

Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

LLMeFT-Transformer: Uma abordagem de TDL para detecção de Discurso de Ódio e Toxicidade em Redes Sociais

Ghivvago Damas Saraiva

Número de Ordem PPGCC: M001

Teresina-PI, Dezembro de 2024

Ghivvago Damas Saraiva

LLMeFT-Transformer: Uma abordagem de TDL para detecção de Discurso de Ódio e Toxicidade em Redes Sociais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: COMPUTAÇÃO INTELIGENTE), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Raimundo Santos Moura

Coorientador: Prof. Dr. Rafael Torres Anchieta

Teresina-PI

Dezembro de 2024

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

S243l Saraiva, Ghivvago Damas.
LLMeFT-Transformer: uma abordagem de TDL para detecção de discurso de ódio e toxidade em redes sociais / Ghivvago Damas Saraiva. -- 2024.
78 f.

Dissertação (Mestrado) - Universidade Federal do Piauí. Centro de Ciências da Natureza. Programa de Pós-Graduação em Ciências da Computação, Teresina, 2024.
"Orientador: Prof. Dr. Raimundo Santos Moura. Coorientador: Prof. Dr. Rafael Torres Anchieta".

1. Dispositivos Inteligentes. 2. Processamento de Linguagem Natural. 3. Redes sociais. 4. Toxidade. 5. Tabular Deep Learning. I. Moura, Raimundo Santos. II. Anchieta, Rafael Torres. II. Titulo.

CDD 005.

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Ghivvago Damas Saraiva

LLMeFT-Transformer: Uma abordagem de TDL para detecção de Discurso de Ódio e Toxicidade em Redes Sociais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: COMPUTAÇÃO INTELIGENTE), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 06 de dezembro de 2024:

Documento assinado digitalmente
gov.br
RAIMUNDO SANTOS MOURA
Data: 06/02/2025 16:32:53-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Raimundo Santos Moura
Orientador

Documento assinado digitalmente
gov.br
RAFAEL TORRES ANCHIETA
Data: 06/02/2025 18:12:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rafael Torres Anchieta
Coorientador

Documento assinado digitalmente
gov.br
RODRIGO DE MELO SOUZA VERAS
Data: 12/02/2025 08:52:39-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo de Melo Souza Veras
Examinador interno à instituição

Documento assinado digitalmente
gov.br
VINICIUS PONTE MACHADO
Data: 08/02/2025 09:51:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Vinícius Ponte Machado
Examinador interno à instituição

Documento assinado digitalmente
gov.br
JONICE DE OLIVEIRA SAMPAIO
Data: 12/02/2025 16:41:29-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Jonice de Oliveira Sampaio
Examinador externo à instituição

Teresina-PI
Dezembro de 2024

Dedico esta Dissertação à comunidade acadêmico-científica, à comunidade LGBTQIAPN+ e a todas as pessoas neurodivergentes.

Expresso minha profunda gratidão à comunidade científica nos campos da Computação, Linguística Computacional e Inteligência Artificial, bem como às comunidades acadêmicas da UFPI, UFMA e UFRN, pelo apoio e conhecimento que possibilitaram este trabalho. Dedico esta pesquisa à comunidade LGBTQIAPN+, como um esforço para combater o ódio e a toxicidade nas redes sociais, visando um mundo mais inclusivo e justo. Dedico também estas páginas às pessoas neurodivergentes, especialmente autistas e TDAHs, como prova de que nosso sucesso acadêmico e profissional é alcançável, superando desafios únicos com resiliência e criatividade.

Que esta pesquisa sirva como um lembrete e uma inspiração para que todos, mesmo diante das adversidades, possam trilhar um caminho de conquistas e promover mudanças positivas na sociedade através da ciência e da tecnologia.

Agradecimentos

Expressar gratidão requer profundidade, reflexão e coerência com os valores pessoais. Reconheço que minha jornada de vida e pesquisa foi desafiadora, mas sou grato por cada obstáculo superado.

