressar oralmente e por escrito [30]. Portanto, podemos definir a distância léxica como uma medida que quantifica a diferença ou semelhança entre dois conjuntos de palavras relacionadas ao vocabulário. Essa métrica serve para medir a diferença ou similaridade lexical entre duas ou mais entidades linguísticas (idiomas). No entanto, medir a diferença/similaridade entre idiomas é algo que vai além da simples avaliação de uma métrica sobre conjuntos de palavras, pois envolve analisar a origem dos idiomas e suas mudanças ao longo do tempo.

Assim, é possível utilizar técnicas para calcular a distância léxica entre idiomas. Uma dessas técnicas é a distância de Levenshtein [53]. A distância de Levenshtein é uma métrica que calcula a distância entre cadeias de caracteres entre palavras [45]. Portanto, a distância pode ser calculada como o valor mínimo de operações de edição necessárias para transformar

sequências de caracteres de uma certa palavra em outra. As operações de edição englobam:

- Inserção: custo para adicionar um caractere a uma palavra;
- Exclusão: custo para remover um caractere da palavra;
- Substituição: custo para trocar um caractere por outro.

Logo, a técnica busca o menor caminho para operações de inserção, exclusão ou substituição necessárias para transformar uma cadeia de caracteres da palavra em outra [51]. Por exemplo, ao calcular a distância entre "casa" e "café" seria necessário dois passos de substituição: substituir "s" por "f" e "a" por "é". Portanto, a distância de Levenshtein entre essas duas palavras seria igual a dois.

Nesta dissertação, foi utilizado um corpora contemplando os idiomas italiano, filipino, português, espanhol, alemão, turco e inglês. No entanto, dado que o escopo desta pesquisa não abrange a verificação da distância léxica entre idiomas, e considerando a complexidade e inviabilidade de realizar tal análise, este trabalho baseou-se nos resultados de outras pesquisas que examinaram as diferenças linguísticas entre esses idiomas, incluindo a distância léxica.

Serva e Petroni (2008) conduziram um experimento abrangendo 50 idiomas [87]. Esses autores analisaram a distância léxica de idiomas por meio da métrica de Levenshtein. A Figura 2.5 apresenta a distância léxica encontrada para alguns dos idiomas indo-europeus analisados. O eixo y mostra a similaridade do tronco linguístico entre os idiomas e o eixo x representa a mudança dessa similaridade ao longo do tempo. É possível perceber que os idiomas português (*portuguese*), espanhol (*spanish*) e italiano (*italian*) apresentam uma distância mais próxima entre si. Em contraste, o idioma inglês (*english*) exibe uma distância consideravelmente maior em relação a esses idiomas. Entretanto, é importante notar que o inglês compartilha um tronco linguístico com o alemão (*german*), o que resulta em uma proximidade linguística mais significativa entre ambos.

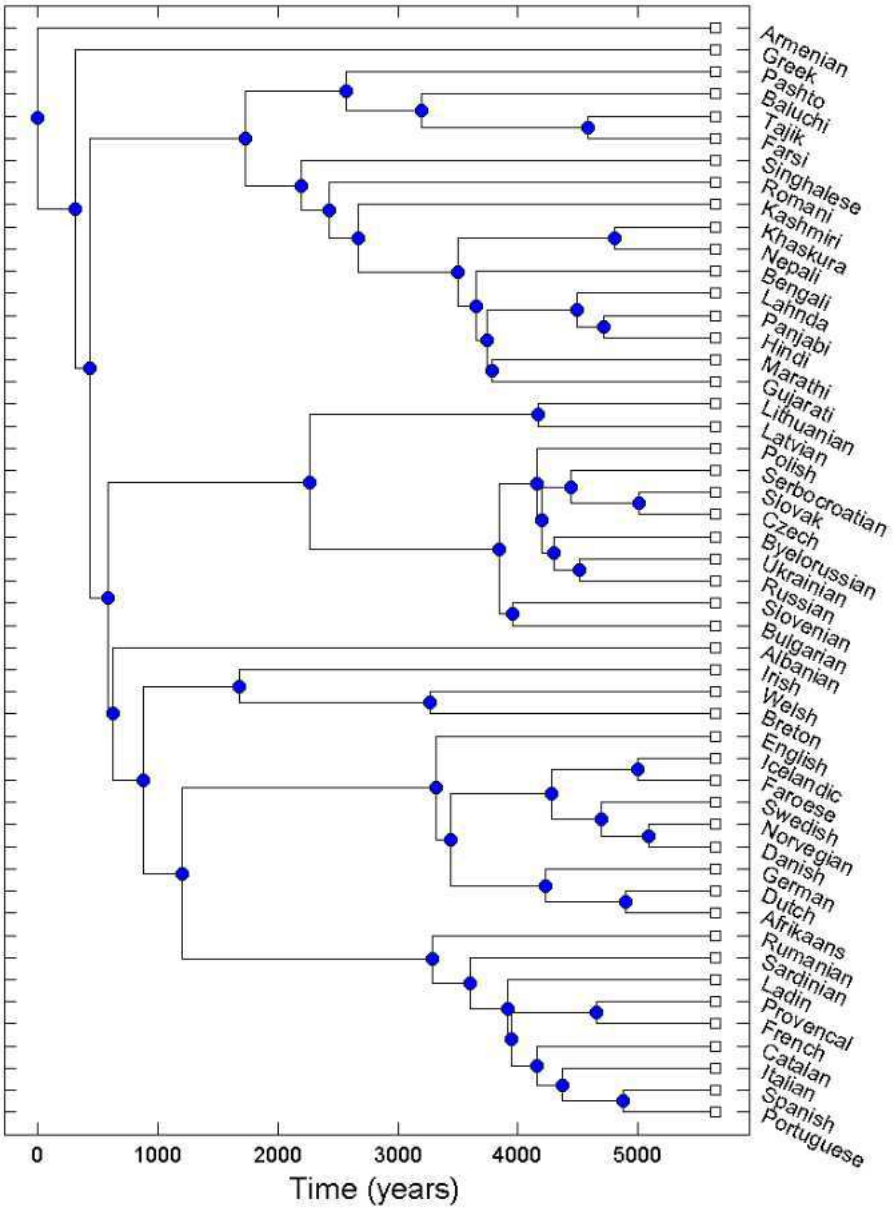

Figura 2.5: Distância léxica dos idiomas indo-europeus. Fonte: Serva e Petroni (2008).

Quanto ao idioma filipino, Gray e Jordan (2001) conduziram uma análise abrangendo idiomas das Filipinas, Taiwan, Polinésia, entre outros [40]. A Figura 2.6 apresenta a proximidade entre esses idiomas, evidenciando a afinidade do idioma filipino com línguas como o tagalo (*tagalog*) e cebuano, por exemplo.

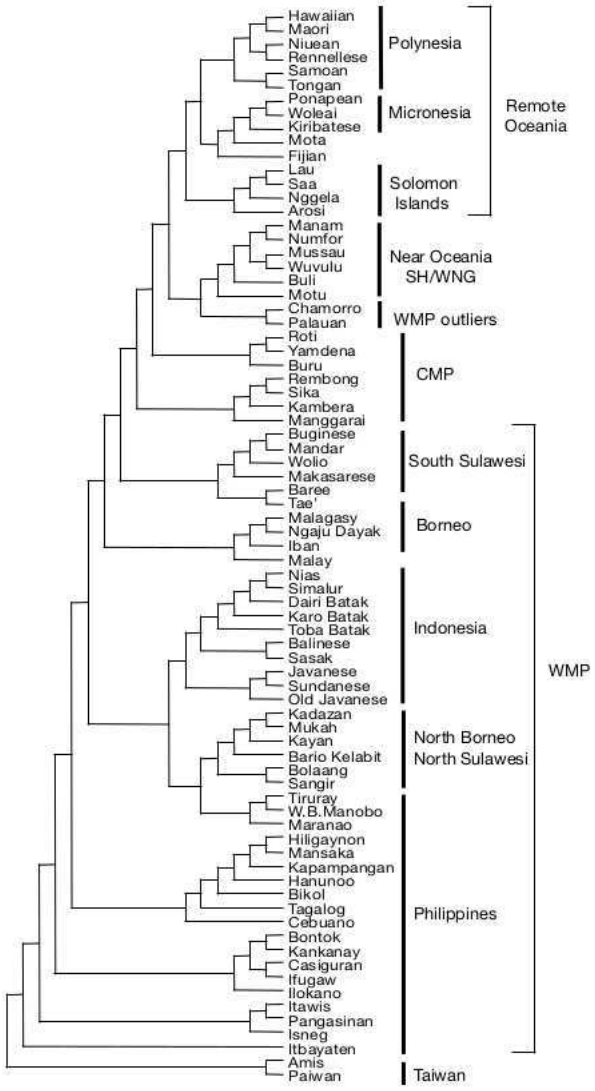

Figura 2.6: Distância das línguas austronésias (adaptado pelo autor). Fonte: Gray e Jordan (2001).

Savelyevr e Robbeets (2020) investigaram a origem do idioma turco ao longo do tempo [82]. A Figura 2.7 ilustra a origem do idioma turco (*turkish*). Os valores entre cada árvore representam o grau de confiança da similaridade do tronco linguístico entre os idiomas. Quanto mais próximo do valor um, maior é a confiança na similaridade. Porém, quanto mais próximo de zero, menor é a confiança na similaridade do tronco linguístico. Adicionalmente, nota-se a proximidade do idioma turco com o gagauz e o azerbaijano (*azeri*). No entanto, é observável que o turco se mostra mais distante da língua tuvana (*tuvan*) e tofalar (*tofa*), entre outras.

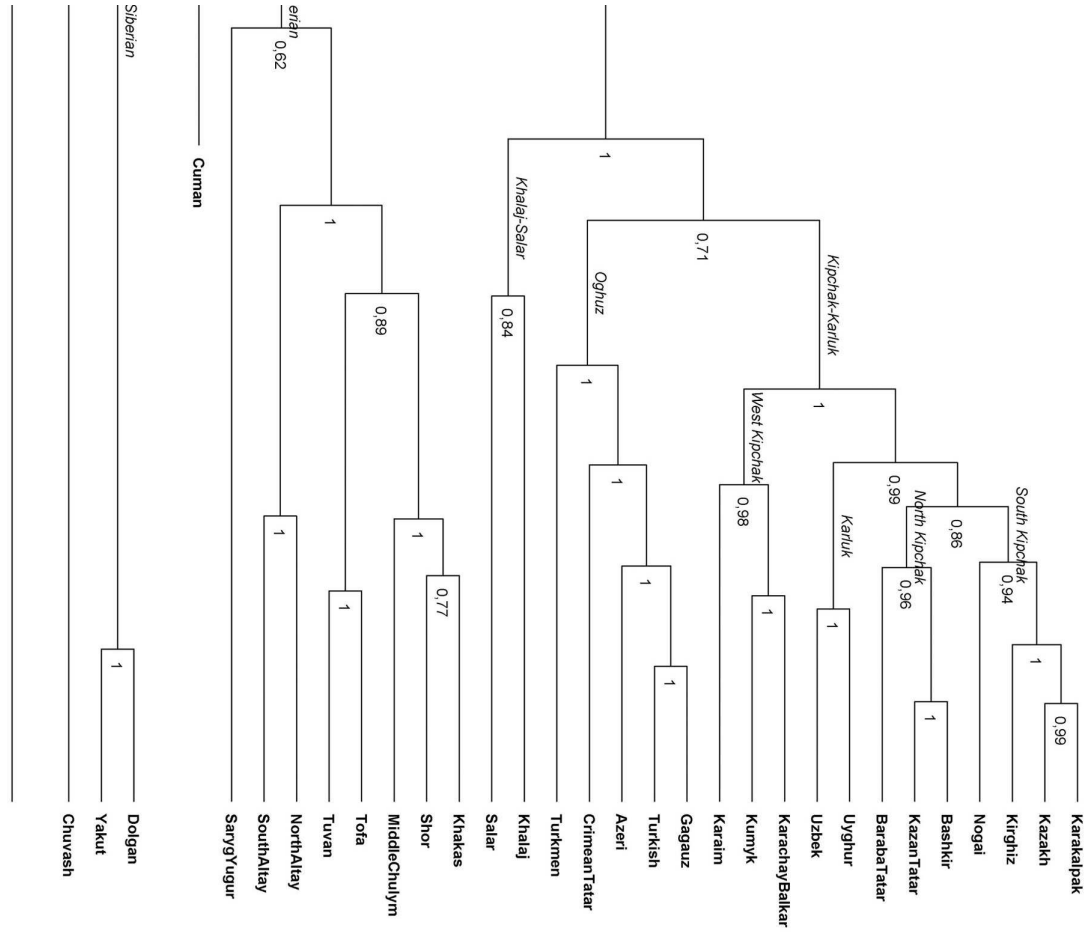

Figura 2.7: Origem linguística do idioma turco (adaptado pelo autor). Fonte: Savelyevr e Robbeets (2020).

2.8 Teste de Significância

Um dos desafios enfrentados em algoritmos de AM está relacionado à comparação do desempenho entre modelos distintos. Muitas vezes métricas tradicionais de comparação (precisão, revocação e medida-F1) são utilizadas para medir a eficiência entre diferentes modelos. No entanto, apesar dessas métricas serem boas para comparar a eficiência individual de cada modelo, elas não conseguem medir a significância estatística dos resultados apresentados entre dois modelos distintos. A justificativa para isso se dá devido à influência da aleatoriedade que os parâmetros utilizados nos modelos podem causar no resultado final [26; 55; 94].

Sendo assim, isso faz com que os valores resultantes das métricas obtidas pelo modelo

na etapa de classificação esteja interligada a parâmetros inerentes à estrutura do modelo. Um exemplo seria a aleatoriedade de inicialização dos pesos do modelo e também aos elementos (imagens, vetores, textos, etc) fornecidos como entrada. Além disso, outro fator está relacionado aos dados de pré-treino que podem variar de modelo para modelo, bem como aos hiper-parâmetros utilizados como taxa de aprendizado, quantidade de neurônios, quantidade de épocas, etc. De uma forma geral, estes fatores permitem uma gama maior de configurações e aumentam as variações de resultados produzidos pelos diferentes modelos. Portanto, a eficiência dos modelos pode variar conforme a aleatoriedade dos elementos passados a eles, podendo assim interferir nos resultados e na comparação final.

Para comparar a eficiência entre dois algoritmos de AM, podemos utilizar outros meios de comparação como o *Almost Stochastic Order* (ASO). Dror et al. (2019), propôs essa abordagem para calcular a significância estatística entre modelos de PLN [26]. A abordagem leva em consideração a variação entre dois elementos a serem avaliados e realiza o teste de significância tomando como base o princípio relacionado à ordem estocástica. O cálculo realiza a comparação relacionada a dois elementos X e Y, levando em conta que nem sempre o elemento X será superior a Y, podendo esse mesmo elemento variar em certos casos. Dessa forma, os valores de dominância entre X e Y são considerados até o momento em que seja violada a ordem estocástica por um determinado limiar.

Em suma, o ASO produz um valor que representa um limiar superior (θ). Esse valor pode ser utilizado para comparação entre dois elementos. Se o valor θ retornado for inferior a 0,5, podemos dizer que o elemento X pode ser estocasticamente considerado dominante sobre Y em grande parte dos casos. Sendo assim, podemos considerar o elemento X superior ao elemento Y. O valor retornado pode ser utilizado também como um valor de confiança, isto é, quanto menor o seu valor maior a certeza em aceitar X sendo dominante em relação a Y.

No entanto, vale destacar que o ASO não calcula estatisticamente valores p. Sendo assim, as hipóteses nula e alternativa podem ser compreendidas conforme mostra a Equação 2.5. Portanto, dadas as hipóteses nula (H0) e alternativa (H1), devemos aceitar H0 se o valor θ for superior ou igual a 0,5. Caso o valor seja inferior a 0,5, então aceitamos H1 e rejeitamos H0.

$$\begin{aligned} H_0 &: \theta \geq 0,5 \\ H_1 &: \theta < 0,5 \end{aligned} \tag{2.5}$$

Ulmer et al. (2022) elaboraram uma ferramenta de livre acesso que usa a abordagem ASO para comparar modelos de PLN [94]. Assim, essa ferramenta foi utilizada neste trabalho para realizar os cálculos de comparação entre determinados modelos.

2.9 Considerações Finais

Este capítulo abordou os principais conceitos que fundamentam esta pesquisa. Foram discutidos conceitos relacionados ao PLN e AM, com ênfase no AM supervisionado. Também foram fornecidos detalhes sobre a principal arquitetura utilizada nos modelos desta pesquisa, conhecida como *transformer*. Além disso, destacou-se a definição de CLL e a sua importância na abordagem deste trabalho. Foram apresentadas as principais métricas utilizadas para a avaliação dos modelos. Adicionalmente, foi descrito o conceito sobre distância léxica entre idiomas, que é abordado nesta pesquisa. Por fim, foi apresentado o conceito sobre teste de significância para comparação de modelos. O próximo capítulo abordará as pesquisas que estão relacionadas a esta dissertação.

Capítulo 3

Trabalhos Relacionados

Este capítulo visa explorar os estudos existentes e relacionados a discurso de ódio, proporcionando uma análise abrangente do cenário atual de pesquisa. Inicialmente, será abordado o trabalho anterior ao qual este trabalho possui como base. Posteriormente, serão abordadas as pesquisas relacionadas à mesma área de pesquisa deste trabalho. Em seguida, serão citadas as pesquisas relacionadas a CLL voltado à detecção de discurso de ódio. Por fim, serão apresentadas as considerações finais deste capítulo.

3.1 Trabalho de Firmino

Este trabalho tem como ponto de partida a pesquisa de doutorado conduzida por Firmino (2021) [31]. Em sua investigação, o autor utilizou CLL para detectar discurso de ódio. Nos experimentos, o autor optou por utilizar apenas um idioma para treino do modelo e um idioma de destino para teste, por exemplo: italiano como fonte de treino para o modelo e o idioma português como alvo de classificação. Além disso, nos experimentos foram utilizados dados com as distâncias léxicas próximas. É importante destacar também que foram conduzidos experimentos exclusivamente em modelos baseados em codificadores.

O presente trabalho diferencia-se da pesquisa anterior pela incorporação de modelos fundamentados em decodificadores e codificadores, proporcionando uma análise comparativa dos resultados desses dois tipos de modelos. Adicionalmente, esta pesquisa aborda experimentos e análises relacionadas às disparidades léxicas, abordando dados de linguagens que possuem distância léxica maior, bem como dados de linguagens com distância léxica menor.

Dessa forma, para essa investigação foram empregados sete idiomas distintos: português, filipino, italiano, inglês, espanhol, alemão e turco.

Além disso, outra diferença relacionada ao trabalho anterior, é de que nesta dissertação foram realizados vários experimentos com múltiplos idiomas, que serviram como idioma fonte na fase de treinamento dos modelos, bem como a elaboração de experimentos com a classificação individual para alvos distintos (português, italiano e inglês). Ademais, explora-se a inclusão de corpora com idiomas que possuem um tronco linguístico mais ampliado, objetivando analisar a eficiência dos modelos diante do acréscimo de corpora mais abrangentes, especialmente quando relacionados a idiomas distintos com disparidades léxicas maiores. Em contraste, também foi utilizado corpora com múltiplos idiomas lexicalmente próximos, a fim de analisar a eficiência do modelo quando se trata de corpora que contemple idiomas distintos com disparidades léxicas menores.

Por fim, além dos pontos mencionados anteriormente, destaca-se que este trabalho tem também como contribuição adicional a disponibilização de um corpus devidamente rotulado e relacionado a discurso de ódio no idioma português, agregando contribuição ao campo de estudo e oferecendo um corpus para futuras investigações.

3.2 Trabalhos Referentes a Discurso de Ódio

Detectar automaticamente textos que apresentem discurso de ódio não é trivial, dado que construir um modelo capaz de compreender e classificar textos em múltiplos idiomas apresenta desafios significativos. Além disso, boa parte dos modelos requerem textos que contenham instâncias desse tipo de problema para treino do modelo em classificação supervisionada. Ao longo dos anos, diversas pesquisas têm sido conduzidas para mitigar esses desafios.

Waseem e Hovy (2016) desenvolveram um corpus composto por 16.000 textos, dos quais 3.383 são referentes a conteúdo racista e 1.972 a conteúdo sexista [102]. Na pesquisa, os autores empregaram Regressão Logística (RL) para classificar os textos, alcançando um desempenho de 73,93% na medida-F1. No entanto, é importante notar que os experimentos e análises dos resultados foram realizadas utilizando exclusivamente a RL, sem efetuar comparações com outros algoritmos de AM.

Fortuna e Nunes (2018) conduziram uma análise sistemática sobre o estado-da-arte da detecção automatizada de discurso de ódio [35]. No estudo, foi abordada a complexidade inerente a esse problema, encontrado em diversas maneiras nas plataformas sociais. Ademais, o artigo destacou algumas limitações fundamentais no campo de pesquisa, como a falta de padronização dos termos utilizados, a escassez de corpus anotado com qualidade, a dificuldade em lidar com o sarcasmo e a ambiguidade presentes em textos, bem como os desafios éticos e legais envolvidos.

A pesquisa de Fortuna e Nunes contribuiu significativamente para o entendimento das complexidades e desafios enfrentados para detectar discurso de ódio. Contudo, é importante notar que, mesmo depois de conduzido esse estudo, tais desafios continuam a ser obstáculos para o avanço efetivo nesse domínio. O reconhecimento dessas limitações destaca a necessidade contínua de abordagens inovadoras e desenvolvimento de ferramentas para detectar automaticamente discurso de ódio.

Davidson et al. (2017) destacam a complexidade em distinguir e classificar sentenças de discurso de ódio comparada a outras ofensas comuns [20]. Os autores fizeram uso do site Hatebase.org, que reúne uma série de termos considerados ofensivos. Essa coleção serviu para coletar os textos analisados na pesquisa. Após coletarem os textos, foram selecionados aleatoriamente 25.000 deles, posteriormente submetendo esses dados à plataforma CrowdFlower para serem manualmente anotados. Na pesquisa, modelos tradicionais de classificação foram empregados (Regressão Logística, *Naive Bayes*, *Decision tree*, *Random Forest* e SVM), resultando em um F-score máximo de 90%. Embora os autores tenham conduzido experimentos com diversos modelos, a informação relacionada ao balanceamento de classes nos dados coletados e rotulados não foi fornecida.

Frenda et al. (2019) examinaram os aspectos computacionais e as distinções entre sexismo e misoginia [36]. Os autores empregaram corpora relacionados à misoginia e ao sexismo em seus experimentos [102]. Uma ferramenta de AM chamada *Support Vector Machine* (SVM) foi designada para identificar instâncias de discurso de ódio em textos. Entretanto, esses experimentos foram limitados ao idioma inglês apenas, e nenhuma investigação envolvendo dados multilíngues foi conduzida. Adicionalmente, os autores empregaram exclusivamente a métrica de acurácia, deixando de empregar métricas como precisão, revocação e medida-F1 para verificar a eficiência do modelo.

Essa abordagem metodológica apresenta algumas limitações. Ao restringir a análise apenas ao inglês, a pesquisa não aborda a complexidade da detecção em um contexto multilíngue, o que é crucial dada a diversidade linguística nas plataformas digitais. Além disso, a avaliação baseada apenas na métrica de acurácia pode não fornecer uma visão completa da eficiência do modelo, pois não considera as nuances de falsos negativos e falsos positivos, elementos críticos em tarefas para detectar discurso de ódio. Faz-se necessário a aplicação de métricas adicionais para uma avaliação mais abrangente do modelo apresentado.

Del Arco et al. (2021) conduziram uma pesquisa relacionada à detecção de discurso de ódio em redes sociais, focando no idioma espanhol [24]. A pesquisa abordou uma análise comparativa entre modelos de *Deep Learning* e modelos pré-treinados de *Transfer Learning* (TL). Os resultados mais promissores foram alcançados ao empregar modelos de TL. Contudo, vale destacar que o escopo do trabalho se restringiu exclusivamente ao idioma espanhol, sem efetuar comparações entre modelos considerando corpora de origens linguísticas diversas.

Embora os resultados destaquem a eficácia dos modelos do tipo TL em detectar discurso de ódio em espanhol, a falta de comparações interlinguísticas limita a generalização desses achados para outros idiomas. Considerando a diversidade de linguagens presentes nas plataformas digitais, uma investigação mais abrangente que aborde a variação linguística poderia oferecer uma compreensão mais completa das capacidades e limitações dos modelos propostos.

Mathew et al. (2019) examinaram as ramificações da violência induzida por conteúdos que contenham discurso de ódio [59]. Para conduzir essa análise, utilizaram um extenso corpora contendo textos de 341.000 usuários da rede social Gab. O objetivo foi elucidar as tendências predominantes e os comportamentos que influenciam os comentários nos grupos de usuários presentes na referida rede social.

Na análise, foi empregado o modelo DeGroot [38] para identificar as conexões entre os usuários. Os autores chegaram a conclusão que os usuários que promovem discurso de ódio formam uma comunidade densamente interconectada, responsável por gerar aproximadamente 25% do conteúdo no Gab, apesar de representarem apenas 0,67% do total de usuários. Esse estudo revela percepções sobre a dinâmica dos conteúdos publicados pelos usuários na plataforma Gab, destacando a concentração de influência em uma parcela de usuários.

Entretanto, é essencial considerar as limitações do estudo, como possíveis vieses nos dados coletados, bem como a generalização desses resultados para outras plataformas sociais.

Yang et al. (2022) sugeriram um framework denominado *Cross-Domain Knowledge Transfer* (CDKT) para detectar discurso de ódio, tendo como base os *transformers* [105]. O objetivo central desse framework é mitigar as lacunas de definição e semânticas entre a detecção de sarcasmo e discurso de ódio. Os autores chegaram à conclusão que o CDKT apresentou resultados satisfatórios ao reduzir as discrepâncias entre esses dois tipos de conteúdos em diversos corpora. Entretanto, é importante destacar que o estudo carece de uma análise comparativa com outros *transformers*. A inclusão dessa comparação poderia enriquecer a análise do desempenho do CDKT, fornecendo uma perspectiva mais abrangente em relacionada às capacidades e limitações do framework proposto.

Maity et al. (2023) conduziram uma estratégia para identificar e prevenir cometários com discurso de ódio em malaio [58]. Os autores elaboraram um corpus específico em malaio, denominado HateM. Esse corpus contém 4.892 textos que foram manualmente anotados. Ademais, os pesquisadores criaram um modelo denominado XLCaps, que é baseado em *deep learning*. Esse modelo combina o XLNet e o FastText com Bi-GRU, visando lidar efetivamente com os textos relacionados ao idioma malaio.

A iniciativa de criar um corpus dedicado em malaio destaca a importância de adaptar as ferramentas de detecção para considerar especificidades linguísticas e culturais. A aplicação do modelo XLCaps, que integra técnicas avançadas de *deep learning*, marca um avanço importante na procura por soluções eficazes para a problemática de textos com discurso de ódio. No entanto, vale destacar que nenhuma análise foi realizada em um contexto multilíngue. Sendo assim, não é possível saber se a eficiência dos resultados obtidos pode ser alcançada em um contexto com múltiplos idiomas, bem como com disparidades léxicas maiores.

Karim et al. (2021) empregaram técnicas de AM, *deep learning*, e *Pre-Trained Language Model* (PTLM) para detectar discurso de ódio em textos relacionados ao idioma bengalês [49]. Os dados compreendem diversas categorias, incluindo política, religião e geopolítica. Os autores exploraram diversas abordagens de classificação, como *Naive Bayes*, Regressão Logística, CNN, e PTLM. Os resultados mais promissores foram obtidos pelos modelos PTLMs: XLM-R, BERT pré-treinado em bengali e o BERT-Multilingual.

Nos experimentos conduzidos, os autores alcançaram uma medida-F1 de 88% como melhor métrica na classificação relacionada ao idioma bengalês. Contudo, o estudo não abordou a detecção relacionada a outros idiomas. Portanto, a aplicabilidade dos métodos e abordagens propostos pode exigir adaptações significativas quando aplicadas em relação a outros idiomas. Dessa forma, a falta de experimentos relacionados a outros idiomas pode comprometer os resultados alcançados.

Soto et al. (2021) empregaram uma modelo CNN com diferentes *embeddings* para efetuar a classificação de textos que apresentam discurso de ódio [90]. Os pesquisadores geraram *embeddings* referentes ao corpus denominado HSD, além de utilizarem *embeddings* provenientes do NILC [43]. Ademais, os autores testaram outros *embeddings*, incluindo Word2Vec [60], Glove [72] e FastText [48]. A combinação que rendeu o maior resultado foi: Glove de 300 dimensões, corpus HSD, bem como os *embeddings* do NILC. A diversidade de *embeddings* utilizados demonstra a abordagem empregada pelos autores na busca pelo melhor desempenho na classificação de textos. A escolha da combinação específica que resultou no melhor desempenho ressalta a importância de avaliar e selecionar cuidadosamente os componentes do modelo, considerando as características específicas do corpus e das tarefas em questão.

Duzha et al. (2021) fizeram experimentos para detectar discurso de ódio utilizando um corpus no idioma italiano [27]. Na pesquisa, os autores elaboraram um corpus manualmente anotado, composto por 1.264 textos coletados da rede social X. Além disso, utilizaram outros dois corpus, HaSpeeDe-tw e it-HS. Os resultados alcançados na classificação de discurso de ódio por meio de *transfer learning*, foram comparados entre dados políticos e não políticos, explorando a viabilidade da predição, mesmo ao empregar dados não vinculados a contextos políticos.