A complexidade e os desafios dos meus ambientes sociais me motivaram profundamente e forneceram um objeto de estudo pessoal que impulsionou esta pesquisa.

Enfrentar a discriminação e a intolerância desde cedo me ensinou a questionar, ser independente e lutar por meus valores. Como homem negro, gay, agnóstico-ateu e neurodivergente, busquei a liberdade e a tolerância, transformando minhas dificuldades em motivação para o sucesso. Sou grato por essa jornada, que me ensinou a importância da aceitação, do amor-próprio e da diversidade.

Agradeço à Profa. Dra. Márjory Cristiany da Costa Abreu pela orientação inicial e aos colegas da UFRN, especialmente para Francimaria Rayanne dos S. Nascimento, Julliana Caroline G. de A. S. Marques, Roberta Freitas, Camila de Araújo, Rafael Jullian O. do Nascimento e Heloíza F. da S. Santiago, pedindo desculpas por qualquer transtorno. Sou grato aos colegas da UFPI, em especial Zilma Félix G. Araújo e Fábio Cordeiro, pelo apoio. Agradeço ao Prof. Francisco Ricarte (IFPI) pela parceria, aos professores do PPGCC-UFPI e ao Prof. Dr. Anselmo (UFMA) pelas contribuições.

Agradeço profundamente aos meus orientadores, Prof. Dr. Raimundo Santos Moura e Prof. Dr. Rafael Torres Anchiêta, pela paciência, confiança e orientação. Sua influência em minha formação acadêmica e pessoal é imensurável.

Agradeço à CAPES pelo apoio financeiro, fundamental para esta pesquisa. Ressalto a importância do fomento à pesquisa para o avanço científico e a formação de profissionais no Brasil.

Esta pesquisa é um reflexo da minha jornada e da força que encontrei em mim, uma homenagem à busca por conhecimento e uma prova de que as adversidades podem ser transformadoras. Espero que minha história inspire outros a encontrarem força em suas próprias lutas.

"O verdadeiro progresso da humanidade
se dá pela união do conhecimento científico
com a habilidade de tolerar e aprender com as diferenças."
(Autor Desconhecido)

Resumo

Nos últimos anos, a incidência de ódio e toxicidade nas redes sociais tem crescido exponencialmente, tornando-se um fenômeno cada vez mais complexo que permeia os ambientes virtuais. Esse fenômeno, agora intrínseco às interações sociais online, tem sido alvo de análise e detecção por parte das comunidades científicas, expandindo-se para além das disciplinas tradicionais como Direito, Sociologia, Psicologia e Política, adentrando também nas áreas de Processamento de Linguagem Natural, Aprendizado de Máquina e Linguística. Nesse contexto, este estudo propõe uma abordagem para identificar comentários tóxicos ou que contenham discurso de ódio nas redes sociais, utilizando uma combinação de *Tabular Deep Learning* e Processamento de Linguagem Natural. O objetivo é ampliar um modelo preditivo baseado em Transformer, o FT-Transformer, para detectar com precisão satisfatória se um comentário textual, como um tweet, possui conteúdo tóxico ou de ódio. Uma parte crucial do trabalho é a geração e processamento das características textuais, geradas a partir de modelos pré-treinados como Sentence-Transformer e BERT (tais como BERTimbau e Albertina PT-BR) e outros modelos de *embeddings*, utilizando-os como poderosos codificadores de texto. Assim, um aspecto importante da pesquisa envolve o uso de modelos de *embeddings* modernos como incorporadores e modelos de linguagem avaliando seu desempenho com o FT-Transformer, um modelo tabular baseado em transformador. O cenário experimental utiliza a versão binária do conjunto de dados ToLD-Br e a nossa abordagem alcançou uma precisão de 76% e uma pontuação F1-score de 75% usando o modelo de *embedding* da OpenAI text-embedding-3-large. Os resultados obtidos indicam o desempenho singular da abordagem, que pode ser aprimorada com a incorporação de novas características, como o uso de técnicas como o RAG (*Retrieval-Augmented Generation*) que podem melhorar sua capacidade de lidar com as nuances da linguagem e a manutenção das características mais significativas em um objeto textual. Portanto, acreditamos haver oportunidades para aprimorar e expandir a capacidade do modelo em classificar comentários com ódio e toxicidade, destacando ainda mais a abordagem como uma alternativa inovadora em relação às abordagens tradicionais.