Os maiores resultados foram alcançados por modelos de *deep learning* fundamentados na arquitetura *transformer*, mostrando a eficácia de modelos fundamentados nesta arquitetura para esse domínio de pesquisa. No entanto, vale salientar que nenhuma análise referente a outros idiomas foi realizada, restringindo os resultados apenas ao idioma italiano. Sendo assim, não há como afirmar se o modelo proposto apresentará a mesma eficiência quando aplicado a outros idiomas.

Yohannes e Amagasa (2022) abordaram a problemática dos idiomas com recursos li-

mitados [107]. Os pesquisadores apresentaram uma estratégia para reconhecer entidades nomeadas em textos redigidos em Tigrinya, uma língua falada na Etiopia que possui recursos limitados. Eles empregaram um modelo fundamentado na arquitetura *transformer*, denominado TigRoBERTa. A eficiência desse modelo foi comparado com outros métodos existentes, tais como CRF, BiLSTM-CRF e BERT-multilíngue. Os resultados apontaram que TigRoBERTa superou todos esses métodos em acurácia, revocação e medida-F1. O estudo, no entanto, apresenta algumas limitações, como o tamanho reduzido dos dados abordados e a escassez de recursos linguísticos disponíveis para o Tigrinya, bem como a falta de análise para outros idiomas.

A Tabela 3.1 exibe um resumo comparativo dos principais tópicos referentes às pesquisas apresentadas.

Tabela 3.1: Sumário com os principais tópicos referentes aos trabalhos que fizeram detecção de discurso de ódio.

Autores	Uso de *transformers*	Técnica(s) de classificação	Há Análise da Diversidade Léxica
Waseem e Hovy (2016)	Não	Regressão Linear	Não
Davidson et al. (2017)	Não	Regressão Logística, Naive Bayes, *Decicion Tree*, *Random Forest* e SVM	Não
Frenda et al. (2019)	Não	SVM	Não
Del Arco et al. (2021)	Sim	BETO	Não
Mathew et al. (2019)	Não	DeGroot	Não
Yang et al. (2022)	Sim	CDKT	Não

Continua na próxima página

Tabela 3.1 – Continuação da página anterior

Autores	Uso de *transformers*	Técnica(s) de classificação	Há Análise da Diversidade Léxica
Maity et al. (2023)	Sim	XLCaps	Não
Karim et al. (2021)	Sim	XLM-R e BERT	Não
Soto et al. (2021)	Não	CNN	Não
Duzha et al. (2021)	Sim	SVM, *Random Forest* e AlBERTo	Não
Yohannes e Amagasa (2022)	Sim	TigRoBERTa	Não

3.3 Trabalhos Referentes a Cross-Lingual Learning

Um dos desafios no domínio de PLN é a necessidade de um corpus anotado em um idioma específico para realizar a classificação dos dados. O CLL representa uma técnica que se vale de corpora anotados de outros idiomas para a construção de modelos em PLN sem a necessidade de se ter um corpus anotado em um idioma específico. Esta técnica tem se tornado presente em várias tarefas de PLN, abrangendo desde tradução automatizada, classificação de sentimentos, bem como classificação semântica e morfológica [73]. Adicionalmente, a técnica CLL pode ser aplicada juntamente com modelos de AM para detectar discurso de ódio.

Stappen et al. (2020) empregaram o CLL como estratégia para detectar discurso de ódio em diversos idiomas [92]. Os autores incluíram partes do idioma de destino ao realizar o treino do modelo de AM. Como resultado, os pesquisadores alcançaram uma métrica de 69% no F-score. Nos experimentos, os autores combinaram as seguintes estratégias: a técnica chamada de *Attention-Maximum-Average Pooling* para classificação de dados, *embeddings* gerados pelo FastText e um modelo baseado em *transformer* (XLM ou BERT). Na fase de classificação foi aplicada uma ferramenta da Amazon para tradução das informações relacionadas ao idioma inglês. No entanto, é importante destacar que não foram implementadas estratégias para mitigar possíveis erros de tradução pela ferramenta, podendo assim

comprometer a interpretação dos resultados obtidos.

Pamungkas et al. (2021) apresentaram um método para detectar discurso de ódio em diversos idiomas utilizando a técnica CLL [71]. Os autores adotaram o idioma inglês para treino do modelo e outros seis idiomas foram escolhidos como idioma de destino: francês, indonésio, italiano, alemão, espanhol e português. Cada idioma de destino foi classificado individualmente. Diversos modelos foram avaliados, abrangendo desde abordagens tradicionais de AM, como a regressão logística, até modelos fundamentados em *transformers*, como o BERT. Técnicas como o *Joint-Learning* e *Zero-shot* foram aplicadas nos experimentos conduzidos. O modelo *Long Short-Term Memory* foi o mais eficiente. Este modelo utilizou *embeddings* multilíngues provenientes do projeto MUSE do Facebook [52].

A pesquisa destaca a importância da abordagem multilíngue ao efetuar detecção de discurso de ódio, demonstrando a viabilidade e eficácia do método proposto em diferentes contextos linguísticos. Adicionalmente, a análise comparativa entre modelos tradicionais e *transformers* contribui para a percepção das melhores práticas na aplicação de abordagens relacionadas à AM nesse domínio específico. Contudo, vale destacar que nenhuma análise foi realizada referente à inclusão de múltiplos idiomas no treino do modelo, bem como não foi realizada nenhuma análise referente à distância léxica dos idiomas empregados nos experimentos.

Schioppa et al. (2023) sugeriram o pré-treinamento de *Large Language Models* (LLMs) em combinação com o CLL [84]. Os autores adotaram em seus experimentos modelos baseados em *transformer*. A abordagem empregada combina modelagem auto-supervisionada e tradução automática supervisionada. Os autores destacaram que o CLL é capaz de aprimorar as habilidades de aprendizado contextual dos LLMs, especialmente para idiomas que possuem recursos limitados. Os resultados demonstraram que o método abordado superou boa parte dos modelos LLMs e modelos de tradução automática analisados pelos autores. Porém, nenhuma investigação foi realizada referente às disparidades léxicas dos idiomas empregados nos experimentos com os modelos.

Bigoulaeva et al. (2021) conduziram uma pesquisa para detectar discurso de ódio, destacando os desafios apresentados pelo vasto cenário das redes sociais [8]. Os pesquisadores utilizaram dados relacionados ao idioma inglês para treino e alemão para classificação. Os autores empregaram modelos de *deep learning*, integrando estratégias como *Joint Learning*,

Zero-shot e *Bilingual Word Embedding*. Contudo, é importante destacar que a pesquisa não realizou experimentos com corpus cuja distância léxica seja mais abrangente, o que pode impactar a generalização dos resultados para idiomas mais distantes linguisticamente. Ademais, nenhuma análise foi realizada referente ao uso de múltiplos idiomas para o treino do modelo.

Asai et al. (2023) apresentaram um novo método para medir a eficiência relacionada a modelos de AM pré-treinados em tarefas de CLL [5]. Os autores criaram um corpus denominado BUFFET e compararam alguns modelos de AM, como BERT, XLM-R e GPT-3, utilizando diferentes abordagens de adaptação, como engenharia de *prompts* e *fine-tunings*. Eles identificaram alguns fatores que influenciam a aprendizagem por meio de CLL quando aplicado em poucas amostras, como similaridade linguística, tamanho do vocabulário e qualidade dos *prompts*. Nenhuma investigação foi feita referente às disparidades léxicas entre os idiomas, bem como nenhuma análise foi realizada no que diz respeito ao uso de múltiplos idiomas no treino do modelo.

Corazza et al. (2020) sugeriram uma arquitetura de rede neural para detectar textos que contêm discurso de ódio em diversos idiomas, tais como inglês, italiano e alemão [18]. Os autores fizeram seus experimentos sob uma perspectiva arquitetural (LSTM e GRU) e extração de atributos (n-gramas, emojis e *word embeddings*). A melhor métrica obtida pelos autores foi uma medida-F1 de 82,3% para detectar discurso de ódio em inglês. Os autores identificaram a importância de vários componentes, como os embeddings, a aplicação de recursos adicionais (baseados em texto ou emoção), a normalização de emojis e hashtags, para a elaboração de modelos mais eficazes para detectar discurso de ódio.

Adicionalmente, os pesquisadores relatam que fizeram a transcrição dos emojis para texto em inglês por meio de uma biblioteca Python e posteriormente utilizaram a ferramenta de tradução do Google para traduzir os termos para alemão e italiano. Porém, não mencionaram o nível de acerto dessa tradução, podendo haver erros na tradução para os idiomas alemão e italiano. A aplicação de tradução nos experimentos impossibilita compreender a eficiência do modelo quando confrontado com dados sem o uso desse recurso. Ademais, a falta de comparação com modelos multilíngues adicionais limita a percepção da eficácia da proposta nesse tipo de modelo. Por fim, a análise do comportamento referente ao modelo não foi realizada quando confrontado com dados que têm uma distância léxica mais ampla.

Mozafari et al. (2022) sugeriram uma abordagem voltada a CLL para detectar discurso de ódio, bem como textos ofensivos, em múltiplos idiomas [65]. A abordagem adotada se chama *Model-Agnostic Meta-Learning* (MAML). Em resumo, MAML é um método que possibilita a realização do treino do modelo em dados que contemplem características relacionadas e, posteriormente, ajustado rapidamente para novas tarefas com poucos exemplos de treinamento, resultando em uma rápida generalização e adaptação a novos cenários de aprendizado. Os autores empregaram como experimento base o modelo XLM-R, este modelo foi empregado como uma referência para comparar a eficiência dos resultados aplicando a estratégia proposta (MAML). Os autores chegaram a conclusão que o método proposto demonstrou resultados melhores quando comparados ao modelo base XLM-R.

Apesar dos resultados terem sido superiores ao XLM-R, nenhuma comparação foi realizada com outros modelos codificadores ou modelos baseados em decodificadores para verificar se os resultados se mantinham superiores. Além disso, a abordagem da estratégia sugerida necessita de um corpora extenso o que pode se tornar uma complicação em situações onde não é possível obter corpora suficiente. Adicionalmente, nenhuma análise foi realizada sobre o impacto da distância léxica dos idiomas utilizados nos experimentos.

Zia et al. (2022) conduziram um experimento fundamentado na abordagem *Zero-Shot Learning* para detectar discurso de ódio em múltiplos idiomas [110]. Para esse propósito, os autores empregaram um modelo codificador baseado em *transformers* denominado XLM-R. A pesquisa concentra-se em transferência de aprendizado *zero-shot*, utilizando um idioma para treino do modelo e outro idioma de destino para classificação. Os autores apontaram em seus resultados que o método demonstrou melhorias no desempenho na detecção realizada pelo modelo.

No entanto, apesar de utilizarem mais de um idioma, os autores empregaram apenas o inglês no treinamento do modelo. Assim, a eficiência do modelo com outros idiomas permanece incerta. Além disso, não foi realizada nenhuma análise sobre o impacto nos resultados referente à distância léxica dos idiomas utilizados. Ademais, apenas o modelo codificador XLM-R foi abordado nos experimentos. Assim, permanece incerto se os resultados alcançados com o *Zero-Shot Learning* seriam replicáveis ao utilizar outros modelos codificadores ou modelos com arquiteturas baseadas em decodificadores.

Nozza (2021) abordou a classificação de discurso de ódio em múltiplos idiomas, desta-

cando as limitações dos modelos nesse tipo de tarefa [67]. Nos experimentos, foi utilizada uma adaptação do modelo BERT denominado mBERT. Os pesquisadores submeteram o modelo à estratégia CLL chamada *Zero-Shot* para verificar se existe aprimoramento do modelo. No geral, os resultados apontaram melhorias em certos casos dependendo do idioma adotado como alvo.

No estudo também foi apontado que abordagens para detectar discurso de ódio em múltiplos idiomas são desafiadoras devido às diferenças linguísticas, tendo sido ressaltada a importância de considerar cuidadosamente essas nuances ao projetar modelos para detectar discurso de ódio, bem como a necessidade de abordagens mais complexas para tratar essa questão. Contudo, vale destacar que o estudo possui algumas limitações, como a falta de comparação com outros modelos codificadores, bem como com modelos decodificadores. Ademais, não foi explorada e nem analisada a eficiência do modelo com respeito à distância léxica entre os idiomas utilizados, assim o estudo apresenta uma lacuna referente aos resultados do modelo quando empregado em idiomas que tenham uma distância léxica mais ampla.

Jiang e Zubiaga (2021) sugeriram um modelo voltado a CLL para detectar discurso de ódio em múltiplos idiomas [47]. O modelo, denominado CCNL-Ex, demonstrou um desempenho significativo para dados em italiano, inglês e espanhol. A abordagem utilizou uma rede chamada *Capsule Network* juntamente com CLL para melhorar a capacidade do modelo em detectar discurso de ódio. Foi utilizado um idioma para treino do modelo e um idioma de destino para classificação. No entanto, o idioma que foi elencado como destino foi traduzido especificamente para o mesmo idioma do modelo. Essa tradução foi realizada por meio de um tradutor automático da Google. Para medir a eficiência do modelo foram efetuados experimentos base com outros modelos tais como BERT, XLM-R, SVM, dentre outros.

Os experimentos comprovaram que o CCNL-Ex superou os modelos base. Apesar dos resultados, o trabalho possui algumas limitações relacionadas à tradução dos dados, pois a necessidade desse recurso para poder lidar com múltiplos idiomas pode introduzir erros na análise, dependendo da qualidade da tradução utilizada. Adicionalmente, nenhuma análise foi realizada referente à eficiência do modelo sem a aplicação da tradução. Portanto, não é possível saber a eficiência do modelo quando confrontado com idiomas sem serem traduzi-

dos. Ademais, nenhuma análise foi realizada quanto ao impacto e complexidade empregada ao modelo referente à distância léxica dos idiomas.

A Tabela 3.2 exibe um resumo comparativo dos principais tópicos referentes às pesquisas revisadas sobre CLL.

Tabela 3.2: Sumário das pesquisas relacionadas à CLL.

Autores	Uso de *transformers*	Técnica(s) de classificação	Há Análise da Diversidade Léxica
Stappen et al. (2020)	Sim	XLM e BERT	Não
Pamungkas et al. (2021)	Sim	BERT e LSTM	Não
Schioppa et al. (2023)	Sim	mT5	Não
Bigoulaeva et al. (2021)	Não	BiLSTM	Não
Asai et al. (2023)	Sim	BERT, XLM-R e GPT-3	Não
Corazza et al. (2020)	Não	LSTM e GRU	Não
Mozafari et al. (2022)	Sim	MAML e XLM-R	Não
Zia et al. (2022)	Sim	XLM-R	Não
Nozza et al. (2021)	Sim	mBERT	Não
Jiang e Zubiaga (2021)	Sim	CCNL-Ex, BERT, XLM-R, SVM	Não

3.4 Considerações Finais

Neste capítulo, foram apresentados os trabalhos relacionados ao mesmo tema de pesquisa desta dissertação de mestrado, evidenciando alguns desafios nessa área de pesquisa. Adicionalmente, foram apresentados trabalhos que utilizam CLL para detectar discurso de ódio, mostrando sua eficácia, especialmente quando utilizado em modelos baseados em *transformer*. A Figura 3.1 mostra o resumo geral dos trabalhos relacionados por categoria.

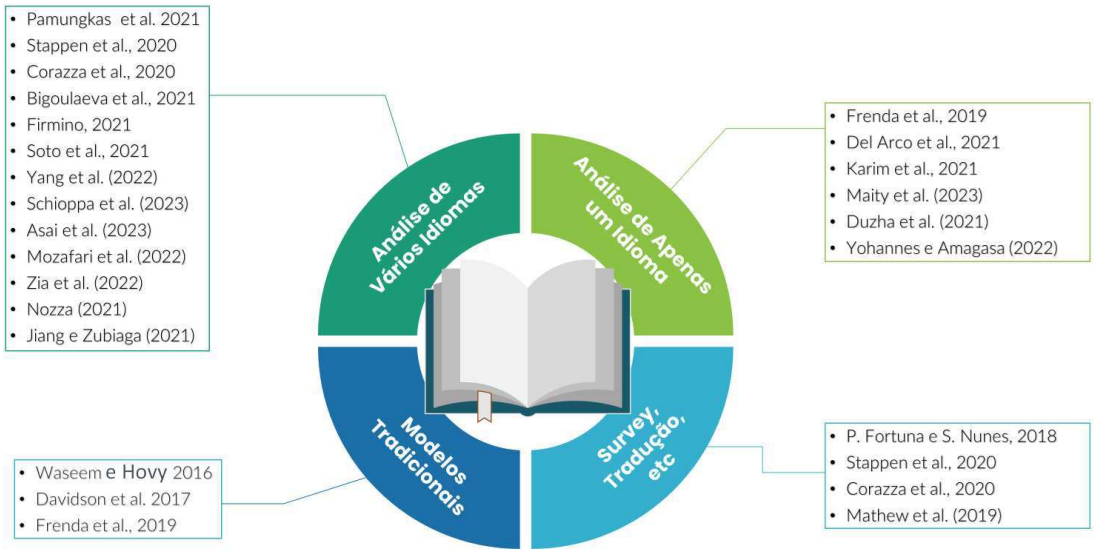

Figura 3.1: Resumo dos trabalhos relacionados por categoria.

Notavelmente, boa parte das pesquisas relacionadas concentram-se em idiomas que apresentam considerável proximidade léxica. Contudo, nenhum deles tentou abordar experimentos considerando uma maior diversidade léxica entre os dados, bem como a utilização de múltiplos idiomas no treinamento do modelo. Ademais, não houve comparações relacionadas ao resultado de modelos codificadores e decodificadores quando aplicados em dados que possuam distância léxica mais abrangente.

A análise da distância léxica é um fator importante ao utilizar CLL, pois, dependendo dos idiomas empregados no treinamento, pode facilitar a transferência de aprendizado para o modelo. Isso pode resultar em uma classificação mais eficiente. Além disso, o uso de múltiplos idiomas permite que o modelo seja adaptado a diferentes idiomas de destino, proporcionando uma detecção eficaz mesmo em idiomas com recursos textuais escassos. Por fim, a análise da distância léxica permite entender quais idiomas são mais adequados para melhorar a eficiência do modelo, considerando as nuances linguísticas e culturais específicas de cada idioma, reduzindo assim a ocorrência de falsos positivos e negativos.

Portanto, este presente trabalho visa suprir esses aspectos identificados. O próximo capítulo aborda os materiais e métodos empregados neste trabalho, bem como a metodologia adotada para realizar os experimentos.

Capítulo 4

Materiais e Métodos

Este capítulo aborda os materiais e a metodologia utilizada para detectar discurso de ódio em textos por meio de modelos baseados na arquitetura de *transformers*. Além disso, são apresentadas as informações relacionadas aos corpora utilizados nos experimentos desta dissertação. Esta seção está estruturada da seguinte maneira: Na Seção 4.1, é detalhada a metodologia empregada; Na Seção 4.2, descreve-se cada corpus utilizado nos experimentos; Na Seção 4.3, são apresentados os modelos de IA utilizados; Na Seção 4.4, são apresentadas as técnicas utilizadas nos experimentos; Na Seção 4.5, é descrita a forma de avaliação dos modelos. Por fim, a Seção 4.6 contempla as considerações finais deste capítulo.

4.1 Metodologia

A detecção eficaz de discurso de ódio demanda a implementação de uma série de etapas metodológicas. No âmbito desta dissertação, uma das etapas envolve a obtenção de corpora anotados em diversos idiomas, essenciais para realização do treino, validação, assim como o teste dos modelos.

Adicionalmente, destaca-se a etapa concernente à utilização de modelos baseados em AM, os quais são treinados para identificar padrões associados à presença de discurso de ódio nos textos analisados. Nesse contexto, optou-se pela utilização de modelos fundamentados na arquitetura de *transformers*, reconhecida por sua eficácia em tarefas complexas de PLN.

Além da aplicação direta de modelos fundamentados em AM, a pesquisa também concentra-se em detectar discurso de ódio empregando a estratégia CLL. Este enfoque de-

manda a implementação de técnicas específicas, as quais têm por finalidade aprimorar a eficácia dos modelos adotados para classificação. Ademais, foram empregadas métricas para medir a eficiência dos modelos em detectar discurso de ódio. Por fim, os resultados provenientes dos modelos foram analisados. A Figura 4.1 apresenta uma síntese visual da metodologia adotada, ilustrando as diferentes etapas e interações entre os elementos envolvidos. Inicialmente, temos a obtenção de corpora. Em seguida, os corpora são submetidos aos modelos de IA. Por conseguinte, os modelos são submetidos às estratégias de CLL, a saber: Zero-Shot Transfer (ZST), Joint Learning (JL), Cascade Learning (CL), Joint Learning/Cascade Learning (JL/CL) e Joint Learning/Cascade Learning+ (JL/CL+). Para cada estratégia, são coletadas as métricas de avaliação dos modelos e, por fim, os resultados são analisados.

No escopo das próximas seções, cada componente da metodologia adotada será detalhado, visando proporcionar uma compreensão mais abrangente do processo metodológico empregado nesta dissertação.

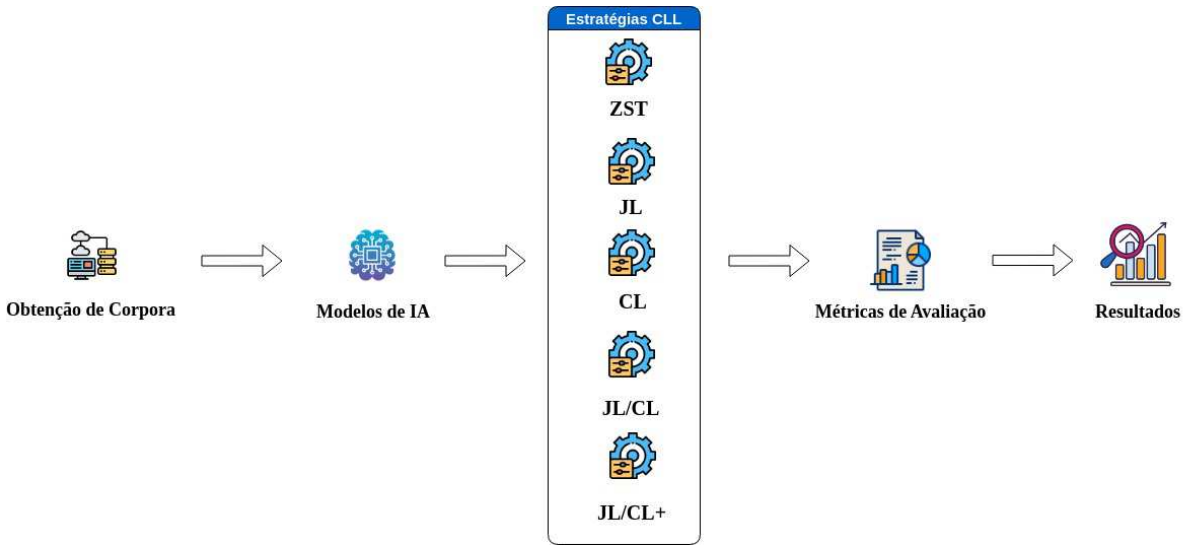

Figura 4.1: Representação da metodologia adotada neste trabalho.

4.2 Obtenção de Corpora

A etapa inicial da metodologia adotada neste estudo consiste na obtenção de corpora para a elaboração das fases de treino, validação, assim como o teste dos modelos propostos. Diver-

sos métodos estão disponíveis para obter os dados necessários, e escolher o procedimento adequado pode variar conforme a estratégia adotada.

Um método para obtenção dos dados envolve a coleta direta por meio de ferramentas computacionais especializadas, seguida pela subsequente rotulagem dos textos adquiridos. Esta abordagem permite uma adaptação mais direta aos objetivos específicos da pesquisa, fornecendo controle sobre o procedimento de coleta e anotação, embora exija um investimento de recursos computacionais e tempo.

Alternativamente, outra estratégia viável é usar corpora previamente coletado em pesquisas anteriores realizadas por diferentes autores. Sendo assim, não é necessário realizar a coleta de dados por ferramentas computacionais. Esta abordagem, embora menos personalizada, oferece a vantagem da disponibilidade imediata de dados já anotados e prontos para uso, economizando recursos e tempo.

No desenvolvimento desta pesquisa, foi realizada a coleta de dados referentes ao idioma português, os quais foram subsequentemente submetidos a uma anotação manual. Os dados foram obtidos utilizando um *crawler*[1]. Esse *crawler* foi empregado para coletar dados da rede social X. Posteriormente, os dados foram manualmente rotulados como contendo discurso de ódio ou não contendo discurso de ódio. Todavia, dado que o escopo desta dissertação demanda a incorporação de múltiplos corpora referentes a distintos idiomas, foi necessário utilizar dados preexistentes de contribuições de outros autores [99; 41; 11; 89; 93; 37]. Desta maneira, optou-se por empregar ambos os métodos de obtenção de corpora, envolvendo dados provenientes de outros estudos relacionados, bem como a coleta direta dos dados.

Dessa forma, a coleta de dados direcionados ao idioma português e sua subsequente anotação manual visaram garantir a especificidade e relevância dos dados para os objetivos desta pesquisa. Já a necessidade de abranger diversos idiomas demandou a incorporação de dados provenientes de outras pesquisas relacionadas, enriquecendo a diversidade linguística e contextual dos corpora utilizados nesta dissertação. Portanto, a combinação destes métodos de obtenção de dados permitiu uma abordagem abrangente e multilíngue, essencial para a eficácia da análise de discurso de ódio em diferentes contextos culturais e linguísticos.

Além disso, para atender aos requisitos de escopo desta pesquisa, foram selecionados

[1] https://github.com/twintproject/twint

dados exclusivamente relacionados ao campo político. Adicionalmente, além de atender aos critérios específicos desta pesquisa, a escolha de dados vinculados a um único campo, facilita a especialização do modelo, permitindo uma melhor transferência de aprendizado atrelado ao discurso de ódio presente em textos. Dessa forma, proporciona ao modelo a aprendizagem de padrões específicos desse domínio. Adicionalmente, essa estratégia facilita a transferência de aprendizado entre diferentes idiomas por meio da técnica CLL, pois os dados compartilham um contexto comum relacionado ao mesmo domínio (política), contribuindo para uma melhor eficiência do modelo em detectar discurso de ódio em um contexto multilíngue.

Os dados empregados neste estudo abarcam múltiplos idiomas, a saber: português, espanhol, italiano, filipino, inglês, turco e alemão. A Tabela 4.1 contempla os detalhes específicos relativos a cada corpus utilizado. Nas subseções seguintes, são apresentados detalhes mais específicos de cada corpus, delineando características de cada um deles para o entendimento e análise do corpus utilizado nesta dissertação.

Tabela 4.1: Detalhes dos dados empregados nesta dissertação.

Corpus e/ou Idioma	Textos Sem Discurso de Ódio	Textos Com Discurso de Ódio	Total de Textos	Total de Palavras
Evalita 2018 (italiano)	1.941 (48,5%)	2.059 (51,5%)	4.000	46.009
WASSA 2021 (inglês)	2.648 (88,36%)	352 (11,64%)	3.000	76.131
POLLY (alemão)	4.318 (98,5%)	62 (1,5%)	4.380	94.583
MisoCorpus (espanhol)	2.272 (56,96%)	1.717 (43,04%)	3.989	92.458
Filipino	9.864 (53,42%)	8.600 (46,58%)	18.464	266.067
Turco	5.055 (66,80%)	2.512 (33,20%)	7.567	169.563
Português	2.860 (89,37%)	340 (10,63%)	3.200	47.787

4.2.1 Corpus no Idioma Inglês

Grimminger e Klinger (2021) desenvolveram um corpus intitulado WASSA 2021 [41]. Este corpus é constituído por 3.000 textos em inglês, os quais foram extraídos de plataformas de redes sociais. Os dados englobam comentários postados por usuários durante o período da eleição presidencial dos Estados Unidos em 2020 entre os candidatos Biden e Trump. Os textos foram rotulados manualmente, sendo que 352 textos foram rotulados como contendo discurso de ódio, enquanto 2.648 foram rotulados como não contendo discurso de ódio. A Tabela 4.2 mostra exemplos de sentenças com e sem discurso de ódio desse corpus.

Tabela 4.2: Exemplos de textos com e sem discurso de ódio no corpus inglês.

	Texto em Inglês	Tradução Livre
Textos com discurso de ódio	"Joe Biden is a disgusting vile sick individual"	"Joe Biden é um indivíduo nojento, vil e doente."
	"...fuck you maga culties!"	"...Vão se foder, seus cultistas. MAGA!"
	"I really can't with this motherfucker anymore."	"Não aguento mais esse filho da puta."
	"You are a disgusting and disrespectful individual."	"Você é um indivíduo nojento e desrespeitoso."
	"Anyone who supports Joe Biden is a complete moron!"	"Qualquer pessoa que apoia Joe Biden é um completo idiota."
Textos sem discurso de ódio	"If Joe Biden is 0.00000000000000001% better I will vote for him."	"Se Joe Biden for 0.00000000000000001%, eu irei votar nele."
	"To the new stewards of the highest elected offices... Congrats Joe and Kamala."	"Parabéns aos novos administradores dos mais altos cargos eleitos: Joe e Kamala."