Palavras-chave: Ódio, Toxicidade, Processamento de Linguagem Natural, Tabular Deep Learning, FT-Transformer.

Abstract

In recent years, the prevalence of hate speech and toxicity on social media has grown exponentially, becoming an increasingly complex phenomenon that pervades virtual environments. This study proposes a novel approach to identify toxic or hate speech comments on social media using a combination of tabular deep learning and natural language processing. The objective is to enhance a transformer-based predictive model, FT-Transformer, for accurately detecting toxic or hateful content in textual comments. A key aspect of this work is the generation and processing of textual features using pre-trained models such as Sentence-Transformer, BERT (e.g., BERTimbau and Albertina PT-BR) and other modern embedding models. These models serve as powerful text encoders, providing rich representations of textual data. The experimental setup employs the binary version of the ToLD-Br dataset. Our approach achieved an accuracy of 76%, and a macro F1 score of 75% using the OpenAI *text-embedding-3-large* embedding model. The results demonstrate the effectiveness of our proposed approach, which can be further improved by incorporating additional features and techniques such as Retrieval-Augmented Generation (RAG) to enhance its ability to handle linguistic nuances. This research highlights the potential of tabular deep learning and language models in combatting online hate speech.

Keywords: Hate, Toxicity, Natural Language Processing, Tabular Deep Learning, FT-Transformer.

Lista de ilustrações

Figura 1 – Linha do tempo com os principais estudos e avanços do PLN (fonte: o autor). 12
Figura 2 – Modelo de Camadas do Processamento de Linguagem Natural (Adaptado de Dale (2010), Schulz (2021), Caseli, Nunes e Pagano (2023)). . . 13
Figura 3 – Exemplo de Método Ensemble: Gradiente Boost Classifier (fonte: o autor). 19
Figura 4 – Exemplo de Redes Neurais Artificais Rasa e Profunda (fonte: Packt - Why deep learning?). 20
Figura 5 – Arquitetura original do modelo Transformer proposto por Vaswani et al. (2017). 22
Figura 6 – Cronografia das Representações textuais distribuídas e simbólicas. (fonte: Liu e Sun (2023). 27
Figura 7 – Pipeline da abordagem LLMeFT-Transformer (fonte: o autor). 41
Figura 8 – Feature Tokenizer (FT) porposto por Gorishniy et al. (2021). 44
Figura 9 – Arquitetura PyTorch Frame adaptado de Hu et al. (2024). 45
Figura 10 – Matriz de confusão para as classes tóxica e não tóxica. 52

Lista de tabelas

Tabela 1 – Trabalhos relacionados à Detecção Discurso de Ódio e Toxicidade. Legenda: F_1 : F-score, A: Acurácia e P: Precisão. 37
Tabela 2 – Discurso Tóxico e de Ódio: corpora disponíveis em Português-BR. . . . 39
Tabela 3 – Características do ToLD-Br - Versão Binária 42
Tabela 4 – Características do ToLD-Br - Versão Multiclasse 42
Tabela 5 – Lista de Modelos de Embeddings. 47
Tabela 6 – Parâmetros do Modelo 50
Tabela 7 – Resultados com diferentes Text Embedders. 51
Tabela 8 – Detecção de Ódio e Toxicidade: comparativo entre abordagens com o ToLD-Br. 51