Continua na próxima página

Tabela 4.2 – Continuação da página anterior

Textos sem discurso de ódio	Texto em Inglês	Tradução Livre
	“God be with us!”	“Deus esteja conosco!”
	“Happy birthday know that you’re as bright and as brilliant as the candles on your cake.”	“Feliz aniversário, saiba que você é tão brilhante e brilhante quanto as velas do seu bolo.”
	“I know I agree entirely. He said greatest president and i wanted to know why.”	“Eu sei que concordo inteiramente. Ele disse o melhor presidente e eu queria saber por quê.”

A Figura 4.2 apresenta três nuvens de palavras, exibindo a ocorrência das palavras no corpus. A subfigura (a) ilustra as cem palavras de textos sem discurso de ódio que possuem mais ocorrências, revelando algumas palavras de caráter neutro, como *“time”* (tempo), *“election”* (eleição), *“president”* (presidente), entre outras.

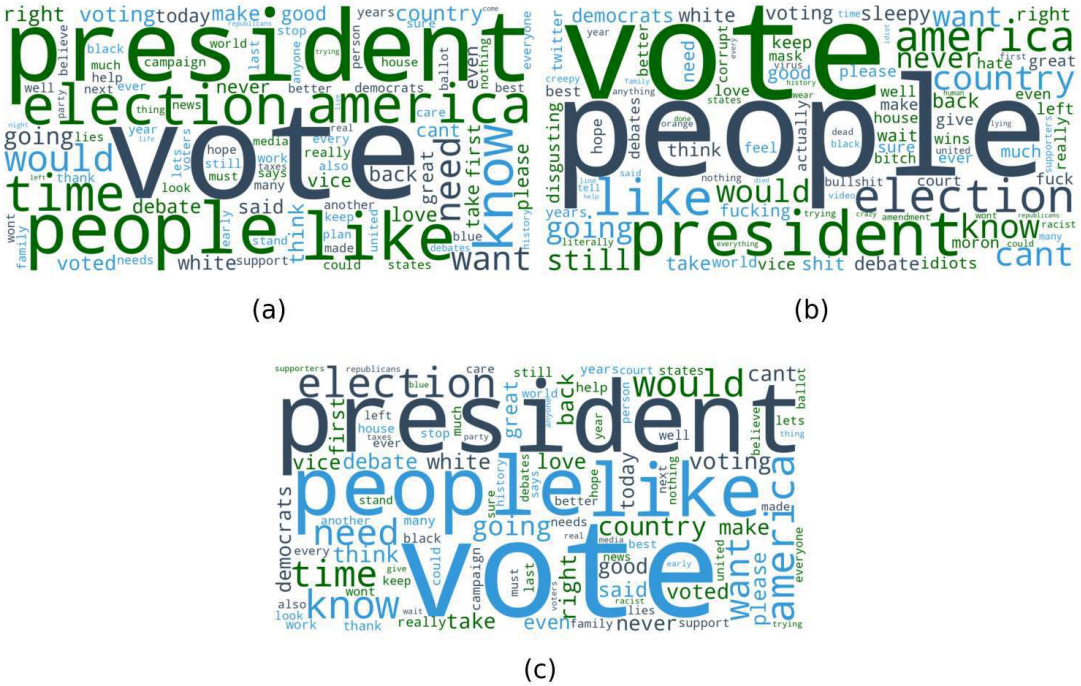

Figura 4.2: As cem palavras com maior ocorrência no corpus inglês.

Em contrapartida, a subfigura (b) exibe as cem palavras com mais ocorrências em textos com discurso de ódio, revelando termos de baixo calão, como *"fuck"* (foda-se), *"disgusting"* (nojento), *"moron"* (idiota), entre outros. A subfigura (c) destaca as cem palavras com maior ocorrência no corpus todo, evidenciando como maior ocorrência as palavras *"people"* (pessoas), *"vote"* (voto), *"president"* (presidente), entre outras.

A Figura 4.3 apresenta uma visão quantitativa dos textos do corpus. Apesar da quantidade de textos que não possuem discurso de ódio ser superior a quantidade contendo discurso de ódio, observa-se que boa parte dos textos que não possuem discurso de ódio em inglês estão em um intervalo situado entre 10 e 30, enquanto os textos com discurso de ódio predominantemente estão no intervalo entre 10 e 40, ou seja, com tamanhos ligeiramente superiores comparativamente aos textos que não contêm discurso de ódio.

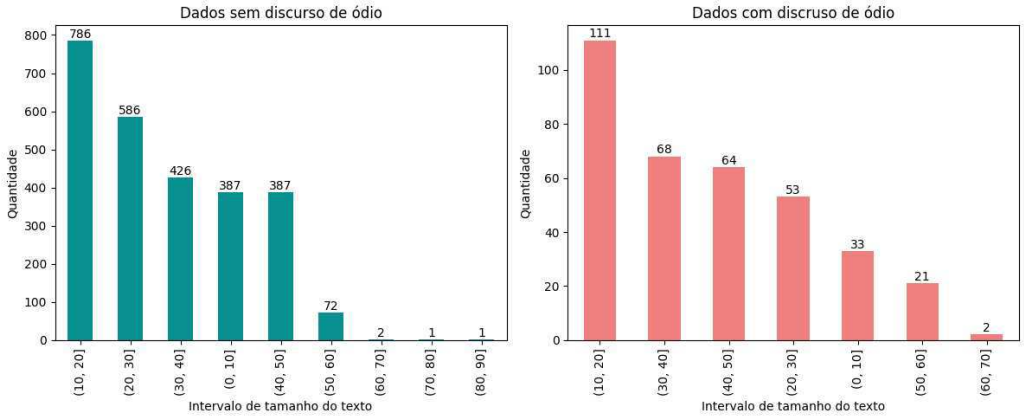

Figura 4.3: Intervalo do tamanho das sentenças em inglês.

Essa observação é corroborada pela Figura 4.4, que apresenta a média do tamanho das sentenças presentes no corpus. Nota-se uma média ligeiramente superior nos textos que contêm discurso de ódio comparado aos textos sem discurso de ódio, evidenciando uma diferença média de 2,91 palavras a mais para os textos com discurso de ódio. Esses resultados indicam que, neste corpus específico, indivíduos propensos a expressar discurso de ódio em inglês, relacionado à política, têm preferência por redigir textos mais extensos quando comparados àqueles que produzem textos sem discurso de ódio.

A Figura 4.5 apresenta um *boxplot* do corpus em inglês e seu desvio padrão. Observa-se que o desvio padrão é relativamente alto em comparação à média, tanto nas sentenças com

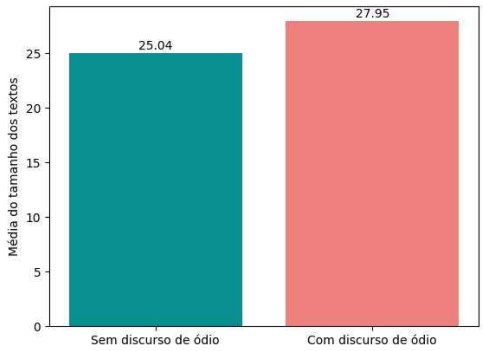

Figura 4.4: Média do tamanho das sentenças em inglês.

discurso de ódio quanto nas sentenças sem discurso de ódio. Um ponto a ser destacado é que, nas sentenças sem discurso de ódio, é possível identificar uma quantidade maior de *outliers* em comparação com as sentenças com discurso de ódio. Adicionalmente, as sentenças com discurso de ódio apresentam um maior tamanho em relação às sentenças sem discurso de ódio, pois o terceiro quartil está próximo de 40, enquanto o terceiro quartil das sentenças sem discurso de ódio encontra-se em um valor menor do que 40.

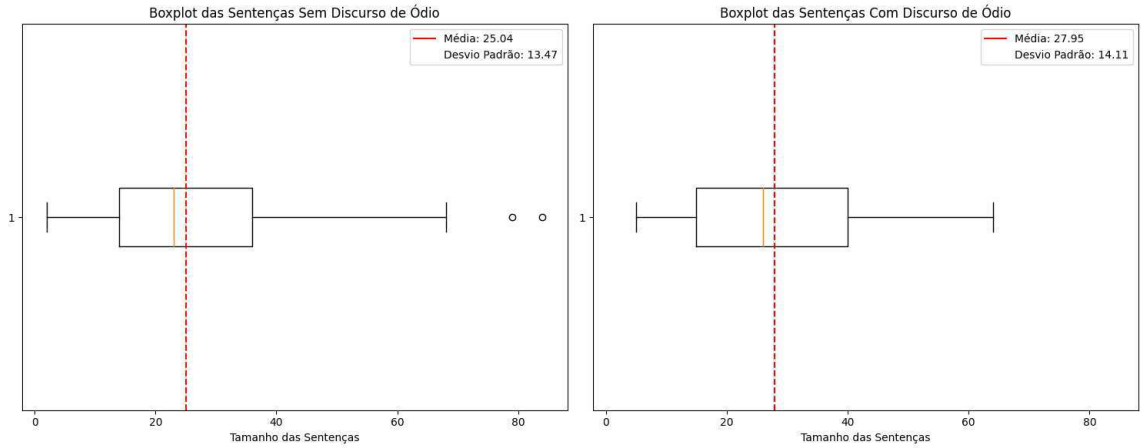

Figura 4.5: *Boxplot* das sentenças em inglês.

4.2.2 Corpus no Idioma Italiano

O corpus denominado Evalita 2018 contempla textos provenientes dos comentários feitos por usuários italianos da rede social Facebook. O corpus possui 17.000 textos, no total. A

coleta desse corpus foi conduzida pelo *Istituto di Informatica e Telematica* em Pisa [99]. No entanto, apenas parte desse corpus foi rotulada. Sendo assim, dentro desse corpus, encontram-se 1.941 textos rotulados como não contendo discurso de ódio, contrastando com 2.059 textos rotulados como apresentando discurso de ódio. Vale destacar que todas as anotações foram realizadas manualmente. A Tabela 4.3 contempla exemplos de textos com e sem discurso de ódio desse corpus.

Tabela 4.3: Exemplos de textos com e sem discurso de ódio no corpus italiano.

	Texto em Italiano	Tradução Livre
Textos com discurso de ódio	"Sono insopportabili per quello che dicono perchè non appartengono alla vita che noi cittadini italiani..."	"São insuportáveis pelo que dizem porque não pertencem à vida que nós, cidadãos italianos..."
	"Maledetti porci schifosi parassiti e ne ho ancora tanti."	"Malditos porcos parasitas nojentos e ainda tenho muitos mais."
	"La Malpezzi è come la rozza blee schifo."	"Malpezzi é um nojento rude."
	"Mamma mia che schifo il partito del PD il gruppo dei somari."	"Meu Deus, o partido do PD, o grupo de burros, é nojento."
Textos sem discurso de ódio	"Ti do una notizia .....assolutamente noooooo."	"Vou te dar algumas novidades.....absolutamente nãooooo."
	"D'accordo ma Salvini non era un Ministro, era un semplice parlamentare."	"Ok, mas Salvini não era ministro, era um simples parlamentar."
	"Ci provo capitano. ..grazie a Dio sei arrivato."	"Vou tentar, capitão. ..graças a Deus você chegou."

Continua na próxima página

Tabela 4.3 – Continuação da página anterior

Textos sem discurso de ódio	Texto em Italiano	Tradução Livre
	“No, soprattutto la prima.”	“Não, especialmente o primeiro.”

A Figura 4.6 contempla as nuvens de palavras relacionadas ao corpus no idioma italiano, onde a subfigura (a) corresponde aos dados sem discurso de ódio, a subfigura (b) aos dados com discurso de ódio, e a subfigura (c) representa o corpus completo. Cada nuvem apresenta as cem palavras com maior ocorrência. Na subfigura (a) é possível observar com maior frequência palavras neutras como *“voto”* (voto), *“lavoro”* (trabalhar), *“solo”* (sozinho) e *“grande”* (ótimo). Adicionalmente, na subfigura (b), que representa as palavras presentes nos textos com discurso de ódio, observa-se algumas palavras que denotam insatisfação, tais como *“odiose”* (odiáveis), *“schifo”* (nojo) e *“insopportabili”* (insuportáveis).

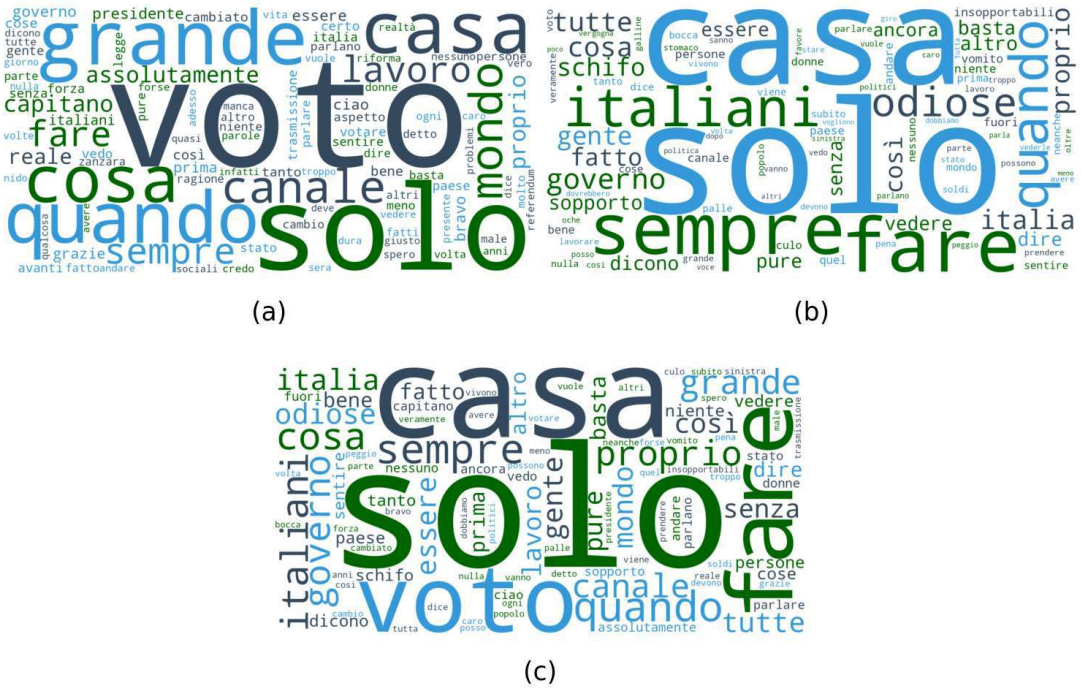

Figura 4.6: As cem palavras com maior ocorrência no corpus de idioma italiano.

A Figura 4.7 proporciona uma visão abrangente dos textos presentes no corpus de idioma italiano. Observa-se que boa parte dos textos desprovidos de discurso de ódio concentram-se

em um intervalo de tamanho compreendido entre 0 e 10. Já os textos que contêm discurso de ódio predominantemente situam-se no mesmo intervalo, embora exibam uma concentração maior no intervalo entre 10 e 20, quando comparados aos textos que não apresentam discurso de ódio no mesmo intervalo.

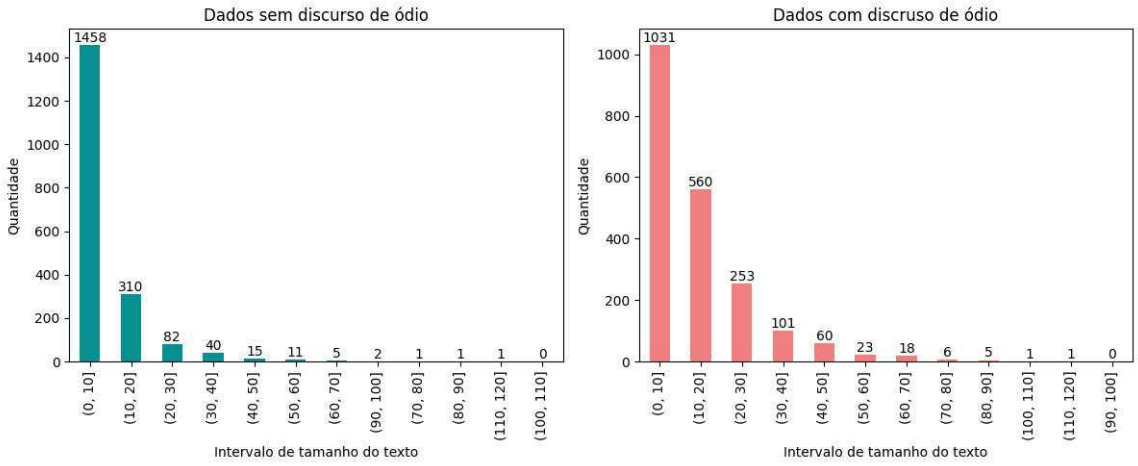

Figura 4.7: Intervalo do tamanho das sentenças em italiano.

A Figura 4.8 ilustra a média do tamanho das sentenças no corpus de idioma italiano. Observa-se que a média do tamanho dos textos contendo discurso de ódio é praticamente o dobro da média do tamanho dos textos sem discurso de ódio, revelando uma diferença média de 6,18 palavras adicionais nos textos com discurso de ódio. Essa observação sugere que, neste corpus específico, indivíduos que manifestaram discurso de ódio em seus textos tendem a redigir textos muito mais extensos em comparação àqueles que produzem textos sem discurso de ódio.

A Figura 4.9 apresenta um *boxplot* para as sentenças do corpus em italiano. Nota-se que o desvio padrão é relativamente alto, indicando uma variação significativa no tamanho das sentenças nesse corpus. Outro ponto a ser destacado é que há uma quantidade maior de *outliers* nas sentenças sem discurso de ódio em comparação com as sentenças com discurso de ódio. Adicionalmente, percebe-se que o tamanho das sentenças com discurso de ódio é maior, pois o terceiro quartil vai até o valor 20, enquanto nas sentenças sem discurso de ódio o terceiro quartil é inferior a 20.

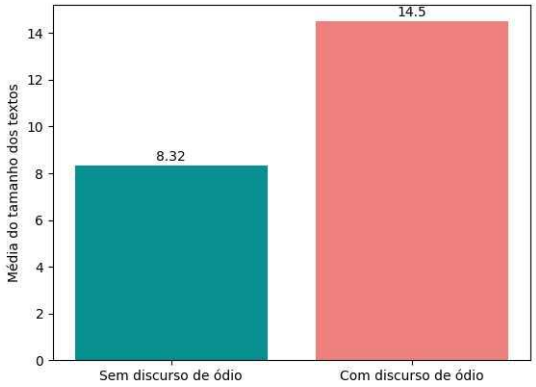

Figura 4.8: Média do tamanho das sentenças em italiano.

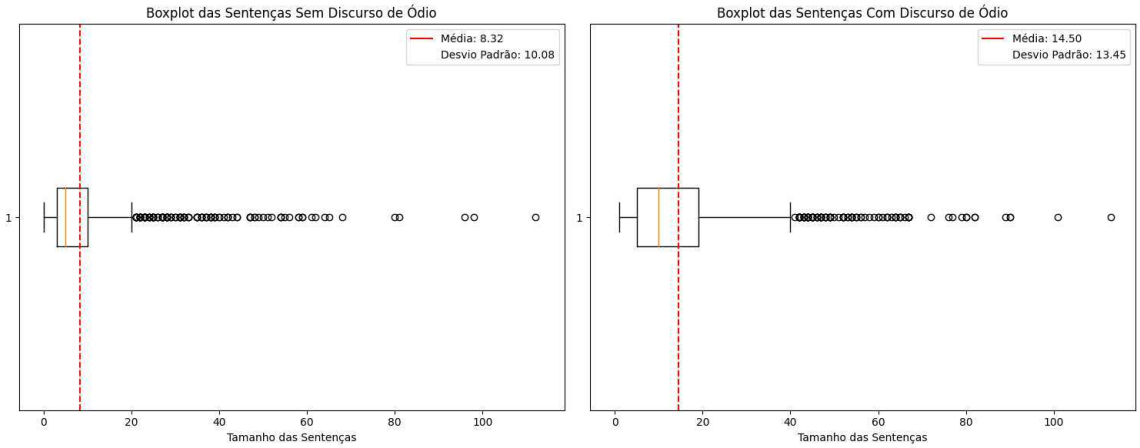

Figura 4.9: *Boxplot* das sentenças no corpus italiano.

4.2.3 Corpus no Idioma Filipino

Cabasag et al. (2019) elaboraram um corpus anotado manualmente no idioma filipino [11]. Este corpus consiste em textos originados de algumas redes sociais, coletados durante o período das eleições presidenciais nas Filipinas em 2016. O corpus é constituído por 8.600 textos que foram categorizados como contendo discurso de ódio, enquanto 9.864 textos foram designados como não contendo discurso de ódio. A Tabela 4.4 mostra exemplos de sentenças com e sem discurso de ódio nesse corpus.

Tabela 4.4: Exemplos de textos com e sem discurso de ódio no corpus filipino.

	Texto em Filipino	Tradução Livre
Textos com discurso de ódio	“putangina mo binay TAKBO PA.”	“sua mãe é uma prostituta, Binay, corra ainda.”
	“Wtf? Tangina mo Binay nagpapaawa kapa! Hahahahaha Nognog??”	“Que porra é essa? Filho da puta, Binay, está se fazendo de coitado! Hahahahaha??”
	“Tangina naman yung ad ni Binay nakabwisit!”	“Filho da puta, o anúncio do Binay é irritante!”
	“Sobrang kapal ng mukha ni Binay. Putangina.”	“Binay é extremamente cara de pau. Filho da puta.”
	“Tangina ni binay talaga.”	“Filho da puta do Binay, realmente.”
Textos sem discurso de ódio	“Salamat sa walang sawang suporta ng mga taga Makati! Ang Pagbabalik Binay In Makati.”	“Obrigado pelo apoio incansável das pessoas de Makati! O retorno de Binay em Makati.”
	“Walang drama. Walang hugot. Trabaho lang. At higit sa lahat, hindi tayo pagnanakawan ni.”	“Sem drama. Sem puxar. Apenas trabalho. E acima de tudo, não seremos roubados.”
	“Buti pa si Binay hahahaha.”	“Binay ainda é bom hahahaha.”
	“Walang hindi importante kay VP Binay.”	“Nada é sem importância para o VP Binay.”

Continua na próxima página

Tabela 4.4 – Continuação da página anterior

Textos sem discurso de ódio	Texto em Filipino	Tradução Livre
	“Pagsuporta ng Gobernador ng Ifugao kay VP Binay, pinasalamatan ng UNA.”	“Apoio do governador de Ifugao ao VP Binay, UNA agradeceu.”

A Figura 4.10 apresenta as nuvens de palavras que fazem referência às cem palavras mais frequentes no corpus em filipino. As subfiguras (a), (b) e (c) referem-se, respectivamente, aos dados sem discurso de ódio, com discurso de ódio e a todo o corpus. É possível notar uma influência do idioma inglês no filipino, devido à presença de palavras como *“president”* (presidente), *“vote”* (voto), *“only”* (somente), etc. Ademais, na subfigura b, que representa as palavras presentes em textos com discurso de ódio, são observadas palavras de insulto, como *“tangina”* (filho de uma prostituta), *“nognog”* (termo racista referente a pessoas negras) e *“corrupt”* (corrupto).

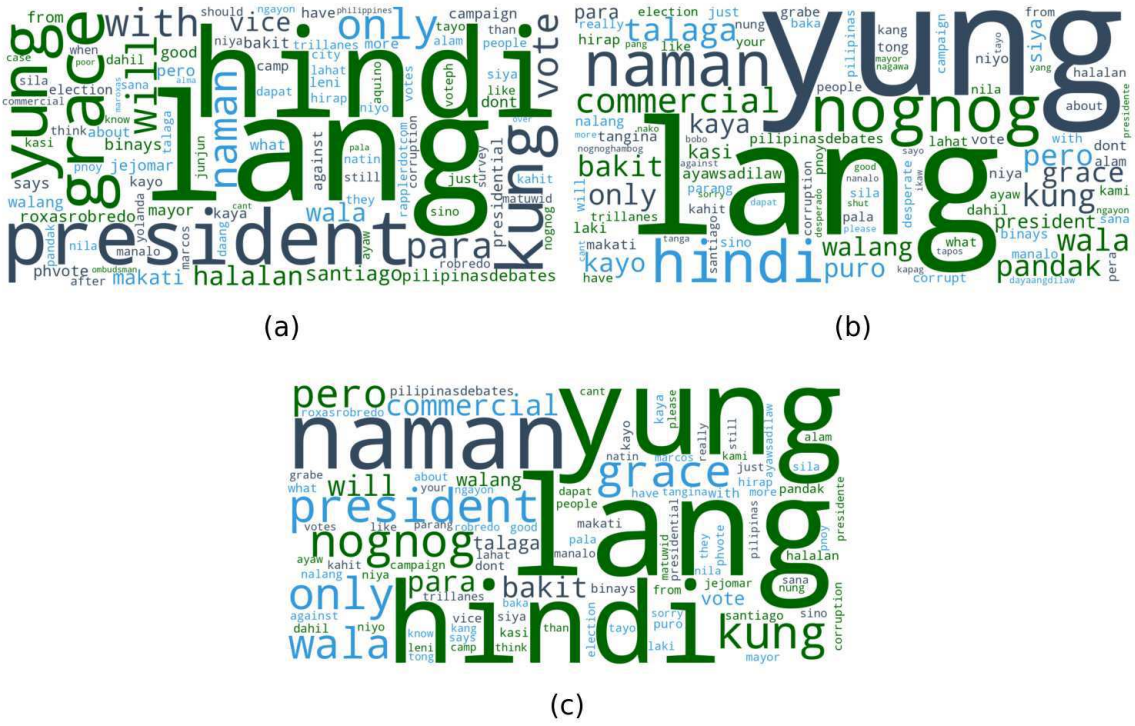

Figura 4.10: As cem palavras com maior ocorrência no corpus filipino.

A Figura 4.11 apresenta a frequência do tamanho das sentenças no corpus. Observa-se que tanto os textos com discurso de ódio quanto os textos sem discurso de ódio estão predominantemente concentrados no intervalo entre 10 e 20. No entanto, as sentenças com discurso de ódio exibem uma quantidade ligeiramente menor no intervalo entre 0 e 10.

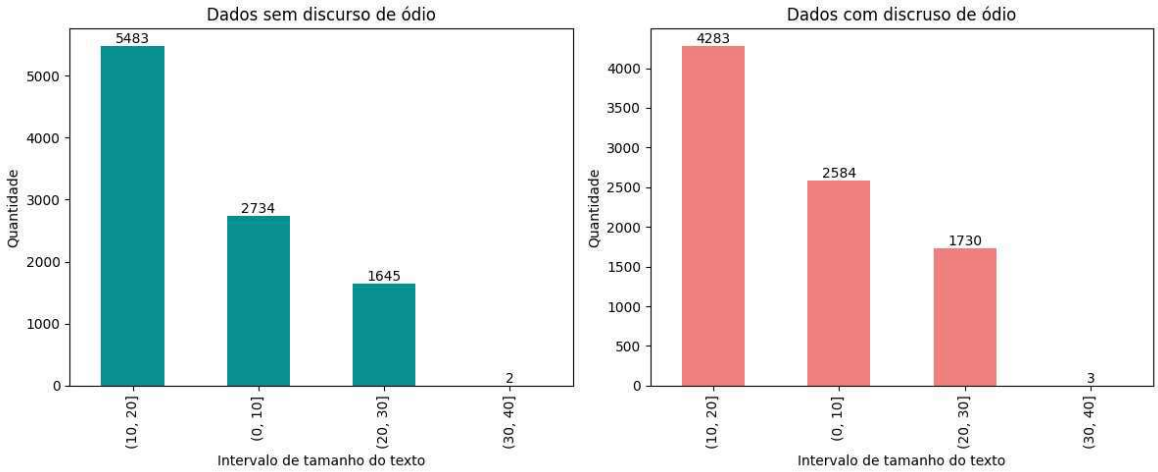

Figura 4.11: Intervalo do tamanho das sentenças em filipino.

Na Figura 4.12, observa-se que a média do tamanho das sentenças, tanto as que possuem discurso de ódio quanto as que não possuem, apresentam praticamente o mesmo valor. Em contraste com os dois corpus anteriores (italiano e inglês), os dados referentes ao idioma filipino exibiram uma proporção mais equilibrada referente ao tamanho das sentenças, diferenciando-se da tendência observada nos idiomas anteriormente apresentados.

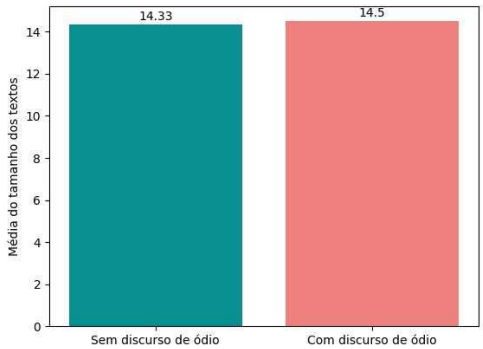

Figura 4.12: Média do tamanho das sentenças em filipino.

A Figura 4.13 exibe o *boxplot* e o desvio padrão para as sentenças em filipino. Percebe-se um valor mediano para o desvio padrão, indicando uma menor variação das sentenças em comparação aos idiomas analisados anteriormente. No entanto, diferentemente dos idiomas anteriores, não há *outliers* nas sentenças sem e com discurso de ódio, existindo um equilíbrio na dispersão do tamanho das sentenças em comparação aos idiomas anteriores. Ademais, assim como no idioma italiano, o tamanho das sentenças com discurso de ódio é maior, pois o terceiro quartil está próximo de 20, enquanto o das sentenças sem discurso de ódio é inferior a 20.