Lista de abreviaturas e siglas

ALBERT	A Lite BERT
AM	Aprendizagem de Máquina
AP	Aprendizagem Profunda
API	Application Programming Interface
AutoInt	Automated Feature Interaction
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
ByT5	Bilingual Tokenized Transformers
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CETIC.br	Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação
CBOW	Continous Bag-of-Words
CLS	Classification Token
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CT Scan	Computed Tomography Scan
DL	Deep Learning
ELMo	Embeddings from Language Models
FT	Feature Tokenizer
FT-Transformer	Feature Tokenizer + Transformer
FTT	FT-Transformer (abreviação específica)
GB Classifier	Gradient Boost Classifier
GBDTs	Gradient-Boosted Decision Trees
GBMs	Gradient-Boosted Models

GloVe	Global Vectors for Word Representation
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
HateBr	Offensive Language and Hate Speech Dataset in Brazilian Portuguese
HGN	Heterogeneous Graph Network
HLPHSD	Hierarchically-Labeled Portuguese Hate Speech Dataset
IA	Inteligência Artificial
LightGBM	Light Gradient Boosting Machine
LLMs	Large Language Models
LSTM	Long Short-Term Memory
ML	Machine Learning
MLM	Masked Language Modeling
MLP	Multi-Layer Perceptron
MLPT	Modelo de Linguagem Pré-treinado
NB	Naive-Bayes
NILC	Núcleo Interinstitucional de Linguística Computacional
NLP	Natural Language Processing
NODE	Neural Ordinary Differential Equations
NSP	Next Sentence Prediction
OffComBR	Offensive Comments in the Brazilian Web Dataset
OLID-BR	Offensive Language Identification Dataset for Brazilian Portuguese
PMI	Pointwise Mutual Information
PT-BR	Português do Brasil
PT-PT	Português de Portugal
PLN	Processamento de Linguagem Natural
RAG	Retrieval-Augmented Generation

ReLU	Rectified Linear Unit
RL	Regressão Logística
RMSE	Root Mean Square Error
RNNs	Recurrent Neural Networks
RNNs	Redes Neurais Recorrentes
SAINT	Selective Attention Integrative Network
SBC	Sociedade Brasileira de Computação
SBERT	Sentence-BERT
SEP	Separation Token
Seq2Seq	Sequence-to-Sequence
STS	Semantic Textual Similarity
SVM	Support Vector Machine
TabNet	Tabular Neural Network
TDL	Tabular Deep Learning
TE	Text Embedder
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TIC	Tecnologias de Informação e Comunicação
ToLD-Br	Toxic Language Dataset for Brazilian Portuguese
Tupy	Brazilian Hate Speech Dataset (TuPy)
TupyE-Dataset	Portuguese Hate Speech Expanded Dataset
TupyE	Portuguese Hate Speech Expanded Dataset (abreviação específica)
Wang2Vec	Word and N-gram embedding model
WE	Word Embeddings
X	X (anteriormente Twitter)
XGBoost	Extreme Gradient Boosting

Sumário

1 INTRODUÇÃO 1
1.1 Desafios, Soluções e Motivação 6
1.2 Objetivos 8
1.3 Contribuições 8
1.4 Organização do texto 9

2 REFERENCIAL TEÓRICO 11
2.1 Processamento de Linguagem Natural 11
2.2 Aprendizagem de Máquina 14
2.2.1 Aprendizagem Clássica 15
2.2.1.1 Gradient-Boosted Decision Trees 17
2.2.1.1.1 Gradient Boost Classifier 17
2.2.2 Aprendizagem Profunda ou *Deep Learning* 20
2.2.2.1 Transformer 21
2.2.2.2 BERT e outros LLMs 22
2.2.2.3 Aprendizagem Profunda Tabular 24
2.3 Representação de Texto 25

3 TRABALHOS RELACIONADOS 31
3.1 Ódio e Toxicidade — abordagens para detecção 31
3.1.1 Datasets 38

4 LLMEFT-TRANSFORMER 41
4.1 Base de Dados 41
4.2 FT-Transformer 43
4.2.1 Feature Tokenizer 43
4.2.2 Transformer 43
4.2.3 LLMeFT-Transformer 44
4.2.4 *Embeddings* Textuais 45
4.2.4.1 *Text Embedder*: A Interface entre Texto e o Modelo 46
4.3 Considerações Finais 47

5 EXPERIMENTOS E RESULTADOS 49
5.1 Experimentos 49
5.2 Resultados 50
5.3 Considerações Finais 53

6 CONCLUSÕES, LIMITAÇÕES E TRABALHOS FUTUROS 55
6.1 Conclusões 55
6.2 Limitações 56
6.3 Trabalhos Futuros 56

REFERÊNCIAS 59

1 Introdução

As mídias sociais são uma ferramenta poderosa, com o potencial de conectar pessoas e compartilhar informações. Através das constantes transformações promovidas com uso da tecnologia e por meio da Internet, as redes sociais criaram não só espaço para a disseminação de informação e conteúdo, mas também espaço para validação da opinião. Essa democratização no acesso e na capacidade de geração de conteúdo é de fato a maior causa das reviravoltas em toda a cultura e comportamento que indivíduos e comunidades têm vivenciado nas últimas décadas (SILVA et al., 2016). Destas reviravoltas, a disseminação de desinformação, a violência e discurso de ódio são algumas das mais preocupantes.

No Brasil, as mídias sociais são extremamente populares, com mais de 165 milhões de usuários ativos. Alguns dados publicados em maio de 2023 na pesquisa TIC Domicílios 2022[1] pela CETIC.br apontaram que mais de 139 milhões de indivíduos engajaram em alguma plataforma e geraram algum tipo de conteúdo, sua maioria em aplicativos de mensagens e redes sociais. Isso significa que muitas pessoas estão expostas ao conteúdo compartilhado ou produzindo algum tipo de conteúdo nessas plataformas. Os dados publicados também relatam o salto na proporção de conteúdos publicados de autoria dos usuários (textos, imagens ou vídeos), de 31% (2021) para 43% (2022). Nessa pesquisa, do total de usuários que utilizaram a Internet para comunicação ou geração de conteúdo por meio principalmente de redes sociais supera os 80% dos entrevistados. Diante de tais números, o coordenador da pesquisa TIC Domicílios, Fábio Storino, ponderou que em 2022 houve uma explosão de postagens curtas e rápidas, facilitadas por aplicativos de redes sociais e mensagens instantâneas e estimuladas pela retomada das atividades presenciais (CETIC.BR, 2023).

Um problema presente nas redes sociais nos últimos anos é o discurso de ódio, este é, de acordo com dados de um levantamento publicado pela Central Nacional de Denúncias da Safernet[2], o mais preocupante e urgente de ser tratado. De acordo com esse mesmo levantamento, o número de crimes de ódio na Internet triplicou nos últimos seis anos, apenas no primeiro semestre de 2022 foram registrados cerca de 23.947 denúncias, 67.5% a mais em relação ao primeiro semestre do ano de 2021.

Dados publicados recentemente pela Secretaria da Segurança Pública do Piauí (SSP-PI) no 2° Boletim de Dados de Violência Contra à Pessoa LGBTQIAPN+ (PIAUí, 2024), revelam um cenário preocupante para a comunidade no estado. Só em 2023, foram denunciados 68 crimes de discriminação e registradas 10 mortes de pessoas LGBTQIAPN+.

[1] Acesso às tecnologias de informação e comunicação no domicílio (TIC Domicílios — 2022 Domicílios)
[2] Crimes de ódio na internet tiveram aumento de quase 70% no primeiro semestre - Rádio Senado

A maioria das vítimas era jovem, entre 20 e 39 anos, e negra, evidenciando a intersecção entre violência de gênero, racismo e LGBTfobia. Os crimes de natureza financeira foram os mais frequentes, representando mais de 40% dos casos. O boletim também destaca a violência letal, com 40% das mortes ocorrendo na faixa etária de 40 a 49 anos, a maioria das vítimas sendo negras e os crimes cometidos principalmente com arma de fogo em residências, estabelecimentos comerciais e vias públicas. Outro dado também importante para esta pesquisa é de que 7,35% dos crimes foram consumados no ambiente virtual e a homofobia foi a motivação preponderante em 73% dos casos de crimes motivados por preconceito [3].