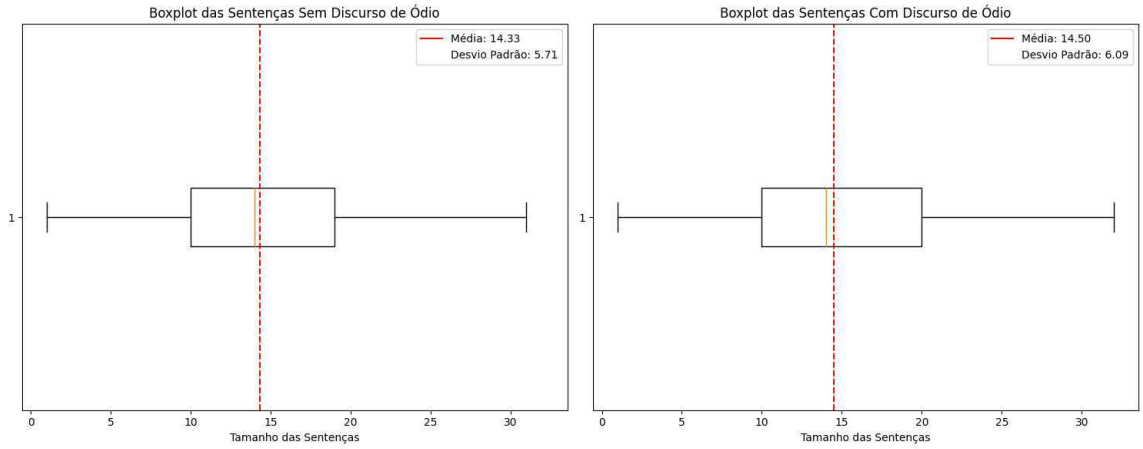

Figura 4.13: *Boxplot* das sentenças em filipino.

4.2.4 Corpus no Idioma Alemão

Smedt et al. (2018) produziram o corpus POLLY [89]. Os dados que o constituem foram coletados durante as eleições alemãs de 2017, abrangendo um total de 125.000 textos relacionados à política. No âmbito desta pesquisa, optou-se por empregar uma versão re-anotada do corpus POLLY, fornecida pelo projeto *German Hate Speech Corpus* da Universidade de Würzburg. Esta versão revisada está disponível publicamente [104] e compreende 4.318 textos categorizados sem discurso de ódio, enquanto 62 textos foram designados como apresentando discurso de ódio. A Tabela 4.5 exibe exemplos de sentenças com e sem discurso de ódio nesse corpus.

Tabela 4.5: Exemplos de textos com e sem discurso de ódio no corpus alemão.

	Texto em Alemão	Tradução Livre
Textos com discurso de ódio	“Hahahahaha, was ist das? Ich ficke diese Huren. Er ist ein Junkie, Hurensohn.”	“Hahahahaha, o que é isso? Eu fodo com esses filhos da puta, ele é um viciado, filho da puta.”
	“Wer ist der Hurensohn da links? Bieber Felix kenne ich, aber wer ist dieser Sportlehrer? Hahahah.”	“Quem é o filho da puta ali à esquerda? Conheço o Felix, mas quem é esse professor de educação física? Hahahah.”
	“Hahahahh, bist du blöd? Kein Chef kündigt dich. Dein Chef ist ein Hurensohn, er braucht dich...”	“Hahahahh, você é burro? Seu chefe não vai te demitir, ele é um filho da puta e precisa de você...”
	“Hahahahh, ich schwöre, der Hurensohn ist kein Mensch.”	“Hahahahh, eu juro que aquele filho da puta não é humano.”
	“Sieht wie ein Hurensohn aus.”	“Parece um filho da puta.”
Textos sem discurso de ódio	“Und wieder nur verbale Propagandamittel.”	“E, novamente, apenas meios verbais de propaganda.”
	“Dafür ist es wohl zu spät.”	“Provavelmente é tarde demais para isso.”

Continua na próxima página

Tabela 4.5 – Continuação da página anterior

	Texto em Alemão	Tradução Livre
Textos sem discurso de ódio	"Vielleicht bauen eure importierten Fachkräfte ja ein paar Lehmhütten."	"Talvez os seus trabalhadores qualificados importados construam algumas cabanas de barro."
	"Hunderttausenden Versorgungs Suchenden gefällt das."	"Centenas de milhares de pessoas procuram suprimentos como este."
	"Ja, das bildet ihr euch ein."	"Sim, você está imaginando."

A Figura 4.14 apresenta as nuvens de palavras que fazem referência às cem palavras com maior ocorrência sem discurso de ódio (a), com discurso de ódio (b) e para todo o corpus (c).

Figura 4.14: As cem palavras com maior ocorrência no corpus em alemão.

Na nuvem que abrange palavras com discurso de ódio, são observadas palavras de teor ofensivo, como *"hurnsohn"* (filho da puta), *"nigger"* (termo racista contra pessoas negras), *"fck"* (abreviação da palavra foda-se), entre outras. Em contraste, nas nuvens de palavras sem discurso de ódio (a) e na nuvem que engloba todos os textos (c), são apresentadas palavras mais neutras. As palavras com maior frequência incluem *"deutschland"* (Alemanha), *"immer"* (sempre) e *"mehr"* (mais).

Esse corpus é o mais desbalanceado entre os que foram abordados nesta pesquisa. A Figura 4.15 apresenta os intervalos de tamanhos das sentenças. Percebe-se que o tamanho dos dados sem discurso de ódio está concentrado no intervalo entre 10 e 30, enquanto os dados que possuem discurso de ódio estão concentrados no intervalo entre 10 e 40.

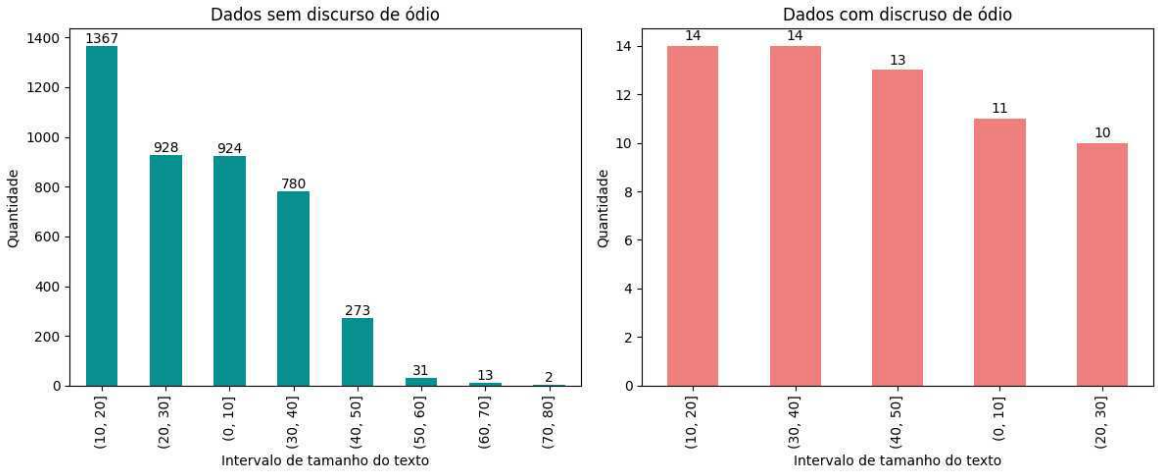

Figura 4.15: Intervalo do tamanho das sentenças em alemão.

Apesar da quantidade de dados contendo discurso de ódio ser menor, é possível notar que a média do tamanho das sentenças é maior quando comparada à média dos textos sem discurso de ódio, conforme ilustrado na Figura 4.16. Essa observação sugere que os usuários alemães que expressaram discurso de ódio em seus textos têm uma propensão a redigi-los de maneira mais extensa em comparação aos usuários que não apresentaram discurso de ódio em seus textos. A Figura 4.17 apresenta o *boxplot* e o desvio padrão das sentenças em alemão. Percebe-se uma quantidade maior de *outliers* nas sentenças sem discurso de ódio em comparação com as sentenças com discurso de ódio. Observa-se um desvio padrão relativamente alto, indicando uma variação considerável no tamanho das sentenças. Assim

como nos idiomas anteriores, as sentenças com discurso de ódio aparentam ser maiores, já que o terceiro quartil vai até o valor 40, enquanto o terceiro quartil das sentenças sem discurso de ódio vai até o valor 30.

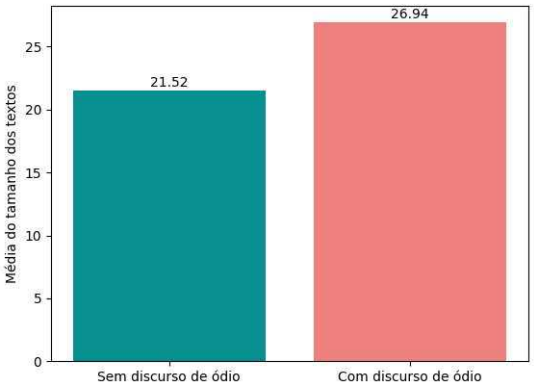

Figura 4.16: Média do tamanho das sentenças em alemão.

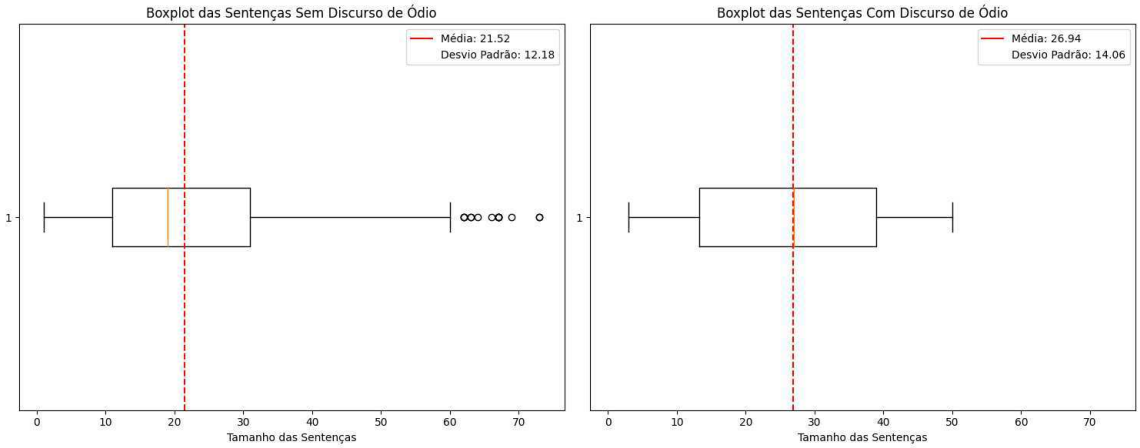

Figura 4.17: *Boxplot* das sentenças em alemão.

4.2.5 Corpus no Idioma Turco

Toraman et al. (2022) elaboraram um corpus no idioma turco, o qual engloba dados compostos por mais de 100.000 textos coletados de algumas redes sociais [93]. Esses textos abrangem diversos domínios temáticos, como política, religião, gênero e esporte. No escopo desta pesquisa, adotou-se o subconjunto centrado no domínio político, composto por 2.512

instâncias rotuladas como contendo discurso de ódio e 5.055 instâncias como não apresentando discurso de ódio. A Tabela 4.6 contempla exemplos de sentenças com e sem discurso de ódio no corpus turco.

Tabela 4.6: Exemplos de textos com e sem discurso de ódio no corpus turco.

	Texto em Turco	Tradução Livre
Textos com discurso de ódio	"Başta Kılıçdaroğlu, Erdoğan düşmanlığı sizi terörist yaptı, terörist. Allah belanızı versin!!"	"A hostilidade para com Erdoğan fez de você um terrorista, que Deus o amaldiçoe!!"
	"İslam düşmanı katil."	"Assassino islamofóbico."
	"Katil Trump ve akıl hocasının derisini yüzeceğiz."	"Vamos esfolar o assassino Trump e seu mentor."
	"Tabi ki yasalara saygımız sonsuz ama bunların cezası tutukluk yapmayan bir silah ve bir mermi olmalıdır."	"É claro que temos muito respeito pela lei, mas a punição deles deveria ser uma arma e uma bala que não emperre."
	"Benim gibi hırsız, katil, ahlaksız ve onursuz bir vatan haini domuzdan güzel ülkem kurtulacak..."	"Meu lindo país será salvo de um ladrão, assassino, porco traidor imoral e desonroso..."
Textos sem discurso de ódio	"Hakkari'de PKK'lı teröristlere ait silah ve mühimmat ele geçirildi."	"Armas e munições pertencentes a terroristas do PKK foram apreendidas em Hakkari."

Continua na próxima página

Tabela 4.6 – Continuação da página anterior

	Texto em Turco	Tradução Livre
Textos sem discurso de ódio	"Evde parti yapanı yakalamak yerine, keşke silah atanları yakalasaydınız."	"Eu gostaria que você tivesse pego aqueles que estavam jogando armas em vez de pegar aqueles que estavam festejando em casa."
	"Silah sıkanlar parti var diye ihbar edesim var, muhbir damgası yerim diye tırsıyorum."	"Tenho que denunciar quem atira porque tem festa, mas tenho medo de ser rotulado de informante."
	"Yılbaşı gecesi saatler 00.00 gösterdiğinde camların önünde silahla bekleyen tek ülke Türkiye'dir."	"A Turquia é o único país que espera com armas em frente às suas janelas às 00h00 da véspera de Ano Novo."
	"Her yerde silah sesi, yeni yıla silahla giren tek ülke biziz herhalde."	"Ouvem-se tiros por todo o lado. Acho que somos o único país que começou o novo ano com armas."

A Figura 4.18 apresenta as nuvens de palavras referentes às cem palavras com maior ocorrência sem discurso de ódio (a), com discurso de ódio (b) e todo o corpus (c). Em todos eles, notam-se palavras frequentes como *"haini"* (traidor), *"vatan"* (pátria) e *"terörist"* (terrorista). Nas palavras com discurso de ódio (b), observam-se algumas palavras negativas, como *"erefsiz"* (sem vergonha), *"katil"* (assassino), *"orospu"* (prostituta), *"tecavüz"* (estupro), *"haini"* (traidor), *"terörist"* (terrorista), entre outras.

Figura 4.18: As cem palavras com maior ocorrência no corpus em turco.

A Figura 4.19 apresenta o intervalo de tamanho das sentenças em turco. Observa-se que boa parte dos textos, tanto os com discurso de ódio quanto os sem discurso de ódio, está concentrada no intervalo entre 10 e 40.

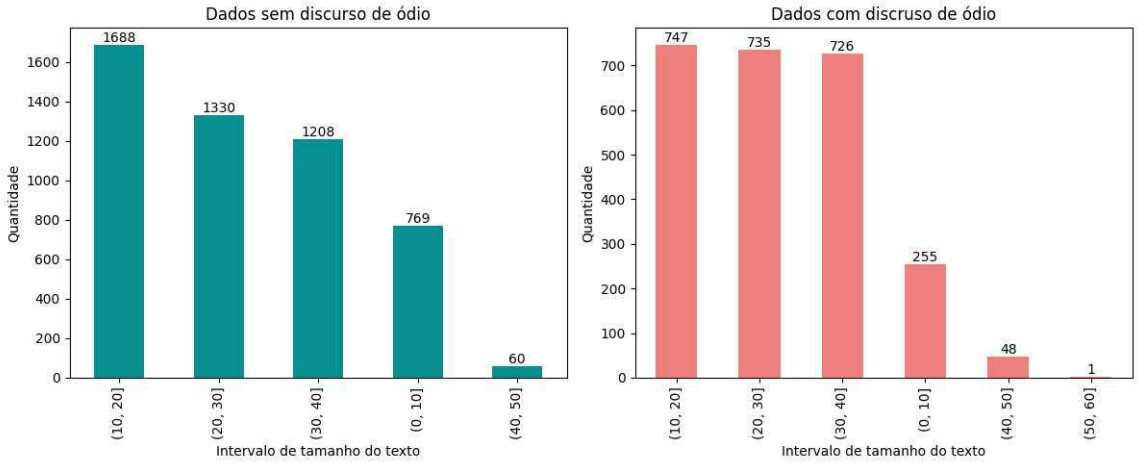

Figura 4.19: Intervalo do tamanho das sentenças em turco.

No entanto, ao analisar separadamente cada gráfico, percebemos que proporcionalmente o total de sentenças com discurso de ódio é maior no intervalo de 20 a 40 em comparação aos dados sem discurso de ódio. Na Figura 4.20, observa-se que a média do tamanho das sentenças com discurso de ódio é ligeiramente maior, o que sugere uma propensão dos usuários turcos que praticam discurso de ódio a escreverem textos mais extensos.

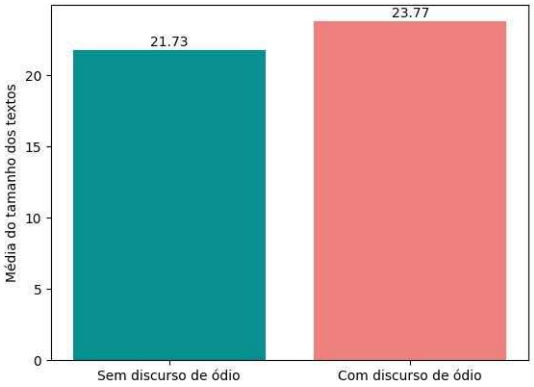

Figura 4.20: Média do tamanho das sentenças em turco.

A Figura 4.21 mostra o *boxplot* e o desvio padrão para as sentenças do corpus em turco. Nota-se um desvio padrão moderado, indicando uma variação relativa das sentenças em relação à média.

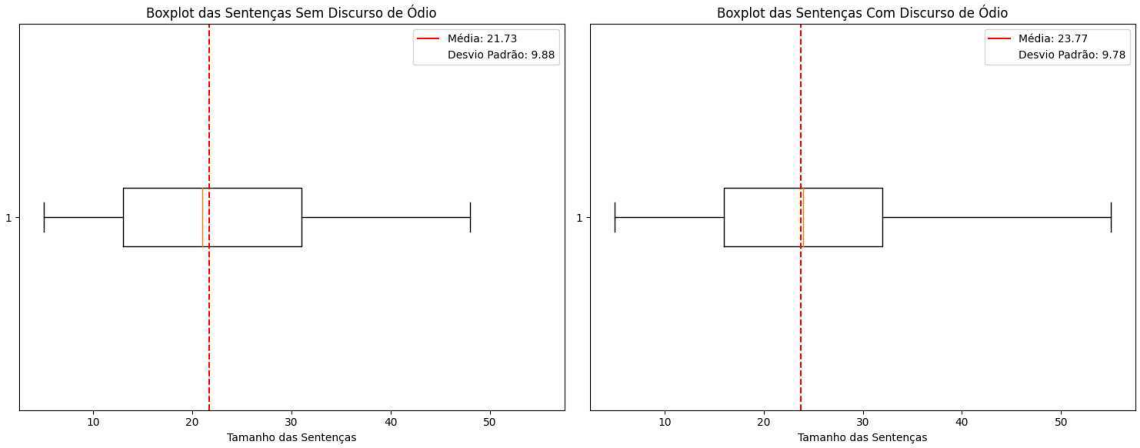

Figura 4.21: *Boxplot* das sentenças em turco.

Outro ponto a destacar é o tamanho das sentenças. Assim como nos idiomas anteriores, as sentenças com discurso de ódio apresentam um tamanho maior em relação às sentenças sem discurso de ódio, pois o terceiro quartil está entre 30 e 40, enquanto o das sentenças sem discurso de ódio está no valor 30.

4.2.6 Corpus no Idioma Espanhol

Díaz et al. (2021) elaboraram um corpus em espanhol denominado MisoCorpus [37]. Esse corpus contempla textos coletados de algumas redes sociais que foram submetidos posteriormente para serem anotados manualmente. Os autores dividiram os dados em três subconjuntos distintos. O primeiro subconjunto engloba dados relacionados a discurso de ódio na esfera política, direcionados especificamente ao público feminino. O segundo subconjunto concentra-se em dados que tem como objetivo serem utilizados para analisar as diferenças linguísticas entre o espanhol falado na Europa e o falado na América Latina. Por fim, o terceiro subconjunto representa um corpus utilizado para analisar mensagens associadas a traços gerais de misoginia. Esta dissertação faz uso exclusivamente do primeiro subconjunto que possui 2.272 textos rotulados como não contendo discurso de ódio e 1.717 textos rotulados como contendo discurso de ódio. A Tabela 4.7 exibe exemplos de sentenças com e sem discurso de ódio para o corpus espanhol.

Tabela 4.7: Exemplos de textos com e sem discurso de ódio no corpus espanhol.

	Texto em Espanhol	Tradução Livre
Textos com discurso de ódio	“Labsoluta hija de la gran puta descerebrada alcohólica y porrera a la señora a la madre del hijo de dios no la ofende ud en mi presencia.”	“Filha absoluta do grande alcoólatra desmiolado e puta vagabunda, a senhora, mãe do filho de Deus, não a ofenda na minha presença.”
	“Hija de zorra.”	“Sua filha de uma vadia.”
	“Y la ZORRA eres TÚ.”	“A vadia é você.”

Continua na próxima página

Tabela 4.7 – Continuação da página anterior

	Texto em Espanhol	Tradução Livre
Textos com discurso de ódio	"La que faltaba la puta abortista feminazi dando otra la lata."	"A que faltava, a puta abortista feminazi dando trabalho de novo."
	"Tu ni para puta vales."	"Você nem pra ser puta serve."
	"Calla ya perra, si no vales ni para esconderte, no tienes ni puta gracia y tenemos que subvencionarte."	"Cala a boca, cadela, você nem serve para se esconder, não tem graça nenhuma e ainda temos que te sustentar."
Textos sem discurso de ódio	"Tú luchas contra el cambio climático? Yo lucho por llegar a fin de mes."	"Você luta contra as mudanças climáticas? Eu luto para sobreviver."
	"La esperanza para enfrentar la crisis climática viene de las personas."	"A esperança de enfrentar a crise climática vem das pessoas."
	"Ayer entré por primera vez en mi vida en Primark. Compré. Perdóname Greta."	"Ontem entrei na Primark pela primeira vez na minha vida. Eu comprei. Perdoe-me Greta."
	"Los mejores memes de Greta Thunberg antes de su llegada a la Cumbre del Clima de Madrid."	"Os melhores memes de Greta Thunberg antes de sua chegada à Cúpula do Clima de Madri."
	"El mundo necesita a una Greta Thunberg, pero no de esta manera."	"O mundo precisa de uma Greta Thunberg, mas não desta forma."

A Figura 4.22 apresenta as cem palavras com maior ocorrência no corpus, subdivididas nas subfiguras (a), (b) e (c) referentes, respectivamente, aos textos sem discurso de ódio, com discurso de ódio e todo o corpus.

Figura 4.22: As cem palavras com maior ocorrência no corpus em espanhol.

Notavelmente a palavra *"puta"* é a mais frequente em todas as subfiguras, caracterizando um termo de insulto. Adicionalmente, nos textos com discurso de ódio, observam-se várias palavras de insulto e linguagem ofensiva, tais como *"patética"* (patética), *"payasa"* (palhaça), *"loca"* (lunática), *"puta"* (puta), *"zorra"* (vadia), entre outras.

Na Figura 4.23, observa-se que as sentenças sem discurso de ódio está concentrada, em sua maioria, no intervalo de 10 a 30, enquanto as sentenças com discurso de ódio predominam no intervalo de 0 a 20. Ao analisar a média do tamanho das sentenças na Figura 4.24, destaca-se que as sentenças sem discurso de ódio apresentam tamanhos maiores. Esse padrão sugere que, nesse contexto específico, os usuários espanhóis têm uma propensão a produzir textos mais longos quando não possuem discurso de ódio.

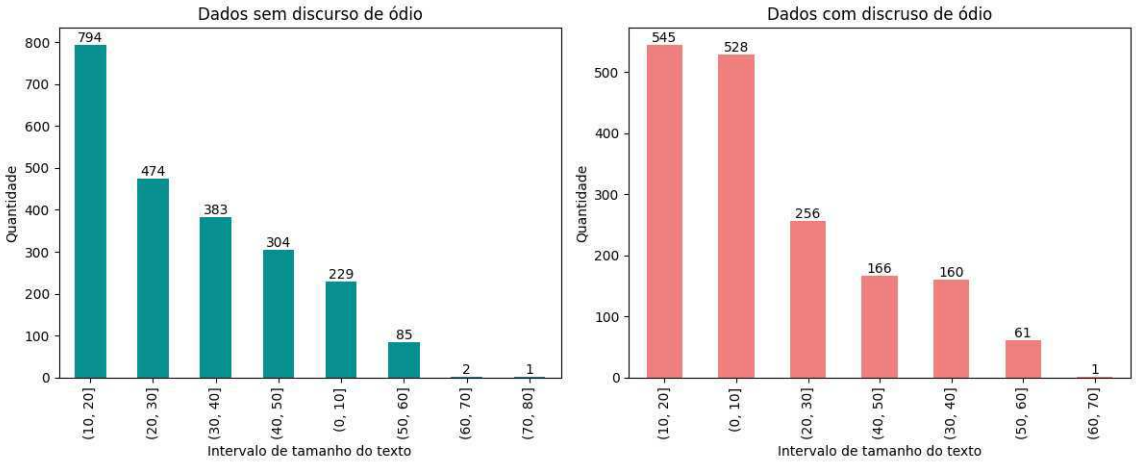

Figura 4.23: Intervalo do tamanho das sentenças em espanhol.

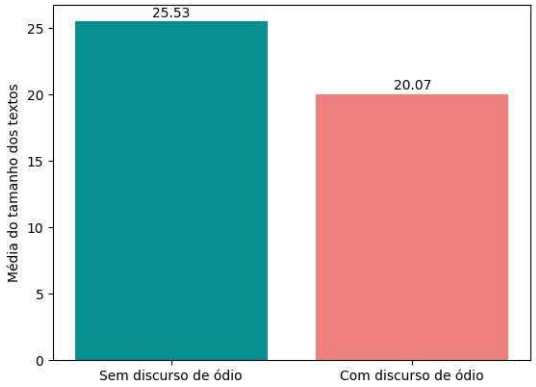

Figura 4.24: Média do tamanho das sentenças em espanhol.

A Figura 4.25 contempla o *boxplot* e o desvio padrão das sentenças em espanhol. Assim como na maioria dos idiomas analisados anteriormente, o desvio padrão apresenta um valor relativamente alto, indicando uma variação considerável em relação à média. No entanto, diferentemente dos outros idiomas, é possível perceber que o tamanho das sentenças em espanhol sem discurso de ódio é superior ao das sentenças com discurso de ódio, pois o terceiro quartil está localizado acima do valor 30, enquanto o das sentenças com discurso de ódio está localizado em um valor menor do que 30.

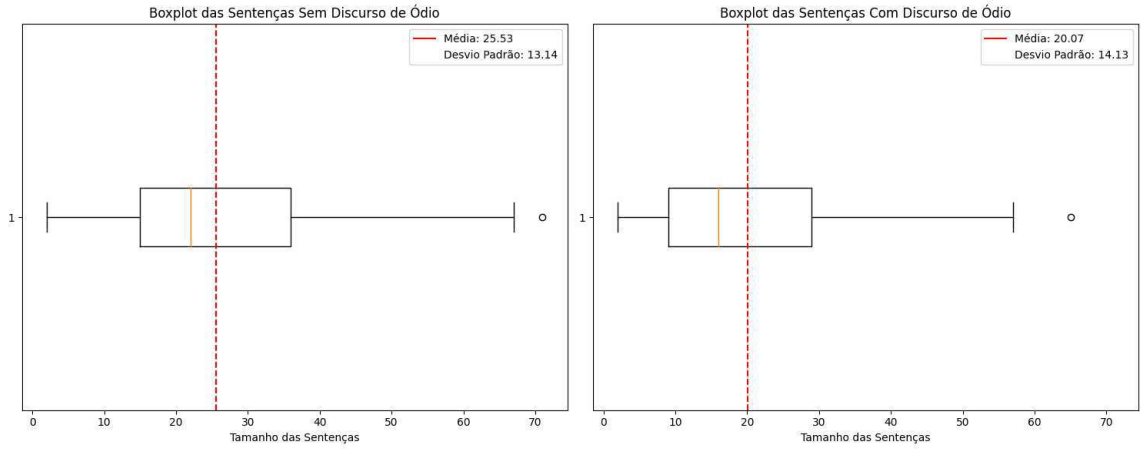

Figura 4.25: *Boxplot* das sentenças em espanhol.

4.2.7 Corpus no Idioma Português

Esse corpus foi criado pelo autor desta dissertação. Os dados foram coletados da rede social X durante as eleições presidenciais no Brasil em 2022 por meio de um *crawler*[2]. Para efetuar a rotulação, adotou-se o seguinte processo: duas pessoas procederam à rotulação dos dados, categorizando-os como contendo ou não discurso de ódio. Esse procedimento de rotulação foi conduzido de maneira individual, garantindo que cada anotador não tivesse conhecimento das anotações realizadas pelo outro. Em situações em que ocorreu discordância nas anotações entre os dois anotadores, uma terceira pessoa foi incumbida de efetuar a anotação, sem ter acesso às anotações dos outros dois anotadores.