Outras estatísticas detalhadas sobre ódio no Brasil informam que mais de 12 mil casos de crimes de ódio foram notificados em 2019, apontando um aumento de 1.95% face ao ano de 2018 (BUARQUE; CRETTON, 2021). Da totalidade de casos, 72,8% estavam relacionados a discriminação-preconceito racial, 14,04% a discriminação sexual, orientação-identidade, 10,65% identidade de gênero, 1,83% intolerância ou preconceito religioso e 0,67% tinham relação direta com a origem das pessoas, sua etnia.

Com base nesses dados, é de extrema importância para a tratativa dessa problemática que as plataformas de mídia social tomem medidas para combater a disseminação de desinformação, violência e discurso de ódio. Isso pode ser feito via uma combinação de medidas, como a verificação de fatos e a moderação de conteúdo mais rigorosa, educação dos usuários e a inoculação psicológica. Também é importante que os governos regulamentem as mídias sociais, de modo a proteger os usuários de danos. A implementação dessas medidas ajudaria a tornar as mídias sociais um espaço mais seguro e saudável para todos. Possibilitando que seus desafios, como a polarização de grupos, à legitimação do medo e à perda da coesão entre as comunidades possam ser enfrentados efetivamente.

A desinformação é um dos principais problemas das redes sociais, logo ao lado do discurso de ódio. Ela pode ser deliberada ou não intencional, mas sempre tem o potencial de causar danos. A desinformação pode levar a decisões desinformadas, à polarização das opiniões e a legitimação do discurso de ódio e da violência. Nesse sentido, a desinformação é um dos mecanismos, ou combustíveis, que alimentam muitos dos comportamentos agressivos nas comunidades virtuais. Gabarron, Oyeyemi e Wynn (2021) relataram que diferentes tipos de informações podem se espalhar fácil e vertiginosamente através das redes sociais, incluindo informações falsas, imprecisas e também fatos. O discurso de ódio não difere, há diversos fatores que influenciam direta e indiretamente na propagação de formas agressivas de discurso, dentre eles estão os fatores culturais, a facilidade na disseminação de informação, o anonimato, e a criação e utilização de robôs. Mesmo utilizando-se de ações de intervenção como reportar ou marcar conteúdo suspeito e campanhas de inoculação psicológica de conteúdos para aumentar a capacidade de resistência ao ódio em escala, o

[3] SSP lança 2º Boletim de Dados Sobre Violência Contra a Pessoa LGBTQIAPN+ no Piauí

ódio e outras formas de discurso tóxico tornam-se praticamente impossíveis de não serem disseminados nos meios digitais de comunicação e informação, fazendo das redes sociais um terreno fértil e com insumos ilimitados para esse tipo de comportamento e conteúdo Roozenbeek et al. (2022), Bastani e Bahrami (2020), Tingley e Wagner (2017), Lanius, Weber e MacKenzie (2021).

O discurso de ódio é uma variante ou forma de *harmful speech* (também conhecido como discurso ofensivo, discurso tóxico ou toxicidade online), assim como o sexting, o cyberbullying, o terrorismo, a misoginia e a xenofobia que geralmente dispara seus ataques a grupos específicos da sociedade (BANKO; MACKEEN; RAY, 2020; KAPANTAI et al., 2021). A toxicidade, de maneira geral, envolve as interações negativas e prejudiciais que contribuem para uma atmosfera online hostil, impedindo o diálogo construtivo e prejudicando a saúde mental dos utilizadores. Neste contexto, a linguagem abusiva empregada, os ataques direcionados e o *teasing/bullying* online podem ter graves impactos psicológicos e emocionais nos usuários (ALTMAN, 1993; SIEGEL; BADAAN, 2020; ALVAREZ; WINTER, 2018; SOROKOWSKI et al., 2020; NIELSEN, 2002; SAHA; CHANDRASEKHARAN; CHOUDHURY, 2019). As vítimas de violência online muitas vezes sofrem de ansiedade, depressão e até transtorno de estresse pós-traumático. Em casos extremos, pode levar a vítima à automutilação, ao suicídio ou a serem vítimas de crimes violentos fomentados pelo discurso de ódio e da desinformação (GELBER, 2019; POLETTO et al., 2020; TALBOT, 2012; MüLLER; SCHWARZ, 2020).

O problema do ódio, assim como o da toxicidade, mesmo sem esses possuírem uma definição universal, pode também ser analisado em torno do grupo de apoio ou suporte (grupo de ódio), trazendo para essa discussão o efeito conhecido como câmara de eco ou *echo chamber* (VICARIO et al., 2016), que se refere ao agrupamento de indivíduos em uma rede na qual postam notícias e informações seletivas, compartilham ideias compatíveis e reforçam seus pontos de vista. O mesmo fato foi exposto e explicado em 1921 por Sigmund Freud em seu trabalho chamado Psicologia de Grupo e a Análise do Ego, que diz:

> "um indivíduo em um grupo está sujeito a mudanças na atividade mental... sua submissão à emoção fica mais intensa, reduzindo sua capacidade mental e indo em direção a outros indivíduos do grupo deixando inibições humanas peculiares" (FREUD, 2014).

Em outras palavras, buscando aceitação e sentimento de pertencimento a este grupo. Müller (2019) também expôs que a própria história guarda momentos em que a falta da mútua compreensão cria um êxodo de indivíduos que buscam no coletivo pensamentos similares aos seus. Mais adiante, Nemer (2018) esclarece que a Internet provê e assegura espaços onde grupos com afinidades comuns, principalmente ideológicas, podem congregar,

promovendo também o compartilhamento de opiniões que não alcançariam uma visibilidade significativa como alcançam na Internet, ou seja, a Internet provê refúgio coletivo.

Geralmente, os esforços para rastrear e lidar com discursos de ódio ou prejudiciais (postagens em geral) antes de assumirem qualquer dimensão séria são muito difíceis devido à falta de todas as informações necessárias, textuais e não textuais, sendo a localidade, informações sobre o perfil, seguidores e audiência, retuítes, contexto externo e o perfil dos alvos exemplos pertinentes dessas informações. Tais informações podem ainda estar relacionadas a uma única ocorrência ou a uma série de ocorrências, e todas essas informações precisam ser analisadas e processadas antes de qualquer ação (ou diretriz). Nos fóruns de discussão e nas mídias sociais, os comentários dos usuários são o centro de toda a dinâmica do espaço e frequentemente o discurso disseminado emprega diferentes falas intolerantes para reforçar os pontos de vista do enunciador e da audiência direta, podendo encorajar inclusive atos de violência(CINELLI et al., 2021; OZ; NURUMOV, 2022; HANSEN, 2023).

Muitas pesquisas concentraram diversos esforços em analisar e determinar os limites e nuances das diversas formas de discurso ofensivo/nocivo, além de descrever esse fenômeno do uso ou comportamento hostil em ambientes online, ou comunidades virtuais. Outra preocupação era encontrar uma definição que pudesse ser universalmente aceita e que não ignorasse as nuances intrínsecas de cada problema. Com base nessas mesmas pesquisas, percebeu-se uma congruência entre os vários autores e pesquisadores acerca da identificação de discurso de ódio ou *hateful speech* (SILVA; FREITAS, 2022; WEITZEL et al., 2023; MIRANDA et al., 2023; TRAJANO, 2023; FIRMINO et al., 2022; CAPOZZI et al., 2018; SCHWEPPE; PERRY, 2022; NOBATA et al., 2016; SALEEM et al., 2017; SCHMIDT; WIEGAND, 2017; SAKSESI; NASRUN; SETIANINGSIH, 2018; SCHOFIELD; DAVIDSON, 2017; DAVIDSON; BHATTACHARYA; WEBER, 2019; SELLARS, 2017; BERKMAN KLEIN CENTER, 2016; SCHOFIELD; DAVIDSON, 2017; SCHMIDT; WIEGAND, 2017; OCAMPO et al., 2023). A partir deste ponto, pode-se definir e resumir descritivamente :