Essa metodologia de anotação dupla, seguida de arbitragem por uma terceira pessoa em casos de discordância, buscou assegurar uma avaliação imparcial do conteúdo, contribuindo para a confiabilidade e validade do corpus utilizado neste trabalho. O corpus resultante compreende um total de 3.200 textos, dos quais 2.860 foram rotulados como não contendo discurso de ódio, enquanto 340 textos foram rotulados como contendo discurso de ódio. A Tabela 4.8 contempla exemplos de sentenças com e sem discurso de ódio desse corpus.

[2]https://github.com/twintproject/twint

Tabela 4.8: Exemplos de textos com e sem discurso de ódio no corpus português.

	Texto em Português
Textos com discurso de ódio	"A cara do mentiroso da república nem arde!!! Canalha!!! Mentiroso."
	"Além de ser um escroto, mentiroso, dissimulado, egocêntrico, inescrupuloso, etc."
	"Ladrão! Mentiroso! Vagabundo!"
	"Chega destes vagabundos, malditos! Lá!!!!"
Textos sem discurso de ódio	"Nordeste tá vindo aí para nos salvar."
	"Lutar pela nossa Democracia e pela nossa Liberdade!!!"
	"Só esperando esse momento!"
	"Nordeste salva a pátria."

A Figura 4.26, contempla as palavras referentes às cem palavras com maior ocorrência, apresentando as subfiguras a, b e c que representam, respectivamente, os dados com sentenças sem discurso de ódio, com discurso de ódio e todo o corpus. Nota-se que as palavras "presidente" e "povo" aparecem com maior frequência. Na subfigura "a" e "c", observa-se a prevalência de palavras mais neutras, tais como "votar", "povo", "turno", "vamos", "governo", entre outras. Conforme esperado, devido à natureza do corpus, a maioria dessas palavras remete ao contexto político. Porém, nas palavras associadas aos textos com discurso de ódio (subfigura b), observam-se termos que denotam agressões verbais, tais como "mentiroso", "genocida", "burro", "vagabundo", entre outros.

Na Figura 4.27, observa-se que as sentenças, tanto os que contêm discurso de ódio quanto os que não contêm, apresentam em sua maioria tamanhos no intervalo entre 0 e 20. Destaca-se, entretanto, que as sentenças com discurso de ódio concentram-se predominantemente no intervalo entre 10 e 20 comparado ao mesmo intervalo dos textos sem discurso de ódio.

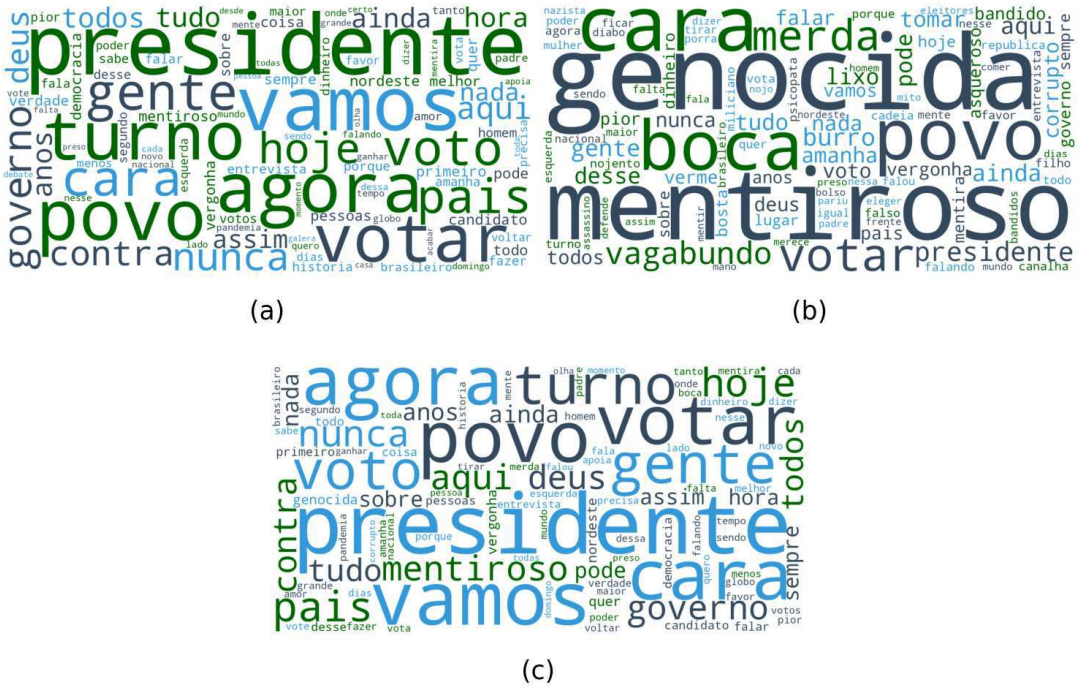

Figura 4.26: As cem palavras com maior ocorrência no corpus com idioma português.

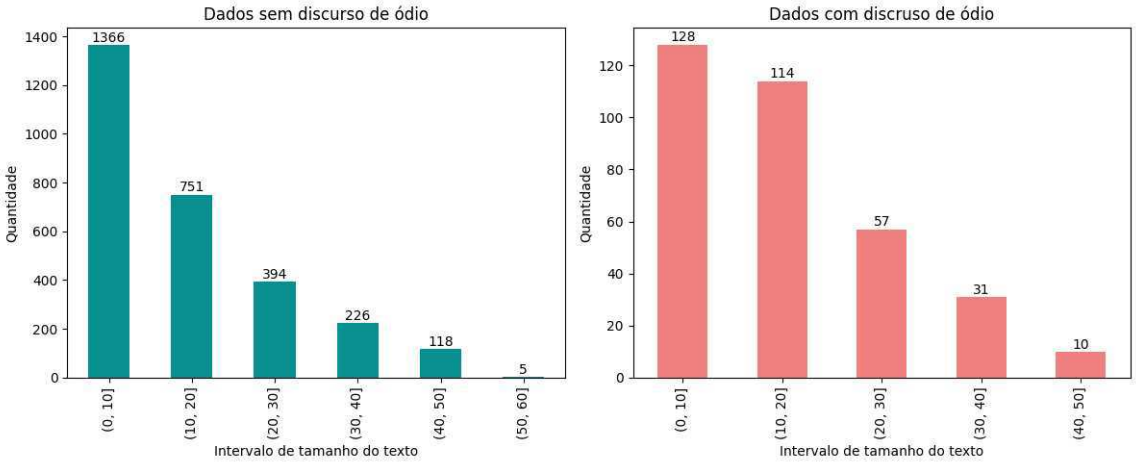

Figura 4.27: Intervalo do tamanho das sentenças em português.

A Figura 4.28 apresenta as médias do tamanho das sentenças com discurso de ódio, bem como as sentenças sem discurso de ódio. Nota-se que a média das sentenças com discurso de ódio é superior à das sentenças sem discurso de ódio. Isso sugere uma tendência na escrita dos textos, indicando que no cenário político brasileiro os usuários têm uma propensão a

produzir textos mais extensos ao manifestarem discurso de ódio.

A Figura 4.29 contempla o *boxplot* e o desvio padrão para as sentenças em português. Assim como na maioria dos idiomas apresentados anteriormente, o corpus em português apresentou um desvio padrão elevado em relação à média, indicando uma variação considerável no tamanho das sentenças. Além disso, o tamanho das sentenças com discurso de ódio mostrou-se maior em relação às sentenças sem discurso de ódio, pois o terceiro quartil está acima de 20, enquanto o das sentenças sem discurso de ódio está mais próximo do valor 20. Esse mesmo comportamento foi observado na maioria dos idiomas analisados anteriormente.

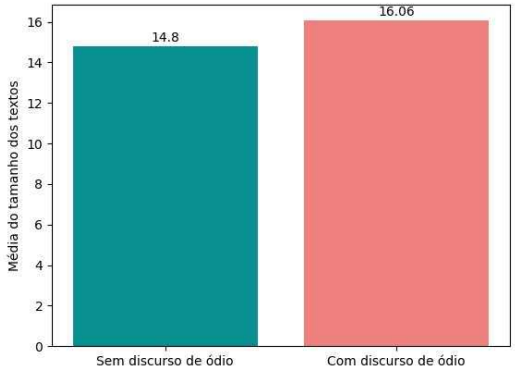

Figura 4.28: Média do tamanho das sentenças em português.

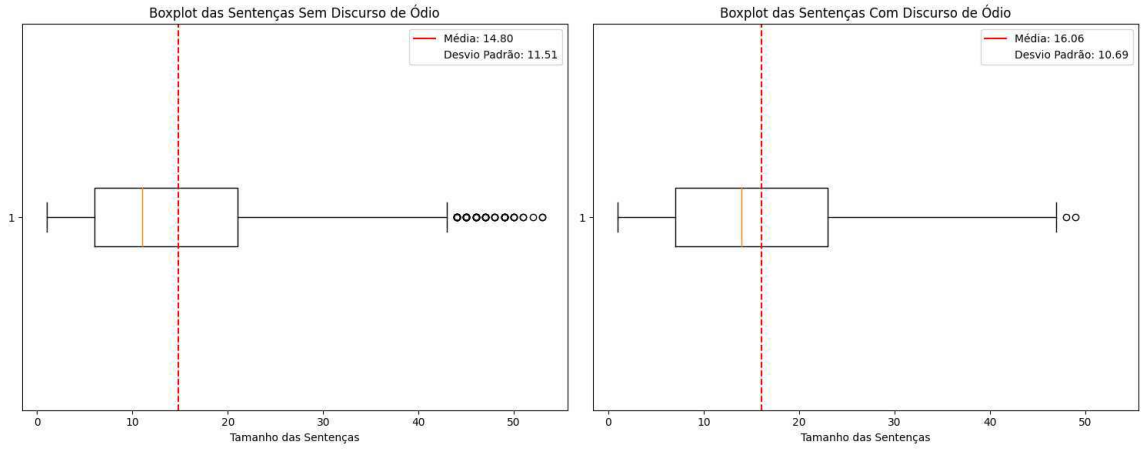

Figura 4.29: *Boxplot* das sentenças em português.

4.2.8 Considerações Finais Sobre os Dados

Notavelmente, na maioria dos idiomas analisados, as sentenças com discurso de ódio apresentaram um tamanho maior em relação às sentenças sem discurso de ódio. A única exceção foi o idioma espanhol, que apresentou um comportamento oposto. Além disso, na maioria dos dados, o desvio padrão apresentou um valor elevado, indicando uma variação considerável no tamanho das sentenças, exceto no corpus filipino, no qual o valor foi menor em relação aos outros idiomas. Por fim, outro ponto a destacar é em relação às palavras presentes nas sentenças em todos os dados, pois as sentenças com discurso de ódio apresentaram palavras de baixo calão e pejorativas em praticamente todos os idiomas analisados enquanto as palavras sem discurso de ódio apresentaram um conteúdo mais neutro.

4.3 Modelos de IA

A escolha de uma ferramenta de IA, principalmente os modelos, desempenha um papel crucial na eficácia em detectar discurso de ódio. Optar por modelos estado-da-arte, que representam o estado atual mais avançado da pesquisa em IA, pode resultar em melhorias significativas na detecção de discurso de ódio. Atualmente, alguns dos modelos mais proeminentes incluem aqueles baseados na arquitetura *Transformer*. Neste trabalho, foram selecionados modelos baseados em codificadores e decodificadores que fazem uso desta arquitetura.

Os modelos baseados em codificadores são o BERT pré-treinado no idioma inglês [25], o BERT pré-treinado no idioma italiano [86], bem como o modelo multilíngue XLM-Roberta [17]. Os modelos decodificadores são o GPT-3 [10] e o GPT-3.5 Turbo [70]. Esses modelos foram escolhidos devido a sua eficácia em tarefas relacionadas a PLN em trabalhos relacionados, incluindo abordagens para detectar discurso de ódio. Dessa forma, a adoção desses modelos busca extrair todo o potencial da IA estado-da-arte para detectar discurso de ódio. Portanto, essa abordagem permite obter resultados mais precisos, contribuindo para o avanço na elaboração de ferramentas mais eficazes para detectar discurso de ódio e promover ambientes online mais seguros e inclusivos.

4.4 Técnicas Provenientes do CLL

Na metodologia deste trabalho, foi empregado o CLL. Este configura-se como um recurso de AM, onde o conhecimento obtido de um ou mais idiomas é empregado para aprimorar a eficiência de um modelo direcionado para classificação em um determinado idioma de destino específico, frequentemente caracterizado por dados ou recursos linguísticos relativamente limitados. Dessa forma, nesta pesquisa, o propósito central da adoção da técnica CLL é desenvolver modelos com capacidade de generalização de um determinado idioma específico, mesmo em situações que envolvem disparidades léxicas, proporcionando aprimoramento nos resultados do modelo quando confrontado com corpora anotados e restritos.

Portanto, por meio da incorporação da abordagem CLL, é possível aplicar técnicas específicas. Nesta dissertação, foram aplicadas as seguintes técnicas [73]: *Zero-shot Transfer* (ZST), *Joint Learning* (JL), *Cascade Learning* (CL), *Joint Learning* combinado com *Cascade Learning* (JL/CL) e a combinação JL/CL+ que utiliza o mesmo conceito da JL/CL, porém com uma validação cruzada na parte do CL. Vale destacar que a validação cruzada foi realizada apenas na técnica JL/CL+, pois foi esta dissertação manteve a metodologia original proposta pelos autores das técnicas [73; 31]. Destarte, a aplicação dessas técnicas visa explorar e potencializar as capacidades do modelo de AM, oferecendo assim um melhor enfoque e especializado para detectar discurso de ódio em dados com diferentes contextos linguísticos e culturais.

Na técnica ZST, corpora de determinados idiomas são empregados durante as etapas de treino e validação do modelo. Já os dados referentes ao idioma de destino são estritamente utilizados durante a fase de teste do modelo. Vale destacar que tanto os idiomas dos corpora de treino quanto o idioma de destino são distintos entre si. A Figura 4.30 contempla uma representação visual dos passos envolvidos nessa abordagem.

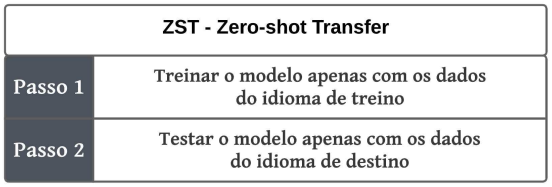

Figura 4.30: Passos para a técnica ZST.

A técnica JL utiliza corpora de determinados idiomas no treino, incluindo também um subconjunto de dados do idioma de destino como parte do corpora de treinamento. Os dados remanescentes do idioma de destino são reservados para o teste do modelo. A Figura 4.31 apresenta uma visão detalhada das etapas envolvidas nesta abordagem.

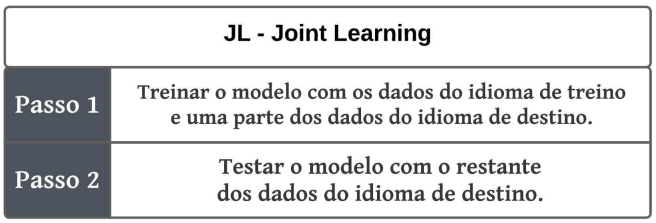

Figura 4.31: Passos para a técnica JL.

Ao direcionar uma porção específica dos dados referentes ao idioma de destino para o treino, essa técnica busca otimizar o aprendizado do modelo, permitindo, assim, que o modelo desenvolva uma representação mais aprofundada e adaptada aos padrões linguísticos do conjunto alvo. A separação dos dados restantes para a fase de teste é essencial para medir a eficiência relacionada à generalização do modelo ao ser testado com dados não utilizados durante o treino.

A técnica CL envolve dois ajustes finos distintos. Na primeira parte dessa técnica, o modelo deve ser submetido a um ajuste fino empregando exclusivamente o corpora referente aos idiomas de treino. Subsequentemente, na segunda parte, realiza-se um outro ajuste fino, no qual um subconjunto dos dados referentes ao idioma de destino é empregado. Os dados remanescentes do corpus referente ao idioma de destino é reservado para medir a eficiência do modelo.

Essa técnica visa aprimorar a capacidade do modelo de generalização para o idioma de destino ao conduzir ajustes finos sequenciais, inicialmente focalizando nos dados referentes aos idiomas treino e, posteriormente, adaptando-se a uma porção específica dos dados referentes ao idioma de destino. A separação dos dados reservados para a fase de teste no segundo ajuste fino é crucial para poder avaliar a eficácia e a adaptabilidade do modelo relacionada aos dados específicos do idioma de destino que não foram empregados na fase de treino. A Figura 4.32 oferece uma representação visual das etapas inerentes a essa estratégia.

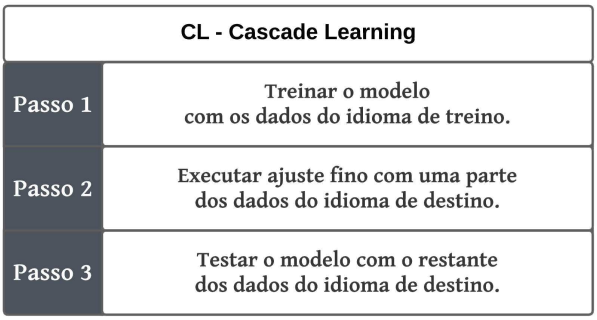

Figura 4.32: Passos para a técnica CL.

Além das estratégias previamente mencionadas, destaca-se a técnica JL/CL. Essa abordagem incorpora uma combinação entre as técnicas JL e CL, significando que uma porção dos dados referentes ao idioma de destino é empregada durante o ajuste fino inicial do modelo, enquanto o restante é reservado para a segunda etapa de ajuste fino aplicado na técnica CL. Posteriormente, realiza-se o ajuste fino do modelo usando uma porção do corpus referente ao idioma de destino, seguido pelo teste do modelo, empregando os dados remanescentes do idioma de destino. A Figura 4.33 proporciona uma visão resumida das etapas relacionadas a essa técnica.

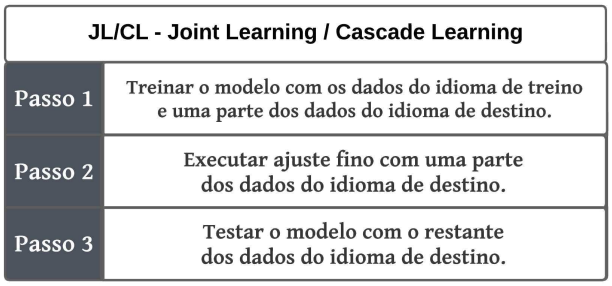

Figura 4.33: Passos para a técnica JL/CL.

Por último, temos a técnica JL/CL+. Nessa abordagem, o símbolo "+" denota a realização de múltiplos ajustes finos durante a etapa CL. Dessa forma, essa estratégia segue o mesmo procedimento empregado na estratégia JL/CL, com a diferença crucial ocorrendo na etapa CL, no qual é realizada uma validação cruzada com os dados provenientes do idioma de destino. A Figura 4.34 apresenta as etapas envolvidas na implementação dessa estratégia.

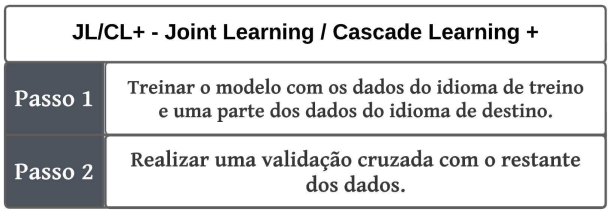

Figura 4.34: Passos para a técnica JL/CL+.

4.5 Métricas e Avaliação dos Resultados

A etapa final da metodologia adotada nesta dissertação consiste em avaliar os modelos implementados por meio das técnicas de CLL. Inicialmente, foi conduzido um experimento base com o intuito de verificar posteriormente se a incorporação de CLL resulta em uma melhoria na eficácia do modelo a ser avaliado, mesmo diante de corpora apresentando disparidades léxicas. Nesse experimento base, não foi utilizado CLL no treino do modelo, empregando apenas um único idioma para o treinamento e para o teste do modelo.

Posteriormente, foram conduzidos experimentos adicionais aplicando CLL combinado com múltiplos idiomas e com as técnicas apresentadas na Seção 4.4. Em seguida, os resultados provenientes das técnicas de CLL foram confrontados com os resultados provenientes do experimento base, visando verificar se a inclusão de CLL promoveu melhorias no desempenho. Adicionalmente, as estratégias foram também objeto de uma comparação entre si, com o intuito de identificar qual delas mostra-se mais eficaz em detectar discurso de ódio.

Portanto, com o propósito de avaliar a eficiência do modelo em contraste ao experimento base, bem como verificar a eficácia na aplicação das técnicas de CLL, foram empregadas as métricas detalhadas na Seção 2.5, incluindo precisão, revocação e medida-F1. No entanto, vale ressaltar que devido ao desbalanceamento dos dados, optou-se pela utilização da medida-F1 *weighted* (ponderada). Essas métricas são essenciais para uma avaliação mais eficiente e quantitativa da eficiência dos modelos, fornecendo percepções sobre a capacidade dos modelos em detectar discurso de ódio em contextos multilíngues e disparidades léxicas.

4.6 Considerações Finais

Neste capítulo, foi descrita a metodologia adotada neste trabalho, abrangendo o procedimento de obtenção de corpora e detalhando os dados empregados nos experimentos. Adicionalmente, foram apresentados os modelos empregados nos experimentos desta dissertação. Ademais, foram discutidas as técnicas derivadas do CLL, utilizadas no aprimoramento dos modelos para detectar discurso de ódio. Por fim, foram apresentadas as métricas e o método de avaliação dos resultados. O próximo capítulo aborda os experimentos conduzidos neste trabalho.

Capítulo 5

Experimentos

Neste capítulo, são apresentados os experimentos executados, os resultados alcançados e sua discussão subsequente. A Seção 5.1 mostra os experimentos com modelos do tipo codificador BERT. A Seção 5.2 contempla os experimentos com modelos do tipo codificador XLM-Roberta. A Seção 5.3 contempla os experimentos com modelos do tipo decodificador. A Seção 5.4 apresenta a análise dos resultados obtidos, bem como exibe um comparativo com outros trabalhos. A Seção 5.5 exibe os resultados observados em um experimento prático utilizando os modelos treinados em português. A seção 5.6 contempla um sumário com todos os experimentos realizados. Por fim, a Seção 5.7 contempla as considerações finais deste capítulo.

5.1 Experimentos com Modelos Monolíngue do Tipo Codificador

A fase inicial dos experimentos envolve os modelos baseados em codificadores. Para essa etapa, foram empregados os modelos BERT pré-treinados em italiano [86], inglês [25] e português [91]. Para todos os modelos, os seguintes parâmetros foram utilizados: taxa de aprendizado igual 1 x e^{-6}, otimizador AdamW com o valor epsilon igual a 1 x e^{-8} e número de épocas igual a três. O otimizador AdamW foi utilizado, pois ele é muito útil em situações onde a generalização do modelo representa prioridade, ajudando a melhorar a generalização do modelo e consequentemente evitando *overfitting*.

Vale destacar que o valor selecionado para as épocas foi determinado após observações experimentais. Foi notado que, acima desse valor, enquanto o erro durante o treino do modelo começava a diminuir, o erro durante a validação começava a aumentar, indicando um início de overfitting. Ademais, essa quantidade de épocas foi adotada tomando como base pesquisas relacionadas, que também empregaram o mesmo valor de épocas em seus experimentos [25; 16; 31]. Assim, esse valor de épocas foi escolhido tomando como base experimentos preliminares para evitar *overfitting*, bem como reforçando-se em trabalhos anteriores.

Vale destacar que o tamanho dos textos fornecidos ao modelo como entrada foram limitados em 128 palavras, visto que nos dados não foram identificados textos maiores que esse tamanho. Isso permite reduzir custo de processamento para o modelo, tanto no treinamento como no teste. Adicionalmente, o Colab[1] foi utilizado para executar os experimentos com o BERT. O Colab é um ambiente gratuito fornecido pela Google. O ambiente possui as seguintes configurações: uma CPU Intel Xeon de 2.30GHz, 12 Gigabytes de RAM e uma placa de vídeo Tesla T4 com 15 Gigabytes de memória GDDR6. Nas subseções seguintes, os experimentos referentes aos modelos codificadores serão detalhados.

5.1.1 Experimento base sem CLL

Uma das investigações realizadas nesta dissertação foi avaliar se existe melhoria na eficiência do modelo base ao utilizar CLL juntamente com a inclusão de múltiplos idiomas de treino diferentes do idioma de destino. Para responder essa questão, foi conduzido um experimento base sem CLL. Posteriormente os resultados desse experimento foram comparados com os obtidos pelos experimentos onde o CLL foi aplicado.

No experimento base, foram empregados modelos BERT nos idiomas italiano, português e inglês. Ademais, nenhuma técnica CLL foi empregada ao experimento. Os dados para cada experimento foram separados da seguinte maneira: 70% dos dados foram reservados para a fase de treino, 10% para a fase de validação do modelo e 20% para a fase de teste do modelo. Essa proporção foi aplicada em cada experimento.

A Tabela 5.1 contempla os resultados provenientes dos experimentos. Observa-se que os valores alcançados nos resultados dos modelos foram bastante semelhantes. O modelo BERT

[1] https://colab.research.google.com/

no idioma português foi superior aos outros, obtendo uma medida-F1 de 87,62%, seguido pelo modelo BERT no idioma inglês, que atingiu 85,00% na medida-F1, e, por último, o modelo BERT no idioma italiano, com 82,00% na mesma métrica.

Além do experimento base, também foi conduzido um experimento com validação cruzada *k-fold* com k = 5. A validação cruzada foi empregada para explorar se era viável obter um desempenho aprimorado além do experimento base inicial, mesmo sem a aplicação do CLL. A Tabela 5.2 contempla os resultados provenientes dessa abordagem. Nota-se que houve um pequeno acréscimo no valor da medida-F1 no modelo inglês, aumentando 0,30% comparado ao experimento base. No modelo italiano, também houve um aumento pequeno, com um aumento de 0,93% na medida-F1 comparado ao experimento base sem validação cruzada. Em contrapartida, no modelo em português, houve um aumento considerável comparado aos modelos inglês e italiano.

Tabela 5.1: Resultados obtidos do experimento base.

Modelo	Corpus de treino	Corpus de teste	Precisão	Revocação	Medida-F1
BERT inglês	inglês	inglês	80,00%	90,00%	85,00%
BERT italiano	italiano	italiano	81,00%	84,00%	82,00%
BERT português	português	português	87,49%	89,06%	87,62%

Tabela 5.2: Resultados obtidos do experimento base com validação cruzada. Cada valor representa a média dos resultados obtidos.

Modelo	Corpus de treino	Corpus de teste	Precisão	Revocação	Medida-F1
BERT inglês	inglês	inglês	81,11%	90,07%	85,30%
BERT italiano	italiano	italiano	83,06%	82,95%	82,93%
BERT português	português	português	90,84%	91,81%	90,95%

5.1.2 Experimentos com estratégias CLL e idiomas com distância léxica maior

Neste experimento, para o treino do modelo foram empregados múltiplos idiomas que possuem uma distância léxica maior comparado ao idioma de destino. Nesse contexto, o Corpora do Idioma Fonte (CIF) representa a combinação dos idiomas e seus respectivos dados que serão utilizados para treino e validação do modelo, enquanto o Corpus do Idioma de Destino (CID) representa o idioma e também o corpus destinado para teste do modelo. Ademais, essas combinações foram empregadas juntamente com as técnicas que utilizam CLL, conforme abordado na Seção 4.4. Nas subseções seguintes, tanto os resultados quanto os detalhes dos experimentos são discutidos.

A) Resultados provenientes da técnica ZST

Nessa estratégia, foram utilizados 90% do CIF para o treino do modelo, enquanto os 10% restantes do CIF foram reservados para validação. Todo o CID foi empregado para teste do modelo. A Tabela 5.3 contempla os resultados provenientes do modelo BERT no idioma inglês. Observa-se que os valores obtidos nos resultados se aproximaram dos valores alcançados no experimento base.

No entanto, mesmo usando a técnica ZST, o resultado ainda ficou abaixo do modelo base. No experimento base, sem o CLL, o valor mais eficaz para a medida-F1 para esse modelo foi de 85,30%, enquanto nesta estratégia o melhor valor obtido foi de 84,00% na mesma métrica.

Tabela 5.3: Resultados obtidos com o BERT no idioma inglês na técnica ZST.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	**inglês**	**86,00%**	**82,00%**	**84,00%**
italiano, filipino e alemão	inglês	87,00%	81,00%	84,00%
italiano, filipino, alemão e turco	inglês	87,00%	78,00%	82,00%

A Tabela 5.4 contempla os resultados provenientes do modelo italiano na técnica ZST. Nota-se que, do mesmo modo que o modelo inglês, o modelo italiano não demonstrou resul-

tados superiores ao modelo base. O valor mais eficaz para a medida-F1 foi de 37,00%, um valor significativamente inferior ao alcançado no experimento base, onde foi possível obter uma medida-F1 máxima de 82,93%.

No entanto, é importante notar que houve um aumento na eficiência do modelo italiano ao incorporar um maior número de idiomas no treino do modelo. A medida-F1 subiu de 35,00% (quando apenas inglês e filipino foram usados como CIF) para 37,00% quando os idiomas inglês, filipino, alemão e turco foram empregados como CIF.

Tabela 5.4: Resultados obtidos com o BERT no idioma italiano na estratégia ZST.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	59,00%	51,00%	34,00%
inglês, filipino e alemão	italiano	68,00%	52,00%	36,00%
inglês, filipino, alemão e turco	**italiano**	**69,00%**	**52,00%**	**37,00%**

B) Resultados provenientes da técnica JL

A técnica JL consiste em utilizar parte do CID, adicionando-o ao CIF, como fonte de treinamento. Nessa abordagem, o CIF foi complementado com 30% do CID para o treino do modelo. Os 70% dos dados remanescentes do CID foram reservados para o teste do modelo. A Tabela 5.5 exibe os valores obtidos nos resultados do BERT no idioma inglês com essa estratégia. É possível perceber uma melhoria, alcançando 87,00% na medida-F1, um valor superior ao obtido no modelo base (85,30%).

Tabela 5.5: Resultados obtidos com o BERT no idioma inglês na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	84,00%	88,00%	86,00%
italiano, filipino e alemão	inglês	86,00%	87,00%	86,00%
italiano, filipino, alemão e turco	**inglês**	**87,00%**	**87,00%**	**87,00%**

Ademais, houve uma melhoria comparado à estratégia ZST, com um aumento de 3%

na medida-F1. Adicionalmente, nota-se que os valores obtidos nos resultados melhoraram conforme mais idiomas foram adicionados como CIF. O valor da medida-F1 aumentou de 86,00% para 87,00% com a inclusão de mais idiomas. Também foi observado um melhor equilíbrio entre a precisão e a revocação com o aumento dos idiomas.

A Tabela 5.6 exibe o mesmo experimento, utilizando o modelo BERT italiano. Observa-se que os valores obtidos nos resultados ultrapassaram os valores obtidos na estratégia ZST, representando uma melhoria de aproximadamente 45% na medida-F1. Ademais, os valores ultrapassaram o resultado do experimento base, indicando que o CLL, nesse caso, contribuiu para melhorar a eficiência do modelo. De maneira geral, a técnica JL mostrou-se mais promissora em comparação à estratégia ZST, tanto no modelo inglês quanto no modelo italiano.

Tabela 5.6: Resultados obtidos com o BERT no idioma italiano na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	**italiano**	**79,00%**	**87,00%**	**83,00%**
inglês, filipino e alemão	italiano	82,00%	82,00%	82,00%
inglês, filipino, alemão e turco	italiano	82,12%	82,10%	82,10%

C) Resultados provenientes da técnica CL

Na técnica CL, foram efetuados dois ajustes finos no modelo. Na etapa inicial de ajuste fino, foram utilizados 90% do CIF para treino do modelo e 10% para validação. Além disso, foram utilizadas 3 épocas para o treinamento. Após o modelo ser treinado, o mesmo foi salvo e submetido a um novo ajuste fino. Nesse segundo ajuste fino, foram utilizados 70% dos dados do CID para treino e 10% para validação, deixando os 20% restantes para teste do modelo.

A Tabela 5.7 mostra os resultados alcançados para o modelo BERT inglês. Observa-se uma melhoria nos resultados com essa estratégia, alcançando 89,00% na medida-F1, um aumento de 2% comparado com a estratégia anterior e de 5% em relação à estratégia ZST. Ademais, os resultados derivados da técnica CL superaram os do experimento base, com um aumento de 3,7%, sugerindo que o CLL, nesse caso, contribuiu para aprimorar a eficiência do modelo, assim como na estratégia anterior (JL).

Tabela 5.7: Resultados obtidos com o BERT no idioma inglês na técnica CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	88,00%	90,00%	89,00%
italiano, filipino e alemão	inglês	87,00%	90,00%	87,00%
italiano, filipino, alemão e turco	**inglês**	**88,00%**	**90,00%**	**89,00%**

A Tabela 5.8 mostra o mesmo experimento aplicado ao BERT italiano. Percebe-se uma melhoria em relação à estratégia JL, alcançando 86,00% como resultado, em contraste com os 83,00% da técnica JL. Ademais, é perceptível que essa estratégia conseguiu superar o valor obtido no experimento base (82,93%). Nos modelos inglês e italiano, essa estratégia mostrou-se mais promissora comparado às estratégias anteriores.

Tabela 5.8: Resultados obtidos com o BERT no idioma italiano na técnica CL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	**italiano**	**86,00%**	**86,00%**	**86,00%**
inglês, filipino e alemão	italiano	82,00%	86,00%	84,00%
inglês, filipino, alemão e turco	italiano	86,00%	85,00%	86,00%

D) Resultados provenientes da técnica JL/CL

Nessa estratégia, foram empregadas as duas técnicas anteriores (JL e CL). Na parte da técnica JL, foram utilizados 90% do CIF e 10% do CID. Ainda na estratégia JL, na parte de validação, foram utilizados 20% do CID combinado com 10% do CIF. Os dados remanescentes do CID foram empregados na parte da técnica CL, sendo 70% desses dados remanescentes utilizados para treino, 10% para validação e 20% para teste.

A Tabela 5.9 mostra os resultados alcançados com essa estratégia para o modelo BERT inglês. Os resultados apresentaram um pior desempenho comparado à estratégia anterior (CL). No entanto, é possível notar que, nessa estratégia, os resultados eram aprimorados conforme eram acrescentados mais idiomas, pois o valor saltou de 84,31% na medida-F1 com

dois idiomas fonte para 86,00% quando foi acrescentado mais um idioma e depois aumentou para 87,10% quando foi acrescentado outro idioma. Ademais, os resultados apresentaram um valor superior ao obtido no experimento base.

Tabela 5.9: Resultados obtidos com o BERT no idioma inglês na estratégia JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	85,00%	89,00%	84,31%
italiano, filipino e alemão	inglês	88,00%	90,00%	86,00%
italiano, filipino, alemão e turco	**inglês**	**87,00%**	**89,00%**	**87,10%**

A Tabela 5.10 exibe os resultados referentes ao modelo BERT italiano. Similarmente ao modelo BERT inglês, o modelo italiano demonstrou um decréscimo de desempenho comparado à estratégia anterior (CL). Porém, os resultados alcançados foram melhores que o experimento base, mostrando assim a eficácia do CLL no aprimoramento do modelo.

Tabela 5.10: Resultados obtidos com o BERT no idioma italiano na técnica JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	85,00%	84,00%	84,26%
inglês, filipino e alemão	italiano	85,00%	84,00%	84,05%
inglês, filipino, alemão e turco	**italiano**	**86,00%**	**85,00%**	**85,25%**

E) Resultados provenientes da técnica JL/CL+

A técnica JL/CL+ possui a mesma configuração da técnica anterior (JL/CL). A única diferença está na parte da técnica CL, na qual é empregada uma validação cruzada *k-fold* com k = 5 para separar os dados em treino, validação e teste. A Tabela 5.11 mostra os resultados alcançados para o modelo BERT inglês. É possível notar que por meio dessa técnica foi possível obter os melhores resultados comparado às estratégias anteriores.

A Tabela 5.12 exibe os resultados referentes ao modelo italiano. Assim como o modelo inglês, os resultados se mostraram superiores comparado às outras estratégias. Ademais,

na estratégia JL/CL+, ambos os modelos obtiveram resultados superiores comparados ao experimento base.

Tabela 5.11: Resultados obtidos com o BERT no idioma inglês na estratégia JL/CL+. Cada valor corresponde a média obtida dos resultados.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	94,6%	94,00%	94,30%
italiano, filipino e alemão	**inglês**	**94,60%**	**95,00%**	**94,80%**
italiano, filipino, alemão e turco	inglês	94,80%	94,00%	94,40%

Tabela 5.12: Resultados obtidos com o BERT no idioma italiano na estratégia JL/CL+.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	92,80%	92,00%	92,40%
inglês, filipino e alemão	italiano	92,70%	92,50%	92,60%
inglês, filipino, alemão e turco	**italiano**	**92,00%**	**94,00%**	**93,00%**

5.1.3 Experimentos com estratégias CLL e idiomas com distância léxica menor

Nos experimentos com a distância léxica menor foram empregados os idiomas cuja distância léxica é a menor possível entre si. Para esse propósito, foi aplicado o modelo BERT português [91]. Ademais, foram empregados os idiomas italiano, espanhol e português, bem como foram aplicadas as mesmas técnicas CLL apresentadas na Seção 4.4. Nas subseções seguintes, serão descritos cada experimento e apresentados os resultados provenientes de cada um deles.

A) Resultados provenientes da técnica ZST

Na estratégia ZST foi utilizado 90% do CIF no treino do modelo e 10% foi empregado para validação. O CID foi utilizado para o teste do modelo. A Tabela 5.13 mostra os resultados

alcançados. É possível perceber que o resultado foi inferior ao resultado do experimento base. Porém, os resultados ficaram próximos.

Tabela 5.13: Resultados obtidos com o BERT português na estratégia ZST.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	85,77%	88,34%	86,67%

B) Resultados provenientes da técnica JL

Nessa estratégia, 30% do CID e todo o CIF foram utilizados no treinamento. Desse total, 10% foi utilizado para validação do modelo. Os dados remanescentes do CID foram designados para testar o modelo. A Tabela 5.14 mostra os resultados alcançados para essa estratégia. Percebe-se que houve uma melhoria comparado à estratégia anterior, pois foi possível obter um resultado melhor. No entanto, mesmo obtendo um resultado melhor, ainda não foi possível superar os resultados alcançados no experimento base.

Tabela 5.14: Resultados obtidos com o BERT português na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	86,61%	89,78%	87,37%

C) Resultados provenientes da técnica CL

Nessa estratégia, foram aplicados dois ajustes finos no modelo. Foram utilizados 90% do CIF para o primeiro treinamento e 10% do CIF foi utilizado para validação. No segundo ajuste fino, 70% do CID foi utilizado para treino e 10% foi designado para validação. Os dados remanescentes do CID serviram para o teste do modelo.

A Tabela 5.15 contempla os resultados alcançados. Pode-se notar uma melhoria comparado às estratégias anteriores, pois os resultados alcançados foram superiores. Contudo, é possível perceber também que os valores obtidos nos resultados foram melhores comparados aos do experimento base que não foi submetido a validação cruzada. No entanto, o resultado

ainda ficou abaixo do experimento base com validação cruzada, porém muito próximo, pois foi possível obter 89,64% na medida-F1 em contraste com 90,95% do experimento base que foi submetido a validação cruzada.

Tabela 5.15: Resultados obtidos com o BERT português na técnica CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	89,64%	90,62%	89,64%

D) Resultados provenientes da técnica JL/CL

Nessa estratégia, foram combinadas as técnicas JL e CL. Na técnica JL, 90% do CIF e 10% do CID foram empregados para o treino do modelo. Adicionalmente, 10% do CIF acrescidos de 20% do CID foram empregados na parte de validação. Os dados remanescentes do CID foram empregados na etapa CL. Nesta etapa, 70% desses dados foram utilizados para treino, 10% para validação e 20% para teste.

A Tabela 5.16 exibe os resultados alcançados usando essa estratégia. Observa-se que os valores obtidos nos resultados foram os melhores comparado às estratégias anteriores. Ademais, foi possível superar o melhor resultado do experimento base, sendo possível obter 92,86% na medida-F1 em contraste com o valor de 90,95% do experimento base. Dessa forma, nessa estratégia, o CLL ajudou a aprimorar o modelo.

Tabela 5.16: Resultados obtidos com o BERT português na estratégia JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	93,54%	92,41%	92,86%

E) Resultados provenientes da técnica JL/CL+

Essa estratégia utiliza as mesmas configurações da técnica anterior (JL/CL). A única diferença está na parte da técnica CL em que é empregada uma validação cruzada *k-fold* com k = 5 para separar os dados em treino, validação e teste. A Tabela 5.17 exibe os resultados

alcançados. Podemos perceber que dentre as estratégias, essa foi a que resultou no melhor resultado, pois superou os resultados das estratégias anteriores, bem como foi superior ao experimento base.

Tabela 5.17: Resultados obtidos com o BERT português na estratégia JL/CL+.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	96,80%	97,00%	96,92%

5.2 Experimentos com Modelo Multilíngue do Tipo Codificador

Esta seção contempla os experimentos realizados com o modelo multilíngue XLM-Roberta [17]. As configurações utilizadas nesse modelo foram as mesmas empregadas no modelo monolíngue. O ambiente e as configurações de *hardware* também foram os mesmos utilizados no modelo monolíngue. Nas subseções seguintes, são apresentados e discutidos os resultados obtidos por esse modelo.

5.2.1 Experimento base sem CLL

Assim como nos modelos codificadores monolíngues, foi realizado um experimento base sem CLL no modelo multilíngue. Os dados foram divididos em 70% para treino, 10% para validação e 20% para teste. A Tabela 5.18 contempla os resultados obtidos no experimento base para os idiomas de destino inglês, italiano e português. Nota-se que o melhor resultado foi alcançado quando foi utilizado o idioma inglês, sendo possível obter 85,02% na medida-F1. O segundo melhor resultado foi alcançado quando foi utilizado o idioma português, obtendo um valor de 81,00% na medida-F1. Por fim, foi possível obter 79,00% na medida-F1 quando foi utilizado o idioma italiano.

Adicionalmente, outro experimento foi realizado, aplicando validação cruzada com $k = 5$. A Tabela 5.19 exibe os resultados obtidos para os idiomas de destino inglês, italiano e português. Observa-se que o melhor resultado foi obtido com o idioma inglês, alcançando

85,30% na medida-F1. O segundo melhor resultado foi alcançado com o idioma português, com um valor de 84,36% na medida-F1. O pior resultado foi obtido com o idioma italiano, alcançando 74,64% na medida-F1.

Tabela 5.18: Resultados obtidos do experimento base para o modelo XLM-Roberta.

Modelo	Corpus de treino	Corpus de teste	Precisão	Revocação	Medida-F1
XLM-Roberta	inglês	inglês	80,70%	89,83%	85,02%
	italiano	italiano	79,00%	79,00%	79,00%
	português	português	75,74%	87,03%	81,00%

Tabela 5.19: Resultados obtidos do experimento base com validação cruzada para o modelo XLM-Roberta. Cada resultado representa a média dos resultados obtidos.

Modelo	Corpus de treino	Corpus de teste	Precisão	Revocação	Medida-F1
XLM-Roberta	inglês	inglês	81,05%	90,03%	85,30%
	italiano	italiano	75,51%	74,81%	74,64%
	português	português	79,88%	89,38%	84,36%

5.2.2 Experimentos com as estratégias CLL e idiomas com distância léxica maior

Esta seção apresenta os resultados obtidos para os idiomas cuja distância léxica é mais ampla. As estratégias de CLL utilizadas foram as mesmas empregadas no experimento monolíngue, assim como os idiomas utilizados para treino e teste. Nas subseções seguintes, são apresentados os resultados obtidos com o modelo codificador multilíngue.

A) Resultados provenientes da técnica ZST

Nessa estratégia, foram utilizados 90% do CIF para o treino do modelo, enquanto os 10% restantes do CIF foram reservados para validação. Para realizar o teste do modelo, foi utilizado todo o CID. A Tabela 5.20 apresenta os resultados obtidos pelo modelo XLM-Roberta

para o idioma inglês. É possível notar que, na técnica ZST, o modelo não conseguiu superar o modelo base. O melhor valor obtido foi de 76,11% com a combinação dos idiomas italiano, filipino e alemão no treino do modelo.

Tabela 5.20: Resultados obtidos com o XLM-Roberta no idioma inglês na técnica ZST.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	87,01%	70,93%	75,94%
italiano, filipino e alemão	**inglês**	**86,91%**	**71,17%**	**76,11%**
italiano, filipino, alemão e turco	inglês	86,97%	70,33%	75,48%

Na Tabela 5.21, temos os resultados do mesmo modelo aplicado ao idioma italiano. É possível perceber que o modelo obteve uma performance inferior se comparado ao modelo anterior. Além disso, o modelo também não conseguiu superar o modelo base. Portanto, tanto no modelo inglês quanto no modelo italiano a técnica ZST não se mostrou eficiente em aprimorá-los.

Tabela 5.21: Resultados obtidos com o XLM-Roberta no idioma italiano na estratégia ZST.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	**italiano**	**76,47%**	**73,58%**	**73,02%**
inglês, filipino e alemão	italiano	73,67%	63,28%	59,38%
inglês, filipino, alemão e turco	italiano	75,51%	70,20%	68,87%

B) Resultados provenientes da técnica JL

Nesse experimento, foi utilizado todo o CIF para treino, acrescido de 30% do CID. Os 70% dos dados restantes do CID foi utilizado para realizar o teste do modelo. A Tabela 5.22 contempla os resultados obtidos utilizando o idioma inglês como alvo. Nota-se que houve uma melhoria nos resultados se comparado à técnica anterior. Além disso, o resultado foi superior ao modelo base, obtendo cerca de 1% a mais na medida-F1. Portanto, nessa técnica o modelo conseguiu ser aprimorado empregando CLL com múltiplos idiomas no treino.

Tabela 5.22: Resultados obtidos com o XLM-Roberta no idioma inglês na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	85,52%	83,84%	84,60%
italiano, filipino e alemão	**inglês**	**85,88%**	**86,13%**	**86,00%**
italiano, filipino, alemão e turco	inglês	85,28%	83,19%	84,12%

A Tabela 5.23 exibe os resultados para o idioma italiano. É possível notar um desempenho superior em relação à técnica ZST, apresentando uma melhoria de cerca de 8% na medida-F1. Ademais, foi possível alcançar um resultado superior comparado ao modelo base. No geral, tanto para o idioma inglês quanto para o idioma italiano, além de uma melhoria em relação à técnica anterior, a técnica JL conseguiu aprimorar o modelo suficientemente em ambos os idiomas para que ele fosse capaz de superar o modelo base.

Tabela 5.23: Resultados obtidos com o XLM-Roberta no idioma italiano na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	**italiano**	**81,24%**	**81,23%**	**81,23%**
inglês, filipino e alemão	italiano	80,19%	80,18%	80,18%
inglês, filipino, alemão e turco	italiano	80,54%	80,56%	80,56%

C) Resultados provenientes da técnica CL

Nessa técnica foi utilizado 90% do CIF para treino e 10% para validação para o primeiro ajuste fino. Em seguida, foi realizado um segundo ajuste fino. No segundo ajuste, foram utilizados 70% dos dados do CID para realizar o treino do modelo e 10% para validação, os 20% dos dados remanescentes foram utilizados para o teste do modelo. A Tabela 5.24 exibe os resultados obtidos para o idioma inglês. Nota-se que houve uma melhoria em relação às técnicas anteriores, com um acréscimo de aproximadamente 5% na medida-F1 comparado à técnica JL e 15% em relação à técnica ZST. Adicionalmente, percebe-se que o resultado foi superior ao obtido no modelo base, demonstrando que, nessa técnica, foi possível aprimorar

o modelo empregando múltiplos idiomas no treino.

Tabela 5.24: Resultados obtidos com o XLM-Roberta no idioma inglês na estratégia CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	90,39%	91,17%	90,69%
italiano, filipino e alemão	**inglês**	**91,20%**	**91,50%**	**91,34%**
italiano, filipino, alemão e turco	inglês	90,50%	91,00%	90,72%

A Tabela 5.25 contempla os resultado obtidos para o idioma italiano. É possível perceber que houve uma melhoria para esse idioma em comparação à técnica anterior (JL), pois foi possível obter cerca de 5% a mais na medida-F1 e 12% em relação à técnica ZST. Portanto, tanto para o idioma inglês quanto o italiano, essa estratégia se mostrou mais efetiva para o modelo XML-Roberta do que as demais apresentadas anteriormente.

Tabela 5.25: Resultados obtidos com o XLM-Roberta no idioma italiano na estratégia CL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	85,27%	85,25%	85,25%
inglês, filipino e alemão	**italiano**	**85,54%**	**85,50%**	**85,49%**
inglês, filipino, alemão e turco	italiano	84,50%	84,50%	84,50%

D) Resultados provenientes da técnica JL/CL

Nesse experimento, na parte JL foi utilizado 90% do CIF e 10% do CID para o treino do modelo. Na parte de validação foram utilizados 20% do CID e 10% do CIF. Os dados remanescentes do CID foram empregados na parte do CL. Sendo 70% desses dados destinados para a parte de treino, 10% para validação e 20% para teste. A Tabela 5.26 contempla os resultados provenientes quando o idioma inglês foi utilizado como destino. Nota-se que, nessa técnica, o modelo obteve um resultado inferior à técnica anterior, obtendo como melhor resultado na medida-F1 o valor de 87,48% contra 91,34% da técnica CL. No entanto, a técnica

JL/CL se mostrou mais eficaz do que as técnicas ZST e JL. Além disso, os resultados foram melhores do que o modelo base.

Tabela 5.26: Resultados obtidos com o XLM-Roberta no idioma inglês na estratégia JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	**inglês**	**87,26%**	**88,86%**	**87,48%**
italiano, filipino e alemão	inglês	85,68%	87,50%	86,23%
italiano, filipino, alemão e turco	inglês	86,33%	88,32%	86,48%

A Tabela 5.27 exibe os resultados obtidos para o idioma italiano como destino. É possível perceber um comportamento similar ao do idioma inglês, pois os resultados foram inferiores à técnica CL, porém superiores às técnicas anteriores. Além disso, foi possível superar o modelo base também. Portanto, nota-se que a aplicação de múltiplos idiomas no treino por meio do CLL foi eficaz em melhorar o desempenho do modelo.

Tabela 5.27: Resultados obtidos com o XLM-Roberta no idioma italiano na estratégia JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	82,22%	81,25%	81,25%
inglês, filipino e alemão	**italiano**	**82,50%**	**82,14%**	**82,17%**
inglês, filipino, alemão e turco	italiano	82,19%	81,79%	81,81%

E) Resultados provenientes da técnica JL/CL+

Nessa técnica foi utilizada a mesma configuração do experimento anterior. A única diferença encontra-se na parte CL, no qual foi realizada uma validação cruzada com $k = 5$. A Tabela 5.28 contempla os resultados para o idioma inglês como alvo. Observa-se que houve uma melhoria em relação a estratégia anterior, sendo possível obter 94,00% como melhor métrica na medida-F1. Adicionalmente, foi possível alcançar valores superiores em relação às outras técnicas apresentadas, bem como do experimento base.

Tabela 5.28: Resultados obtidos com o XLM-Roberta no idioma inglês na estratégia JL/CL+.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	inglês	93,62%	94,08%	93,70%
italiano, filipino e alemão	**inglês**	**93,95%**	**94,41%**	**94,00%**
italiano, filipino, alemão e turco	inglês	93,86%	94,25%	93,86%

A Tabela 5.29 exibe os resultados para o idioma italiano utilizado como destino. É possível perceber que o comportamento dos resultados foi similar aos observados no idioma inglês como alvo. Sendo possível obter como melhor medida-F1 o valor de 92,37%, superando o resultado do modelo base e também os resultados das técnicas anteriores. Portanto, tanto para o idioma alvo inglês quanto o idioma alvo italiano, a técnica JL/CL+ apresentou ser mais eficiente em aprimorar o modelo base se comparado as outras técnicas.

Tabela 5.29: Resultados obtidos com o XLM-Roberta no idioma italiano na estratégia JL/CL+.

CIF	CID	Precisão	Revocação	Medida-F1
inglês e filipino	italiano	91,80%	91,56%	91,54%
inglês, filipino e alemão	italiano	90,75%	90,40%	90,38%
inglês, filipino, alemão e turco	**italiano**	**92,53%**	**92,37%**	**92,36%**

5.2.3 Experimentos com as estratégias CLL e idiomas com distância léxica menor

Esta seção contempla os resultados obtidos para os idiomas cuja distância léxica é menor. As estratégias CLL utilizadas foram as mesmas empregadas no experimento com maior distância léxica. Além disso, os idiomas utilizados foram os mesmos. Nas subseções seguintes, são apresentados os resultados obtidos para esse experimento.

A) Resultados provenientes da técnica ZST

Nessa técnica foi utilizado 90% dos dados do CIF no treino do modelo e 10% para a validação. Para o teste do modelo foi utilizado todo o CID. A Tabela 5.30 exibe os resultados para o idioma português como destino. É possível perceber um comportamento diferente nos resultados quando comparado com o experimento em que a distância léxica era maior. Neste experimento, foi possível obter um resultado superior ao modelo base já a partir da primeira técnica.

Tabela 5.30: Resultados obtidos com o XLM-Roberta português na estratégia ZST.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	86,45%	84,09%	85,14%

B) Resultados provenientes da técnica JL

Nessa técnica foram utilizados todos os dados do CIF acrescidos de 30% do CID para o treino do modelo. Desse total, 10% foi designado para validação do modelo. A Tabela 5.31 contempla os resultados obtidos nessa estratégia. Nota-se uma melhoria em relação a estratégia anterior, sendo possível obter cerca de 2,8% a mais na medida-F1. Além disso, foi possível superar o resultado do modelo base.

Tabela 5.31: Resultados obtidos com o XLM-Roberta português na estratégia JL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	87,55%	88,35%	87,92%

C) Resultados provenientes da técnica CL

Para essa técnica, inicialmente utilizou-se 90% do CIF para treinamento e 10% para validação durante o primeiro ajuste fino. Posteriormente, realizou-se um segundo ajuste fino, onde 70% dos dados do CID foram empregados para o treinamento do modelo e 10% para a validação, enquanto os 20% restantes dos dados foram reservados para o teste do modelo.

A Tabela 5.32 exibe os resultados para essa técnica. Observa-se uma redução na medida-F1 (82,10%) comparado às técnicas anteriores (87,92% e 85,14%). No entanto, mesmo tendo alcançado um resultado inferior, ainda foi possível superar o resultado obtido no experimento base. Sendo assim, é possível perceber que nessa técnica o uso de CLL com múltiplos idiomas no treino permitiu aprimorar o modelo.

Tabela 5.32: Resultados obtidos com o XLM-Roberta português na estratégia CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	89,07%	87,50%	82,10%

D) Resultados provenientes da técnica JL/CL

Neste experimento, na fase JL, utilizou-se 90% do CIF e 10% do CID para treinar o modelo. Para a etapa de validação, foram empregados 20% do CID e 10% do CIF. Os dados restantes do CID foram destinados para a etapa CL, onde 70% desses dados foram usados para o treinamento, 10% para a validação e 20% para o teste. A Tabela 5.33 contempla os resultados obtidos por meio dessa técnica. É possível perceber que, além de ter superado o resultado do modelo base, essa técnica forneceu o melhor resultado comparado às técnicas anteriores, obtendo 91,96% na medida-F1.

Tabela 5.33: Resultados obtidos com o XLM-Roberta português na técnica JL/CL.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	91,94%	91,96%	91,96%

E) Resultados provenientes da técnica JL/CL+

Por fim, temos a última técnica, a JL/CL+. Nessa técnica, foi utilizada a mesma configuração da técnica JL/CL, porém com a aplicação de uma validação cruzada com $k = 5$ na parte da técnica CL. A Tabela 5.34 exibe os resultados alcançados por meio dessa técnica. É possível notar que o resultado foi superior ao modelo base, bem como essa técnica apresentou o

melhor resultado comparado às técnicas anteriores, sendo possível obter 95,40% na medida-F1. Outro ponto a destacar é que o mesmo comportamento foi observado nos experimentos anteriores, tanto no codificador monolíngue quanto no experimento multilíngue com maior distância léxica.

Tabela 5.34: Resultados obtidos com o XLM-Roberta português na técnica JL/CL+.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	95,61%	95,70%	95,40%

5.3 Experimentos com Modelos do Tipo Decodificador

Esta etapa de experimentos contempla a utilização de modelos baseados em decodificadores. Para esses experimentos, foram empregados os modelos GPT-3 ADA e GPT-3.5 Turbo da OpenAI [70]. Em todos os modelos foram realizados ajustes finos. As configurações de cada ajuste fino foram as mesmas fornecidas e recomendadas pela OpenAI, o que inclui a quantidade de épocas de treinamento igual a 4 e a taxa de aprendizado igual a 1 x e^{-1}. Por ser um modelo decodificador, é necessário o uso de um *prompt* tanto no treinamento quanto nos testes. O *prompt* pode ser uma pergunta, uma afirmação ou uma instrução dirigida ao modelo. A Tabela 5.35 exibe os *prompts* empregados nos experimentos, abrangendo tanto o treinamento quanto os testes. No treino, o modelo recebeu um *prompt* que incluía uma pergunta e uma frase, acompanhados da respectiva resposta. Nos testes, foi fornecido apenas o *prompt* com a pergunta e a frase para análise, cabendo ao modelo gerar a resposta.

Tabela 5.35: *Prompts* fornecidos no treino e teste do modelo.

Etapa	Prompt
Treino	Responda apenas com Sim ou Não. A seguinte frase possui discurso de ódio? “Frase a ser analisada”. Resposta: “Sim ou Não”
Teste	Responda apenas com Sim ou Não. A seguinte frase possui discurso de ódio? “Frase a ser analisada”.

5.3.1 Experimento base sem CLL

Igualmente ao experimento com codificadores, no experimento com decodificadores também foram executados experimentos sem CLL para poder comparar posteriormente se ao adicionar CLL é possível aprimorar o modelo. Os resultados provenientes do experimento base com decodificadores estão presentes na Tabela 5.36. Nota-se que o modelo no idioma português alcançou o melhor valor na medida-F1, atingindo 92,65% nessa métrica, seguido pelos modelos em inglês e italiano, com 88,83% e 83,21%, respectivamente. Para cada experimento, foi empregado 70% do corpus para treino e 10% para validação. Os 20% dos dados remanescentes foram designados para o teste do modelo.

Tabela 5.36: Resultados obtidos com os modelos decodificadores.

Modelo	Corpus de treino	Corpus de teste	Precisão	Revocação	Medida-F1
GPT-3 ADA	inglês	inglês	88,87%	90,10%	88,83%
GPT-3 ADA	italiano	italiano	83,32%	83,19%	83,21%
GPT-3.5 Turbo	português	português	92,46%	92,96%	92,65%

É importante salientar que o idioma português foi executado em um modelo diferente do inglês e do italiano porque, no decorrer desta pesquisa, a OpenAI descontinuou o modelo GPT-3 ADA em seu sistema, impossibilitando a execução do experimento no idioma português usando esse modelo. Portanto, optou-se pela utilização do GPT-3.5 Turbo, outro modelo disponibilizado pela OpenAI após a descontinuação do GPT-3 ADA.

Ademais, devido à indisponibilidade do modelo ADA, não foi possível efetuar a validação cruzada nesse modelo. Contudo, embora fosse viável executar a validação cruzada no modelo GPT-3.5 Turbo, isso introduziria um viés nos resultados, já que teríamos o modelo no idioma português com validação cruzada, enquanto os modelos em inglês e italiano não teriam essa configuração. Para manter a integridade e a comparabilidade dos resultados, optou-se por não realizar a validação cruzada no modelo em português, mantendo assim a consistência com os experimentos executados nos modelos em inglês e italiano.

5.3.2 Experimentos com estratégias CLL e idiomas com distância léxica maior

Para esse experimento, foram selecionados os idiomas com maior distância léxica. Assim, para fins comparativos, foram empregados os idiomas inglês, turco, italiano e filipino. Os mesmos idiomas utilizados no experimento dos codificadores. Ademais, nos experimentos dos codificadores, a técnica que obteve os resultados mais promissores foi a JL/CL+. Essa técnica demonstrou resultados superiores tanto para os idiomas lexicalmente próximos quanto para os idiomas lexicalmente distantes. Por esse motivo, optou-se pela utilização apenas dessa técnica no experimento com decodificadores. Nas subseções seguintes, os resultados provenientes de cada experimento são apresentados.

Resultados provenientes da técnica JL/CL+

A Tabela 5.37 exibe os resultados derivados do experimento envolvendo dois idiomas como CIF e um idioma como CID. No experimento inicial, o italiano e o filipino foram empregados como CIF, enquanto o inglês serviu como CID. Nesse contexto, os resultados alcançados demonstraram uma melhoria, com uma medida-F1 de 96,43%. Esse resultado representa um aprimoramento substancial comparado ao experimento base. Na configuração subsequente, os idiomas filipino e inglês foram empregados como CIF, enquanto o italiano foi designado como CID. Nessa configuração, foi possível obter uma medida-F1 de 92,05%, um resultado melhor comparado ao experimento base.

Tabela 5.37: Resultados obtidos com o GPT-3 ADA com dois idiomas no treino.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e filipino	**inglês**	**96,50%**	**96,70%**	**96,43%**
inglês e filipino	italiano	92,20%	92,06%	92,05%

A Tabela 5.38 exibe os resultados alcançados no experimento realizado utilizando três idiomas como CIF. Nesse experimento, os mesmos corpora do experimento anterior foram empregados, com a inclusão do corpus em alemão. Observa-se uma redução no desempenho para os idiomas de destino (inglês e italiano) comparado aos resultados anteriores.

Tabela 5.38: Resultados obtidos com o GPT-3 ADA com três idiomas no treino.

CIF	CID	Precisão	Revocação	Medida-F1
italiano, filipino e alemão	**inglês**	**96,34%**	**96,44%**	**96,06%**
inglês, filipino e alemão	italiano	92,08%	91,88%	91,87%

Uma hipótese para essa queda no desempenho pode ser atribuída ao desbalanceamento significativo do corpus em alemão, no qual o número de textos sem discurso de ódio é substancialmente maior que os textos com discurso de ódio. O corpus no idioma alemão consiste em 4.318 textos sem discurso de ódio e 62 textos com discurso de ódio, resultando em uma proporção desbalanceada de 98,50% de instâncias sem discurso de ódio e 1,50% de instâncias com discurso de ódio. Entretanto, é importante destacar que, apesar dessa redução no desempenho, o modelo ainda alcançou resultados superiores comparado ao experimento base. Este resultado evidencia a eficácia do CLL em aprimorar as capacidades de classificação do modelo, mesmo em cenários com corpora desbalanceados.

A Tabela 5.39 exibe os resultados alcançados no experimento subsequente. Neste experimento, o corpus no idioma alemão foi removido do CIF, sendo substituído pelo corpus em turco, dado que este último é menos desbalanceado comparado ao corpus em alemão. Uma observação relevante neste experimento, comparado ao anterior, é a redução da medida-F1 para os idiomas de destino (inglês e italiano).

Tabela 5.39: Resultados obtidos com o GPT-3 ADA com três idiomas no treino sem a inclusão do idioma alemão.

CIF	CID	Precisão	Revocação	Medida-F1
italiano, filipino e turco	**inglês**	**96,31%**	**96,38%**	**95,92%**
inglês, filipino e turco	italiano	91,61%	91,50%	91,44%

A análise dos resultados revela uma leve redução no desempenho ao empregar o inglês como idioma de destino. Uma hipótese para esse declínio pode ser atribuída à proximidade linguística entre os corpora. No experimento anterior onde o alemão foi incluído no treino, apesar da distância léxica considerável entre o inglês e o alemão, ambos compartilham o

mesmo tronco linguístico. Em contrapartida, o turco é lexicalmente mais distante do inglês e não pertence ao mesmo tronco linguístico. Assim, a similaridade do tronco linguístico entre o inglês e o alemão pode ter contribuído para resultados melhores comparado ao corpus turco, apesar do desbalanceamento do corpus em alemão.

No que se refere aos resultados obtidos ao empregar o italiano como idioma de destino, é relevante salientar que a introdução de diferentes corpora não resultou em melhorias significativas. Uma possível hipótese para esse resultado reside na prática comum dos modelos, como o GPT, serem predominantemente pré-treinados com dados no idioma inglês. Dessa forma, ao incluir o italiano como idioma de destino, é possível que tenha sido introduzida uma complexidade adicional ao modelo na classificação, o que poderia ter mitigado qualquer potencial melhoria de desempenho.

A Tabela 5.40 contempla os resultados alcançados no experimento subsequente, onde quatro idiomas foram empregados como CIF. Nessa configuração específica, os corpora alemão e turco foram incluídos como CIF, enquanto o inglês foi designado como CID. Destaca-se que a medida-F1 atingiu 96,58%, representando o melhor valor para essa métrica comparada aos experimentos anteriores.

Tabela 5.40: Resultados obtidos com o GPT-3 ADA com quatro idiomas no treino.

CIF	CID	Precisão	Revocação	Medida-F1
italiano, filipino, alemão e turco	**inglês**	**96,89%**	**96,95%**	**96,58%**
inglês, filipino, alemão e turco	italiano	91,00%	90,81%	90,80%

Uma hipótese para essa melhoria nos resultados pode ser associada à similaridade linguística entre o inglês e o alemão, bem como à inclusão do idioma turco. No entanto, é crucial reconhecer que o corpus no idioma alemão continua desbalanceado, como mencionado em experimentos anteriores. É possível que essa questão de desbalanceamento tenha sido, em certa medida, mitigada pela inclusão do corpus em turco, que mostra um equilíbrio mais favorável comparado ao corpus em alemão. Isso pode ter facilitado ao modelo compreender melhor os padrões presentes nos textos com e sem discurso de ódio.

No que se refere ao italiano como idioma de destino, é pertinente destacar que os valores obtidos nos resultados revelaram uma diminuição comparado aos experimentos anteriores.

Enquanto o modelo com inglês como idioma de destino apresentou uma melhoria, a inclusão de idiomas adicionais resultou em uma redução na eficiência do modelo quando utilizado o italiano como idioma de destino.

A queda no desempenho sugere que a inclusão de idiomas no treino pode ter causado apenas um impacto marginal na eficiência do modelo italiano especificamente. Essa observação apoia a hipótese anterior de que boa parte dos modelos, como o GPT, são predominantemente pré-treinados usando o inglês como língua principal. Como resultado, a disparidade linguística entre os dados dos idiomas de treino e os dados do idioma de destino, introduz um nível mais elevado de complexidade para o modelo ao realizar tarefas de classificação em idiomas diferentes do inglês.

No entanto, é importante observar que, considerando todos os experimentos realizados, todos superaram o experimento base. Os resultados indicam que usar CLL, bem como a técnica JL/CL+ com múltiplos idiomas no treino, é uma estratégia promissora para aprimorar modelos baseados em decodificadores para detectar discurso de ódio.

5.3.3 Experimentos com estratégias CLL e idiomas com distância léxica menor

Nesse experimento, foram empregados idiomas com uma distância léxica menor. Portanto, foram selecionados os idiomas italiano, espanhol e português. Os mesmos utilizados no experimento dos codificadores. Igualmente ao experimento anterior com decodificadores, nesse experimento, apenas a técnica JL/CL+ foi empregada, pois obteve os resultados mais promissores nos experimentos que foram detalhados nas seções anteriores. Na subseção seguinte, os resultados provenientes do experimento são apresentados.

Resultados provenientes da técnica JL/CL+

A Tabela 5.41 exibe os resultados alcançados com os idiomas italiano e espanhol como CIF e o idioma português como CID. É notável que os valores obtidos nos resultados foram inferiores quando comparados aos resultados com maior distância léxica. Enquanto no experimento com maior distância léxica o melhor resultado foi 96,58% na medida-F1 para o idioma alvo inglês, o resultado com menor distância léxica atingiu 93,88% na mesma métrica para o idi-

oma alvo português. Apesar do resultado ter sido inferior, o modelo decodificador com a distância léxica menor conseguiu superar o experimento base. Isso evidencia que as técnicas de CLL podem aprimorar a eficiência do modelo.

Tabela 5.41: Resultados obtidos com o GPT-3.5 Turbo na estratégia JL/CL+ e português como idioma de destino.

CIF	CID	Precisão	Revocação	Medida-F1
italiano e espanhol	português	94,56%	94,30%	93,88%

5.4 Análise dos Resultados

Os resultados dos modelos do tipo codificador demonstraram uma investigação detalhada do impacto ao utilizar CLL para melhorar a eficiência dos modelos, especialmente em tarefas voltadas a PLN. O experimento base, sem CLL, proporcionou uma direção para comparação com as estratégias que abrangem CLL. Os resultados indicaram desempenho considerável dos modelos codificadores BERT e XLM-Roberta nos idiomas italiano, inglês e português, com métricas de precisão, revocação e medida-F1 variando de 80% a 96%.

Algumas estratégias CLL foram aplicadas e comparadas com o modelo base. No que se trata da técnica ZST, a mesma não conseguiu superar o resultado do experimento base, com exceção do experimento com a distância léxica menor no modelo XLM-Roberta. Embora tenha havido melhorias em certos casos, os resultados gerais não foram superiores ao modelo base. Isso sugere que simplesmente adicionar mais idiomas sem uma estratégia específica de transferência de aprendizado pode não ser suficiente para melhorar a eficiência do modelo.

Já a técnica JL demonstrou melhorias significativas com relação à estratégia ZST. Ao incorporar parte dos dados do idioma de destino no treinamento, os modelos apresentaram um desempenho melhorado, especialmente quando mais idiomas foram adicionados como fonte de treinamento. Isso sugere que, dependendo da estratégia, a combinação de múltiplos idiomas durante o treino do modelo pode melhorar a capacidade do modelo de generalizar para idiomas diferentes.

Em relação à técnica CL, de maneira geral, mostrou-se ainda mais eficaz do que a técnica

JL, com exceção do experimento com menor distância léxica no modelo XLM-Roberta, no qual o resultado foi inferior à técnica JL. Isso sugere que, para modelos codificadores multilíngues, essa técnica pode não favorecer a melhoria do desempenho desses tipos de modelos. No entanto, com exceção desse caso isolado, essa técnica se mostrou mais eficiente nos outros experimentos com codificadores. Sendo assim, para esses casos, ao realizar ajustes finos com dados do CID após o treinamento inicial com dados do CIF, os modelos apresentaram melhorias adicionais no desempenho. Isso pode indicar que o ajuste fino com dados do CID pode adaptar melhor o modelo às nuances específicas do idioma de destino.

Já as estratégias JL/CL e JL/CL+, que combinam as técnicas JL e CL, apresentaram resultados superiores comparados às outras estratégias. Principalmente a técnica JL/CL+, sendo esta a melhor estratégia dentre todas as apresentadas. Ao combinar múltiplos idiomas atrelados à abordagem da técnica JL aliado aos ajustes finos da técnica CL, os modelos alcançaram os melhores desempenhos em todas as métricas avaliadas.

Outro ponto a destacar é o desempenho dos modelos monolíngues em comparação com o modelo multilíngue. No geral, os resultados obtidos pelos modelos monolíngues foram superiores aos do modelo multilíngue. Uma hipótese para isso pode estar relacionada à especialização do modelo no idioma de destino, pois os modelos monolíngues são pré-treinados especificamente com o mesmo idioma utilizado no destino. Portanto, os pesos aprendidos pelo modelo podem estar adaptados ao idioma de destino. Por outro lado, no modelo multilíngue, os pesos contemplam valores agregados de múltiplos idiomas, o que pode dificultar uma melhor adaptação ao idioma de destino e, consequentemente, dificultar a obtenção de melhores resultados.

Considerando os modelos baseados em decodificadores, apenas a técnica JL/CL+ foi aplicada, por esta ter sido a melhor nos experimentos dos codificadores. No geral, os resultados variaram conforme a inclusão dos idiomas no treino do modelo, bem como a inclusão dos idiomas de destino. Observou-se que usar múltiplos idiomas no treino trouxe melhorias significativas no desempenho, especialmente quando os idiomas compartilhavam o tronco linguístico semelhante.

No entanto, quando o italiano foi empregado como idioma de destino o modelo demonstrou um decréscimo nos resultados, conforme mais idiomas foram adicionados ao treino. Uma hipótese sugerida para essa redução pode estar na forma como os modelos são treina-

dos, pois pelo fato do inglês ser um idioma mais conhecido, no geral esse idioma acaba sendo mais predominante na maioria do treinamento de modelos, mesmo em modelos multilíngues como o GPT.

Portanto, a redução no desempenho devido o italiano ser usado como idioma de destino pode ter ocorrido devido à complexidade adicional introduzida pela diversidade linguística. Um ponto importante a ser destacado é que, apesar das variações nos resultados nos decodificadores, todos os experimentos usando CLL superaram o experimento base, demonstrando a eficácia dessa abordagem no aprimoramento dos modelos para detectar discurso de ódio.

No quesito recursos, vale destacar também o tempo de treino dos modelos codificadores em comparação com os modelos decodificadores. Nos modelos codificadores, o treino e os ajustes finos foram mais rápidos e consumiram bem menos *hardware* em comparação com os modelos decodificadores, que se mostraram mais demorados e consumiram mais recursos. No entanto, levando em consideração os resultados obtidos pelos modelos, vale destacar que os resultados alcançados nesta pesquisa mostraram-se superiores aos de outras pesquisas que empregaram o mesmo corpus como idioma de destino. A Tabela 5.42 exibe os resultados provenientes de outras pesquisas para o corpus no idioma inglês.

Tabela 5.42: Resultados obtidos em outros trabalhos para o corpus inglês.

Autor	Melhor Resultado na Métrica Medida-F1
Xinlei He et al. (2024) [44]	83,30%
Grimminger e Klinger (2021) [41]	74,00%
Li et al. (2021) [56]	74,52% (média macro)
Clark et al. (2021) [15]	85,70%
Este trabalho	**96,58%**

Destaca-se que dentre os trabalhos apresentados, este trabalho alcançou a melhor métrica para o mesmo corpus inglês. A Tabela 5.43 exibe os resultados provenientes de outras pesquisas para o corpus no idioma italiano. É possível notar que, dentre os resultados apresentados, este trabalho alcançou a melhor métrica comparado aos demais trabalhos apresentados para o mesmo corpus italiano.

Tabela 5.43: Resultados obtidos em outros trabalhos para o corpus italiano.

Autor	Melhor Resultado na Métrica Medida-F1
Fagni et al. (2019) [29]	77,50%
Vigna et al. (2018) [99]	78,45%
Fortuna et al. (2018) [33]	72,30%
Bosco et al. (2018) [9]	82,88% (média macro)
Cimino et al. (2018) [13]	86,00%
Este trabalho	**93,00%**

Sendo assim, tendo como base os resultados apresentados neste trabalho, podemos responder as questões de pesquisa propostas no Capítulo 1:

Questão 1: É possível aumentar a eficiência do modelo base ao utilizar CLL atrelado ao treinamento em múltiplos idiomas que sejam distintos do idioma de destino?

Os experimentos com modelos codificadores revelaram que certas estratégias baseadas em CLL, combinadas com a inclusão de vários idiomas no treino, resultaram em melhorias significativas no desempenho comparado ao modelo base. Entre essas estratégias, destaca-se a JL/CL+, a qual, quando combinada com múltiplos idiomas no treino, permitiu que o modelo BERT alcançasse uma medida-F1 de até 96,92%, superando o experimento base. Da mesma maneira, os resultados dos modelos decodificadores também evidenciaram melhorias ao utilizar CLL aliado a inclusão de múltiplos idiomas no treino do modelo. Assim, esta abordagem possibilitou alcançar uma medida-F1 de até 96,58%, também superando o experimento base. Portanto, com base nos resultados apresentados, podemos concluir que utilizar CLL combinado com o treinamento em múltiplos idiomas distintos do idioma de destino tende a aprimorar a eficiência do modelo em detectar discurso de ódio.

Questão 2: Ao utilizar CLL em modelos codificadores, a eficiência do modelo é melhor quando a distância léxica dos idiomas utilizados no treino é mais próxima ou distante do idioma de destino?

Tomando como base os melhores resultados obtidos com os modelos codificadores, nos quais foram utilizados dados cuja distância léxica dos idiomas de treino era mais distante em relação ao idioma de destino, foi possível obter como melhor medida-F1 os valores 94,80%

e 93,00% para os idiomas de destino inglês e italiano respectivamente. Já na situação em que foram utilizados dados cuja distância léxica dos idiomas de treino era mais próxima do idioma de destino, foi possível obter, como melhor medida-F1, o valor de 96,92% para o idioma de destino português.

No entanto, os valores obtidos entre os modelos foram muito próximos. Portanto, para compará-los foi conduzido um teste de significância empregando a abordagem ASO. Primeiramente, foi efetuado o teste nos modelos que contemplam os resultados com dados que possuem a distância léxica mais ampla, ou seja, os modelos inglês e italiano.

A Tabela 5.44 exibe o resultado entre ambos os modelos. Como é possível perceber, o valor θ foi superior a 0,5. Dessa forma descartamos a hipótese alternativa (H1) e aceitamos a hipótese nula (H0). Sendo assim, podemos concluir que o modelo italiano foi superior ao modelo inglês.

Tabela 5.44: Teste de significância entre o modelo inglês e italiano.

Hipótese	Valor θ
H0: O modelo inglês não foi superior ao modelo italiano	0,65
H1: O modelo inglês foi superior ao modelo italiano	

Agora que sabemos qual o melhor modelo referente à distância léxica mais ampla, vamos compará-lo com o modelo que foi treinado com dados que possuem uma distância léxica mais próxima, ou seja, o modelo português. A Tabela 5.45 exibe o resultado referente ao teste de significância entre o modelo italiano e português. Como é possível notar o valor θ foi inferior a 0,5.

Tabela 5.45: Teste de significância entre o modelo português e italiano.

Hipótese	Valor θ
H0: O modelo português não foi superior ao modelo italiano	0,19
H1: O modelo português foi superior ao modelo italiano	

Assim, descartamos a hipótese H0 e aceitamos a hipótese H1, ou seja, o modelo português foi superior ao modelo italiano. Portanto, com base nos resultados obtidos, podemos

concluir que a inclusão de CLL em modelos codificadores é mais eficiente quando a distância léxica dos idiomas de treino é mais próxima do idioma de destino.

Questão 3: Ao utilizar CLL em modelos decodificadores, a eficiência do modelo é melhor quando a distância léxica dos idiomas utilizados no treino é mais próxima ou distante do idioma de destino?

Assim como na questão anterior, para responder à terceira questão de pesquisa, foi empregado o teste de significância. Portanto, para efetuar a primeira comparação foram selecionados os melhores modelos cuja distância léxica dos idiomas utilizados no treino é mais distante do idioma de destino. Assim, nos modelos com maior distância léxica, as melhores medidas-F1 foram de 96,58% para o modelo inglês e de 92,05% para o modelo italiano.

É possível notar que os valores obtidos entre os dois modelos foram próximos, com uma leve vantagem para o modelo inglês. Para determinar qual dos dois foi superior, foi efetuado um teste de significância. A Tabela 5.46 contempla o resultado desse teste. Como o valor de θ foi inferior a 0,5, rejeitamos a hipótese nula (H0) e aceitamos a hipótese alternativa (H1). Portanto, concluímos que o modelo inglês foi superior ao modelo italiano.

Tabela 5.46: Teste de significância entre o modelo inglês e italiano.

Hipótese	Valor θ
H0: O modelo inglês não foi superior ao modelo italiano	0,23
H1: O modelo inglês foi superior ao modelo italiano	

Para obtermos a resposta para a questão de pesquisa, é necessário comparar o modelo inglês com o modelo que foi treinado com idiomas que possuem a distância léxica mais próxima do idioma de destino, ou seja, o modelo português. Este modelo obteve como melhor medida-F1 o valor de 93,88%.

A Tabela 5.47 contempla os resultados do teste de significância entre esses modelos. Observa-se que o valor de θ foi inferior a 0,5. Assim, rejeitamos a hipótese nula (H0) e aceitamos a hipótese alternativa (H1). Consequentemente, o modelo inglês demonstrou ser superior ao modelo português. Portanto, com base nos experimentos realizados, podemos concluir que a inclusão de CLL em modelos decodificadores é mais eficiente quando a distância léxica dos idiomas de treino é mais distante do idioma de destino.

Tabela 5.47: Teste de significância entre o modelo inglês e português.

Hipótese	Valor θ
H0: O modelo inglês não foi superior ao modelo português	0,37
H1: O modelo inglês foi superior ao modelo português	

5.5 Experimento Prático

Foi conduzido um experimento prático para verificar a eficácia dos modelos desenvolvidos. O experimento envolveu a utilização dos modelos treinados em português para detectar discurso de ódio em um corpus que os modelos nunca tinham visto anteriormente. Para isso, empregou-se o corpus HateBR, elaborado por Vargas et al. (2022) [97]. O corpus consiste em textos no idioma português, coletados do Instagram durante as eleições presidenciais do Brasil em 2022. O corpus foi rotulado manualmente, incluindo 3.500 textos categorizados como não contendo discurso de ódio, 2.798 textos rotulados como contendo conteúdo ofensivo e 702 textos identificados como contendo discurso de ódio.

Como os modelos desenvolvidos nesta dissertação foram treinados para classificar textos com ou sem discurso de ódio, e não especificamente para detectar conteúdo ofensivo, optou-se pela utilização no teste prático os 3.500 textos classificados como não contendo discurso de ódio e os 702 textos identificados como contendo discurso de ódio, totalizando 4.202 textos. Ademais, dado que os dados estão no idioma português, o teste prático foi realizado apenas com os modelos BERT português e GPT-3.5 Turbo, os quais foram treinados para detectar discurso de ódio nesse idioma. É importante destacar que os modelos não foram treinados com esse corpus específico. Portanto, os dados foram apresentados aos modelos exclusivamente para a classificação dos textos fornecidos como entrada.

A Tabela 5.48 exibe os resultados da classificação realizada pelos modelos para o corpus HateBR. Observa-se que as taxas de acerto de ambos os modelos foram muito próximas. No entanto, o modelo BERT obteve uma leve vantagem, com uma taxa ligeiramente superior de acertos em comparação ao modelo GPT.

Tabela 5.48: Resultados obtidos no teste prático com o BERT português e o GPT-3.5 Turbo português.

Modelo	Classificação Correta	Classificação Incorreta	Taxa de Acerto
BERT português	**3.795**	**407**	**90,31%**
GPT 3.5 Turbo português	3.776	426	89,86%

Os resultados do experimento prático também foram utilizados para comparação com outros trabalhos que empregaram o mesmo corpus como teste. A Tabela 5.49 apresenta os resultados de outros autores para o mesmo corpus. É importante destacar que, no experimento prático desta pesquisa, os modelos não utilizaram os dados do corpus no treinamento, enquanto os outros trabalhos relatados empregaram o corpus para treinar o modelo. Contudo, mesmo não tendo utilizado o corpus no treino, observa-se que o experimento prático desta dissertação obteve o melhor resultado dentre os trabalhos apresentados, tanto no modelo BERT quanto no modelo GPT.

Tabela 5.49: Comparação dos resultados do experimento prático com outros trabalhos que utilizaram o mesmo corpus.

Autor	Modelo	Melhor Resultado na Medida-F1
Vargas et al. (2022) [97]	Naive Bayes	78,00%
Assis et al. (2024) [7]	BERTweet.BR	82,20% (média)
Assis et al. (2024) [7]	GPT	76,60% (média)
Vargas et al. (2023) [96]	BoW + MOL	86,00%
Este trabalho	GPT 3.5 Turbo	88,32%
Este trabalho	**BERT**	**88,90%**

5.6 Sumário dos Resultados dos Experimentos

Nesta seção, é apresentado um sumário de todos os experimentos realizados nesta dissertação. O sumário está dividido em experimentos que não utilizaram validação cruzada e aqueles que a utilizaram. Além disso, os resultados obtidos são apresentados com base na distância léxica empregada nos experimentos.

5.6.1 Sumário dos experimentos sem validação cruzada

A Tabela 5.50 contempla os resultados obtidos sem validação cruzada para a distância léxica mais ampla. Já a Tabela 5.51 mostra os resultados obtidos com a distância léxica menor. O melhor resultado de cada técnica está marcado em negrito.

Tabela 5.50: Sumário dos resultados dos experimentos com a distância léxica mais ampla.

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
Base sem CLL	BERT inglês	inglês	inglês	80,00%	90,00%	85,00%
	XLM-Roberta	inglês	inglês	80,70%	89,83%	85,02%
	GPT-3 ADA	**inglês**	**inglês**	**88,87%**	**90,10%**	**88,83%**
	BERT italiano	italiano	italiano	81,00%	84,00%	82,00%
	XLM-Roberta	italiano	italiano	79,00%	79,00%	79,00%
	GPT-3 ADA	italiano	italiano	83,32%	83,19%	83,21%
ZST	**BERT inglês**	**italiano e filipino**	**inglês**	**86,00%**	**82,00%**	**84,00%**

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
ZST	BERT inglês	italiano, filipino e alemão	inglês	87,00%	81,00%	84,00%
	BERT inglês	italiano, filipino, alemão e turco	inglês	87,00%	78,00%	82,00%
	XLM-Roberta	italiano e filipino	inglês	87,01%	70,93%	75,94%
	XLM-Roberta	italiano, filipino e alemão	inglês	86,91%	71,17%	76,11%
	XLM-Roberta	italiano, filipino, alemão e turco	inglês	86,97%	70,33%	75,48%
	BERT italiano	inglês e filipino	italiano	59,00%	51,00%	34,00%
	BERT italiano	inglês, filipino e alemão	italiano	68,00%	52,00%	36,00%
	BERT italiano	inglês, filipino, alemão e turco	italiano	69,00%	52,00%	37,00%
	XLM-Roberta	inglês e filipino	italiano	76,47%	73,58%	73,02%

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
ZST	XLM-Roberta	inglês, filipino e alemão	italiano	73,67%	63,28%	59,38%
	XLM-Roberta	inglês, filipino, alemão e turco	italiano	75,51%	70,20%	68,87%
JL	BERT inglês	italiano e filipino	inglês	84,00%	88,00%	86,00%
	BERT inglês	italiano, filipino e alemão	inglês	86,00%	87,00%	86,00%
	BERT inglês	**italiano, filipino, alemão e turco**	**inglês**	**87,00%**	**87,00%**	**87,00%**
	XLM-Roberta	italiano e filipino	inglês	85,52%	83,84%	84,60%
	XLM-Roberta	italiano, filipino e alemão	inglês	85,88%	86,13%	86,00%
	XLM-Roberta	italiano, filipino, alemão e turco	inglês	85,28%	83,19%	84,12%
	BERT italiano	inglês e filipino	italiano	79,00%	87,00%	83,00%

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL	BERT italiano	inglês, filipino e alemão	italiano	82,00%	82,00%	82,00%
	BERT italiano	inglês, filipino, alemão e turco	italiano	82,12%	82,10%	82,10%
	XLM-Roberta	inglês e filipino	italiano	81,24%	81,23%	81,23%
	XLM-Roberta	inglês, filipino e alemão	italiano	80,19%	80,18%	80,18%
	XLM-Roberta	inglês, filipino, alemão e turco	italiano	80,54%	80,56%	80,56%
CL	BERT inglês	italiano e filipino	inglês	88,00%	90,00%	89,00%
	BERT inglês	italiano, filipino e alemão	inglês	87,00%	90,00%	87,00%
	BERT inglês	italiano, filipino, alemão e turco	inglês	88,00%	90,00%	89,00%
	XLM-Roberta	italiano e filipino	inglês	90,39%	91,17%	90,69%

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
CL	**XLM-Roberta**	**italiano, filipino e alemão**	**inglês**	**91,20%**	**91,50%**	**91,34%**
	XLM-Roberta	italiano, filipino, alemão e turco	inglês	90,50%	91,00%	90,72%
	BERT italiano	inglês e filipino	italiano	86,00%	86,00%	86,00%
	BERT italiano	inglês, filipino e alemão	italiano	82,00%	86,00%	84,00%
	BERT italiano	inglês, filipino, alemão e turco	italiano	86,00%	85,00%	86,00%
	XLM-Roberta	inglês e filipino	italiano	85,27%	85,25%	85,25%
	XLM-Roberta	inglês, filipino e alemão	italiano	85,54%	85,50%	85,49%
	XLM-Roberta	inglês, filipino, alemão e turco	italiano	84,50%	84,50%	84,50%
JL/CL	BERT Inglês	italiano e filipino	inglês	85,00%	89,00%	84,31%

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL/CL	BERT Inglês	italiano, filipino e alemão	inglês	88,00%	90,00%	86,00%
	BERT Inglês	italiano, filipino, alemão e turco	inglês	87,00%	89,00%	87,10%
	XLM-Roberta	**italiano e filipino**	**inglês**	**87,26%**	**88,86%**	**87,48%**
	XLM-Roberta	italiano, filipino e alemão	inglês	85,68%	87,50%	86,23%
	XLM-Roberta	italiano, filipino, alemão e turco	inglês	86,33%	88,32%	86,48%
	BERT italiano	inglês e filipino	italiano	85,00%	84,00%	84,26%
	BERT italiano	inglês, filipino e alemão	italiano	85,00%	84,00%	84,05%
	BERT italiano	inglês, filipino, alemão e turco	italiano	86,00%	85,00%	85,25%
	XLM-Roberta	inglês e filipino	italiano	82,22%	81,25%	81,25%

Continua na próxima página

Tabela 5.50 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL/CL	XLM-Roberta	inglês, filipino e alemão	italiano	82,50%	82,14%	82,17%
	XLM-Roberta	inglês, filipino, alemão e turco	italiano	82,19%	81,79%	81,81%

Tabela 5.51: Sumário dos resultados dos experimentos com a distância léxica menor.

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
Base sem CLL	BERT português	português	português	87,49%	89,06%	87,62%
	XLM-Roberta	português	português	75,74%	87,03%	81,00%
	GPT-3.5 Turbo	**português**	**português**	**92,46%**	**92,96%**	**92,65%**
ZST	**BERT Portu-guês**	**italiano e espanhol**	**português**	**85,77%**	**88,34%**	**86,67%**
	XLM-Roberta	italiano e espanhol	português	86,45%	84,09%	85,14%
JL	BERT Português	italiano e espanhol	português	86,61%	89,78%	87,37%

Continua na próxima página

Tabela 5.51 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL	**XLM-Roberta**	**italiano e espanhol**	**português**	**87,55%**	**88,35%**	**87,92%**
CL	**BERT Português**	**italiano e espanhol**	**português**	**89,64%**	**90,62%**	**89,64%**
	XLM-Roberta	italiano e espanhol	português	89,07%	87,50%	82,10%
JL/CL	**BERT Português**	**italiano e espanhol**	**português**	**93,54%**	**92,41%**	**92,86%**
	XLM-Roberta	italiano e espanhol	português	91,94%	91,96%	91,96%

5.6.2 Sumário dos experimentos com validação cruzada

A Tabela 5.52 contempla os resultados obtidos com validação cruzada para a distância léxica mais ampla e a Tabela 5.53 mostra os resultados obtidos com a distância léxica menor. O melhor resultado de cada técnica está marcado em negrito.

Tabela 5.52: Sumário dos resultados dos experimentos com a distância léxica maior com validação cruzada.

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
Base sem CLL	**BERT inglês**	**inglês**	**inglês**	**81,11%**	**90,07%**	**85,30%**
	XLM-Roberta	inglês	inglês	81,05%	90,03%	85,30%

Continua na próxima página

Tabela 5.52 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
Base sem CLL	BERT italiano	italiano	italiano	83,06%	82,95%	82,93%
	XLM-Roberta	italiano	italiano	75,51%	74,81%	74,64%
JL/CL+	BERT Inglês	italiano e filipino	inglês	94,6%	94,00%	94,30%
	BERT Inglês	italiano, filipino e alemão	inglês	94,60%	95,00%	94,80%
	BERT Inglês	italiano, filipino, alemão e turco	inglês	94,80%	94,00%	94,40%
	XLM-Roberta	italiano e filipino	inglês	93,62%	94,08%	93,70%
	XLM-Roberta	italiano, filipino e alemão	inglês	93,95%	94,41%	94,00%
	XLM-Roberta	italiano, filipino, alemão e turco	inglês	93,86%	94,25%	93,86%
	GPT-3 ADA	italiano e filipino	inglês	96,50%	96,70%	96,43%
	GPT-3 ADA	italiano, filipino e alemão	inglês	96,34%	96,44%	96,06%

Continua na próxima página

Tabela 5.52 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL/CL+	GPT-3 ADA	italiano, filipino e turco	inglês	96,31%	96,38%	95,92%
	GPT-3 ADA	**italiano, filipino, alemão e turco**	**inglês**	**96,89%**	**96,95%**	**96,58%**
	BERT Italiano	inglês e filipino	italiano	92,80%	92,00%	92,40%
	BERT Italiano	inglês, filipino e alemão	italiano	92,70%	92,50%	92,60%
	BERT Italiano	inglês, filipino, alemão e turco	italiano	92,00%	94,00%	93,00%
	XLM-Roberta	inglês e filipino	italiano	91,80%	91,56%	91,54%
	XLM-Roberta	inglês, filipino e alemão	italiano	90,75%	90,40%	90,38%
	XLM-Roberta	inglês, filipino, alemão e turco	italiano	92,53%	92,37%	92,36%
	GPT-3 ADA	inglês e filipino	italiano	92,20%	92,06%	92,05%

Continua na próxima página

Tabela 5.52 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL/CL+	GPT-3 ADA	inglês, filipino e alemão	italiano	92,08%	91,88%	91,87%
	GPT-3 ADA	inglês, filipino e turco	italiano	91,61%	91,50%	91,44%
	GPT-3 ADA	inglês, filipino, alemão e turco	italiano	91,00%	90,81%	90,80%

Tabela 5.53: Sumário dos resultados dos experimentos com a distância léxica menor com validação cruzada.

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
Base sem CLL	**BERT português**	**português**	**português**	**90,84%**	**91,81%**	**90,95%**
	XLM-Roberta	português	português	79,88%	89,38%	84,36%
JL/CL+	**BERT Português**	**italiano e espanhol**	**português**	**96,80%**	**97,00%**	**96,92%**
	XLM-Roberta	italiano e espanhol	português	95,61%	95,70%	95,40%

Continua na próxima página

Tabela 5.53 – Continuação da página anterior

Técnica	Modelo	CIF	CID	Precisão	Revocação	Medida-F1
JL/CL+	GPT-3.5 Turbo	italiano e espanhol	português	94,56%	94,30%	93,88%

5.7 Considerações Finais

Neste capítulo, foram delineados os experimentos que possuem como propósito detectar discurso de ódio. Foram conduzidos experimentos com modelos baseados em codificadores e decodificadores. Esses experimentos, exploraram tanto a aplicação de idiomas que possuem distância léxica mais ampla quanto com distância léxica mais próxima. Os resultados foram discutidos e analisados para responder às questões de pesquisa estabelecidas neste trabalho. Por fim, foi conduzido um experimento prático para verificar a eficiência dos modelos na prática. No próximo capítulo, serão apresentadas as conclusões finais decorrentes desta pesquisa.

Capítulo 6

Conclusão

Neste capítulo, são apresentadas as considerações finais desta pesquisa, resumindo os principais resultados alcançados e destacando sua relevância para o campo de pesquisa relacionado à detecção de discurso de ódio. Ademais, são apontadas as limitações encontradas durante o desenvolvimento do trabalho e identificadas possíveis áreas para investigações futuras, visando aprimorar e expandir os conhecimentos obtidos.

6.1 Considerações Finais

Durante muitos anos, a civilização tem buscado avançar em tecnologias visando automatizar ou facilitar as atividades cotidianas. A comunicação é uma dessas tecnologias, que evoluiu ao longo do tempo, especialmente com o advento da internet. Como resultado, pessoas no mundo todo podem agora se comunicar em tempo real, compartilhando suas opiniões publicamente e exercendo sua liberdade de expressão [59].

Apesar de ser algo fundamental na sociedade, o uso inadequado da liberdade de expressão pode levar a outros problemas, como o discurso de ódio, uma expressão que promove a violência e incita ataques a grupos ou indivíduos. Geralmente, esses ataques são direcionados a diversos domínios, como religião, política, nacionalidade, etc. A propagação de mensagens com discurso de ódio em plataformas online pode causar efeitos psicológicos negativos em indivíduos que são alvos desses ataques, podendo resultar em problemas mais graves, como depressão e ansiedade [68; 100]. Portanto, detectar discurso de ódio é crucial para mitigar o impacto negativo desse fenômeno na sociedade.

Sendo assim, este trabalho explorou a aplicação de técnicas relacionadas à AM para detectar discurso de ódio em textos provenientes de redes sociais. Foram empregados modelos baseados em codificadores e decodificadores, os quais usam uma arquitetura baseada em *Transformers*. Adicionalmente, foram empregadas técnicas baseadas em CLL para aprimorar a detecção de discurso de ódio em diferentes idiomas, bem como foi realizada uma análise sobre os modelos baseados em codificadores e decodificadores para avaliar a adaptação desses modelos na detecção de discurso de ódio.

A primeira parte dos experimentos desta pesquisa concentrou-se na avaliação de modelos baseados em codificadores. Por meio de experimentos detalhados, foi evidenciado empiricamente que a técnica CLL, especialmente quando combinada com múltiplos idiomas no treinamento do modelo, como na abordagem JL/CL+, resultou em melhorias significativas na eficácia dos modelos em detectar discurso de ódio. Na segunda parte dos experimentos, foi realizada uma investigação para modelos baseados em decodificadores. Da mesma maneira que os experimentos com codificadores, os experimentos com decodificadores revelaram que a técnica CLL, especialmente na abordagem JL/CL+, obteve melhorias na eficácia dos modelos na detecção de discurso de ódio.

Ademais, foi observado que tanto nos codificadores quanto nos decodificadores, o uso de múltiplos idiomas como corpora de treino ajudou a aprimorar os modelos na classificação de discurso de ódio, mesmo que o idioma de destino seja diferente dos idiomas utilizados no treinamento. Esse resultado sugere que é viável utilizar corpora com múltiplos idiomas para efetuar a classificação de discurso de ódio, ajudando a superar problemas de escassez de dados em determinados idiomas. Contudo, outro ponto a destacar é o desbalanceamento dos dados. O fato de serem desbalanceados pode influenciar o desempenho do modelo, como aconteceu no modelo GPT, no qual o resultado na medida-F1 foi reduzido ao adicionar o idioma alemão, que é o mais desbalanceado do corpora.

Em suma, os resultados indicaram que ao empregar CLL nos modelos, houve uma melhora na eficiência dos modelos na classificação de discurso de ódio. Adicionalmente, verificou-se que os modelos monolíngues baseados em codificadores foram mais eficientes quando a distância léxica dos idiomas de treino era mais próxima do idioma de destino. Já os modelos baseados em decodificadores, apresentaram uma eficiência melhor quando a distância léxica dos idiomas de treino era mais distante do idioma de destino. Uma hipótese

para o modelo monolíngue ter alcançado um resultado melhor quando a distância léxica foi menor pode estar relacionada à especialização do modelo no idioma de destino. Os modelos monolíngues são pré-treinados especificamente com o mesmo idioma utilizado no destino. Portanto, como os idiomas no treino são próximos lexicalmente do idioma de destino, isso pode ter colaborado para o resultado ter sido melhor para esse tipo de modelo. Já nos modelos decodificadores, uma hipótese pode ser que o fato de serem pré-treinados em múltiplos idiomas tenha favorecido um resultado mais eficiente quando a distância léxica entre os idiomas era mais ampla.

Por fim, este trabalho oferece novas estratégias e possibilidades para a elaboração de ferramentas mais robustas e eficazes para detectar discurso de ódio, demonstrando o potencial de aproveitar a diversidade linguística em sistemas de Inteligência Artificial, bem como destaca-se a importância de utilizar diversos recursos linguísticos e métodos estratégicos como o CLL para ampliar a eficácia dos modelos em detectar discurso de ódio. Destarte, esta contribuição oferece percepções que podem beneficiar pesquisas futuras no campo de detecção de discurso de ódio.

6.2 Limitações

Uma das limitações desta dissertação está relacionada ao escopo da pesquisa, que foi restrito apenas aos corpora políticos. Como resultado, não se pode inferir por meio dos resultados obtidos neste trabalho, se os resultados seriam os mesmos quando os modelos fossem expostos a corpora que abrangem vários tipos de discurso de ódio simultaneamente, incluindo tipos como religião, racismo, sexismo, dentre outros.

Ademais, não foram realizados experimentos com modelos decodificadores pré-treinados em apenas um idioma. Portanto, não se sabe se o modelo seria mais eficiente nesse tipo de situação. Um outro ponto a ser considerado é a subjetividade inerente à rotulagem dos dados, pois devido à participação de diferentes anotadores durante o processo de rotulagem, pode haver um grau de subjetividade introduzido, o que significa que textos específicos classificados como discurso de ódio por um anotador podem não receber a mesma classificação de outro anotador ao rotular o mesmo texto. Assim, essa subjetividade pode afetar positivamente ou negativamente a eficiência dos modelos.

Outro fator a ser considerado está relacionado ao uso de apenas modelos como o BERT, GPT-3 e GPT-3.5 Turbo. Seria necessário efetuar experimentos utilizando as mesmas estratégias voltadas à CLL apresentadas neste trabalho com outros codificadores além do BERT, bem como experimentos com outros modelos decodificadores além do GPT. Dessa forma, poderíamos verificar se é possível obter a mesma eficácia e resultados similares aos obtidos neste trabalho em outros modelos.

6.3 Trabalhos Futuros

Existem diversas direções para trabalhos futuros que podem expandir e aprofundar os resultados obtidos nesta pesquisa, a saber:

- Investigar o uso de decodificadores pré-treinados exclusivamente no mesmo idioma de destino e avaliar se é possível aprimorar os resultados ao lidar com dados com distância léxica mais próxima.
- Explorar a eficácia de codificadores pré-treinados em múltiplos idiomas para investigar se é possível aprimorar os resultados ao lidar com dados com distância léxica mais ampla.
- Realizar experimentos adicionais considerando dados balanceados, o que pode oferecer novas percepções sobre o comportamento dos modelos sob essas circunstâncias.
- Verificar a consistência dos resultados obtidos nesta pesquisa empregando outros modelos baseados em codificadores e decodificadores, tais como: T5 [78], BART [54] e ELECTRA [14].
- Estender o uso da técnica CLL para outras arquiteturas para verificar se os resultados observados nesta dissertação são mantidos.
- Investigar a eficiência dos modelos em corpora de discurso de ódio relacionados a outras temáticas, como sexismo, racismo, xenofobia, religião e outras áreas temáticas relevantes.

- Desenvolver e testar novas estratégias baseadas na técnica CLL para otimizar os resultados alcançados neste trabalho, como por exemplo: aplicar CLL em subgrupos de idiomas para verificar o desempenho do modelo ou utilizar *ensemble* de modelos monolíngues/multilíngues aplicando CLL para verificar a eficiência do modelo.

- Explorar outras arquiteturas de modelos diferentes dos *Transformers* para determinar se oferecem resultados superiores aos encontrados nesta pesquisa, por exemplo: modelos CNN [50] ou o modelo HAN [106].

- Avaliar a aplicabilidade da metodologia adotada nesta pesquisa em outras áreas de classificação, como análise de emoção ou sentimento, *fake news*, entre outros.

Bibliografia

[1] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 141–153. Springer, 2018.

[2] MATT AHLGREN. Mais de 55 estatÍsticas, fatos e tendencias do twitter para 2023. https://www.websiterating.com/pt/research/twitter-statistics/. Acesso em novembro de 2023.

[3] Jay Alammar. The illustrated transformer. https://jalammar.github.io/illustrated-transformer, 2018. Acesso em novembro de 2023.

[4] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2009.

[5] Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. BUFFET: benchmarking large language models for few-shot cross-lingual transfer. *CoRR*, abs/2305.14857, 2023.

[6] April Asgapo. How many tweets are sent out per day? https://truelist.co/blog/how-many-tweets-per-day/. Acesso em novembro de 2023.

[7] Gabriel Assis, Annie Amorim, Jonnatahn Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In Pablo

Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 301–311, Santiago de Compostela, Galicia/Spain, March 2024. Association for Computational Linguistics.

[8] Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. Cross-lingual transfer learning for hate speech detection. In Bharathi Raja Chakravarthi, John P. McCrae, Manel Zarrouk, Rajeev K. Bali, and Paul Buitelaar, editors, *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, LT-EDI@EACL 2021, Online, April 19, 2021*, pages 15–25. Association for Computational Linguistics, 2021.

[9] Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. Overview of the EVALITA 2018 hate speech detection task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[11] Neil Vicente Cabasag, Vicente Raphael Chan, Sean Christian Lim, Mark Edward

Gonzales, and Charibeth Cheng. Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing. *Philippine Computing Journal, XIV No*, 1, 2019.

[12] Gobinda G. Chowdhury. Natural language processing. *Annu. Rev. Inf. Sci. Technol.*, 37(1):51–89, 2003.

[13] Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. Multi-task learning in deep neural networks at EVALITA 2018. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[14] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[15] Thomas Hikaru Clark, Costanza Conforti, Fangyu Liu, Zaiqiao Meng, Ehsan Shareghi, and Nigel Collier. Integrating transformers and knowledge graphs for twitter stance detection. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT 2021, Online, November 11, 2021*, pages 304–312. Association for Computational Linguistics, 2021.

[16] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2426–2430. ISCA, 2021.

[17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Vese-

lin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[18] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Techn.*, 20(2):10:1–10:22, 2020.

[19] Datareportal. Digital 2023: Global digital overview. https://datareportal.com/social-media-users, 2023. Acesso em novembro de 2023.

[20] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017.

[21] Aillkeen de Oliveira., Cláudio Baptista., Anderson Firmino., and Anselmo Cardoso de Paiva. Using multilingual approach in cross-lingual transfer learning to improve hate speech detection. In *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 374–384. INSTICC, SciTePress, 2023.

[22] Aillkeen Bezerra de Oliveira. Hate speech corpus. https://github.com/Aillkeen/hate_speech_corpus, 2024. Disponibilizado em junho de 2024.

[23] Aillkeen Bezerra De Oliveira, Claudio de Souza Baptista, Anderson Almeida Firmino, and Anselmo Cardoso De Paiva. A large language model approach to detect hate speech in political discourse using multiple language corpora. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, page 1461–1468, New York, NY, USA, 2024. Association for Computing Machinery.

[24] Flor Miriam Plaza del Arco, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Syst. Appl.*, 166:114120, 2021.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burs-

tein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[26] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics, 2019.

[27] Armend Duzha, Cristiano Casadei, Michael Tosi, and Fabio Celli. Hate versus politics: Detection of hate against policy makers in italian tweets. *CoRR*, abs/2107.05357, 2021.

[28] Jacob Eisenstein. *Natural Language Processing*. MIT Press, 2018.

[29] Tiziano Fagni, Leonardo Nizzoli, Marinella Petrocchi, and Maurizio Tesconi. Six things I hate about you (in italian) and six classification strategies to more and more effectively find them. In Pierpaolo Degano and Roberto Zunino, editors, *Proceedings of the Third Italian Conference on Cyber Security, Pisa, Italy, February 13-15, 2019*, volume 2315 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[30] Sandra Regina Feiteiro, Soorro Cardoso Silva, and Sandra Regina Feiteiro. Estudo da variação lexical na amazônia paraense: um olhar sobre o atlas linguístico do brasil. *Signum: Estudos da Linguagem*, 18(1):157–181, jan. 2015.

[31] Anderson Almeida FIRMINO et al. Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas. 2022.

[32] Camilla Fonseca. Word embedding: Fazendo o computador entender o significado das palavras. https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057, 2021. Acessado em maio de 2024.

[33] Paula Fortuna, Ilaria Bonavita, and Sérgio Nunes. Merging datasets for hate speech classification in italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[34] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, 2018.

[35] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, 2018.

[36] Simona Frenda, Bilal Ghanem, Manuel Montes-y-Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.*, 36(5):4743–4752, 2019.

[37] José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo Palacios, and Rafael Valencia-García. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.*, 114:506–518, 2021.

[38] Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49, February 2010.

[39] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[40] Russell D. Gray and Fiona M. Jordan. Language trees support the express-train sequence of austronesian expansion. *Nature*, 405(6790):1052–1055, June 2001.

[41] Lara Grimminger and Roman Klinger. Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In Orphée De Clercq, Alexandra Balahur, João Sedoc, Valentin Barrière, Shabnam Tafreshi, Sven Buechel, and Véronique Hoste, editors, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EACL 2021, Online, April 19, 2021*, pages 171–180. Association for Computational Linguistics, 2021.

[42] Bruno Ferrari Guide. *Detecção automática de discurso de ódio punitivista em redes sociais*. PhD thesis, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2022.

[43] Nathan Hartmann, Erick Rocha Fonseca, Christopher Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In Gustavo Henrique Paetzold and Vládia Pinheiro, editors, *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, STIL 2017, Uberlândia, Brazil, October 2-5, 2017*, pages 122–131. Sociedade Brasileira de Computação, 2017.

[44] Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. *CoRR*, abs/2308.05596, 2023.

[45] Wilbert Heeringa. Measuring dialect pronunciation differences using levenshtein distance, ph.d. Master's thesis, University of Groningen, 2004.

[46] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. The problem of identifying misogynist language on twitter (and other online social spaces). In Wolfgang Nejdl, Wendy Hall, Paolo Parigi, and Steffen Staab, editors, *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 333–335. ACM, 2016.

[47] Aiqi Jiang and Arkaitz Zubiaga. Cross-lingual capsule network for hate speech detection in social media. *CoRR*, abs/2108.03089, 2021.

[48] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[49] Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2021.

[50] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[51] Joseph B. Kruskal and Mark Liberman. The symmetric time-warping problem: from continuous to discrete. In David Sankoff and Joseph B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter 4. CSLI Publications, Stanford, CA 94305, 1999.

[52] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[53] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. 10:707–710, 1966.

[54] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[55] Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. *CoRR*, abs/1712.09913, 2017.

[56] Yingjie Li, Chenye Zhao, and Cornelia Caragea. Improving stance detection with multi-dataset learning and knowledge distillation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6332–6345. Association for Computational Linguistics, 2021.

[57] Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Victor Carneiro, and Fidel Cacheda. Early detection of cyberbullying on social media networks. *Future Gener. Comput. Syst.*, 118:219–229, 2021.

[58] Krishanu Maity, Shaubhik Bhattacharya, Sriparna Saha, and Manjeevan Seera. A deep learning framework for the detection of malay hate speech. *IEEE Access*, 11:79542–79552, 2023.

[59] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In Paolo Boldi, Brooke Foucault Welles, Katharina Kinder-Kurlanda, Christo Wilson, Isabella Peters, and Wagner Meira Jr., editors, *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*, pages 173–182. ACM, 2019.

[60] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[61] Tom Mitchell and Maria Hill. *Machine Learning*. McGraw-Hill Science and Engineering and Math, 1st edition, 1997.

[62] Miljana Mladenovic, Vera Osmjanski, and Stasa Vujicic Stankovic. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Comput. Surv.*, 54(1):1:1–1:42, 2022.

[63] Maria Carolina Monard and José Augusto Baranauskas. Conceitos sobre aprendizado de máquina. pages 89–114, 2003.

[64] Mainack Mondal, Leandro Araújo Silva, Denzil Correa, and Fabrício Benevenuto. Characterizing usage of explicit hate expressions in social media. *New Rev. Hypermedia Multim.*, 24(2):110–130, 2018.

[65] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896, 2022.

[66] Yuvaraj Natarajan, Victor Chang, Balasubramanian Gobinathan, Arulprakash Pinagapani, Srihari Kannan, Gaurav Dhiman, and Arsath Raja Rajan. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Comput. Electr. Eng.*, 92:107186, 2021.

[67] Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 907–914. Association for Computational Linguistics, 2021.

[68] Magdalena Obermaier and Desirée Schmuck. Youths as targets: factors of online hate speech victimization among adolescents and young adults. *Journal of Computer-Mediated Communication*, 27(4), 07 2022.

[69] ONU. Pacto internacional sobre direitos civis e políticos. https://oas.org/dil/port/1966 Pacto Internacional sobre Direitos Civis e Políticos.pdf, 1996. Acesso em novembro de 2023.

[70] OpenAI. Openai site. https://openai.com, 2022. Acessado em 9 de fevereiro de 2024.

[71] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544, 2021.

[72] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daele-

mans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[73] Matús Pikuliak, Marián Simko, and Mária Bieliková. Cross-lingual learning for text processing: A survey. *Expert Syst. Appl.*, 165:113765, 2021.

[74] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523, 2021.

[75] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and et al. Improving language understanding by generative pre-training. 2018.

[76] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[77] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[78] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[79] Alpa Reshamwala, Dhirendra Mishra, and Prajakta Pawar. Review on natural language processing. *IRACST – Engineering Science and Technology: An International Journal (ESTIJ)*, 3:113–116, 02 2013.

[80] Maria Isabel Rizzotto and André Luiz Saraiva. Discurso de ódio nas redes sociais digitais: tipos e formas de intolerância na página oficial de jair bolsonaro no facebook. *Revista Eletrônica Científica da UERGS*, 6(1):11–23, 2020.

[81] Stuart Russell and Peter Norvig. *Inteligência Artificial*. Campus, 3rd edition, 2013.

[82] Alexander Savelyev and Martine Robbeets. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53, 02 2020.

[83] David Sayce. The number of tweets per day in 2022. https://www.dsayce.com/social-media/tweets-day/. Acesso em novembro de 2023.

[84] Andrea Schioppa, Xavier Garcia, and Orhan Firat. Cross-lingual supervision improves large language models pre-training. *CoRR*, abs/2305.11778, 2023.

[85] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics, 2017.

[86] Stefan Schweter. Italian bert and electra models. *Zenodo, nov*, 8, 2020.

[87] M. Serva and F. Petroni. Indo-european languages tree by levenshtein distance. *Europhysics Letters*, 81(6):68005, feb 2008.

[88] Rosane Leal da Silva, Andressa Nichel, Anna Clara Lehmann Martins, and Carlise Kolbe Borchardt. Discursos de ódio em redes sociais: jurisprudência brasileira. *Revista direito GV*, 7:445–468, 2011.

[89] De Smedt, Tom, and Sylvia Jaki. The polly corpus: Online political debate in germany. In *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*, 2018.

[90] Claver P. Soto, Gustavo M. S. Nunes, José Gabriel R. C. Gomes, and Nadia Nedjah. Application-specific word embeddings for hate and offensive language detection. *Multim. Tools Appl.*, 81(19):27111–27136, 2022.

[91] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. Bertimbau: Pretrained BERT models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020,*

Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer, 2020.

[92] Lukas Stappen, Fabian Brunn, and Björn W. Schuller. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850, 2020.

[93] Cagri Toraman, Furkan Sahinuç, and Eyup Halit Yilmaz. Large-scale hate speech detection with cross-domain transfer. *CoRR*, abs/2203.01111, 2022.

[94] Dennis Thomas Ulmer, Christian Hardmeier, and Jes Frellsen. deep-significance - easy and meaningful statistical significance testing in the age of neural networks. April 2022.

[95] Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Mach. Learn.*, 109(2):373–440, 2020.

[96] Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[97] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France, June 2022. European Language Resources Association.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N.

Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[99] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In Alessandro Armando, Roberto Baldoni, and Riccardo Focardi, editors, *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, volume 1816 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS.org, 2017.

[100] José Juan Vázquez, Alexia C. Suarez, Alberto E. Berríos, and Sonia Panadero. Intersecting vulnerabilities, intersectional discrimination, and stigmatization among people living homeless in nicaragua. *Social Science Quarterly*, 102(1):618–627, 2021.

[101] Shirui Wang, Wenan Zhou, and Chao Jiang. A survey of word embeddings based on deep learning. *Computing*, 102, 03 2020.

[102] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics, 2016.

[103] Elizabeth Whittaker and Robin M Kowalski. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29, 2015.

[104] Würzburg. German hate speech corpus. https://github.com/cophi-wue/German_HateSpeech_Corpus, 2021. Acesso em novembro de 2023.

[105] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. Multimodal hate speech detection via cross-domain knowledge transfer. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin,

Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4505–4514. ACM, 2022.

[106] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[107] Hailemariam Mehari Yohannes and Toshiyuki Amagasa. Named-entity recognition for a low-resource language using pre-trained language model. In Jiman Hong, Miroslav Bures, Juw Won Park, and Tomás Cerný, editors, *SAC '22: The 37th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, April 25 - 29, 2022*, pages 837–844. ACM, 2022.

[108] Ethan Zhang and Yi Zhang. *F-Measure*, pages 1147–1147. Springer US, Boston, MA, 2009.

[109] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

[110] Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In Ceren Budak, Meeyoung Cha, and Daniele Quercia, editors, *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 1435–1439. AAAI Press, 2022.