Definição. ***Discurso de Ódio*** *é toda e qualquer forma de discurso ou comentário ofensivo ou comunicação hostil que ataque grupos sociais ou indivíduos baseados em características específicas e até especialíssimas a esse grupo ou indivíduo, e aqui se enquadram a intolerância religiosa, o antissemitismo, a gordofobia, o sexismo e misoginia, a homofobia, a xenofobia, o racismo e até a apologia a ditaduras, terrorismo e ao nazifascismo.*

Outros trabalhos trouxeram uma conceituação para o que seria considerado discurso, comentário, texto ou postagem tóxica. De maneira quase que geral, há uma concordância em dizer que o discurso tóxico contém formas de linguagem não-aceitável (inadequada, ofensiva ou abusiva) que pode estar implícita ou explícita, podendo incluir insultos e ameaças direcionadas a alguém ou a um grupo (ZAMPIERI et al., 2019; ALSHAREF et al., 2022; SAIF et al., 2018; SRIVASTAVA; KHURANA, 2019; GUIMARÃES et al., 2020; NGUYEN;

NGUYEN; NGUYEN, 2021; ZHAO; ZHANG; HOPFGARTNER, 2021). Discurso tóxico ou comentários tóxicos podem também ser descritos como comportamentos negativos ou comentários que são grosseiros, desrespeitosos e que principalmente incentivam que alguém, geralmente a parte atacada, a deixar a discussão e ser assim silenciada (SCHMIDT; WIEGAND, 2017; CJADAMS JEFFREY SORENSEN, 2017).

É importante salientar que, na problemática do discurso tóxico e de ódio, o uso ou não de termos ou expressões consideradas parte da linguagem tóxica, ofensiva ou abusiva não pode ser considerado principal fator para descrever se um comentário contém alguma toxicidade ou ódio. Nesta divergência, deve-se atentar principalmente para as formas de discurso ofensivo que exploram os mais diversos estilos linguísticos presentes numa língua (FORTUNA; NUNES, 2018), incluindo o uso do humor ou formas implícitas e delicadas de se descrever (*subtle forms*), como a manipulação psicológica e o *gaslighting* [4] (OCAMPO et al., 2023; PAPCUNOVÁ et al., 2023; ANDERSON; BARNES, 2022; PÉREZ et al., 2023).

Para ilustrar um pouco desse bordeamento irregular, alguns exemplos da dificuldade de lidar com discurso de ódio implícito são apresentados. O primeiro exemplo é a fala do youtuber Bruno Monteiro Aiub, conhecido como Monark, durante o Flow Podcast no dia 07 de fevereiro de 2022, onde foi dito o seguinte: “Eu acho que tinha que ter um partido nazista reconhecido pela lei”, e acrescentou a essa ideia “Se o cara quiser ser um anti-judeu, eu acho que ele tinha direito de ser” (CANDAL LUDMILA, 2022). Neste primeiro exemplo o youtuber e podcaster utiliza em seu discurso a defesa do direito ao movimento nazista (1933-1945), esse que perseguiu, torturou e matou judeus (em maioria), negros, homossexuais, estrangeiros, testemunhas de jeová e outros grupos minoritários na busca pela pureza da raça do povo alemão. Sua fala fez apologia, uso de símbolos e a defesa ao direito a um partido político nazista no Brasil que desde muito antes da época de sua fala é considerado crime segundo a Lei 7.716/1989. Além disso, em 2003, durante o julgamento do caso Ellwanger [5] o Supremo Tribunal Federal (STF) definiu que a liberdade de expressão não abarca apologias ao regime de Adolf Hitler.

O segundo exemplo, foi o cancelamento e banimento de postagens contendo as *hashtags* #Oppenheimer e #Barberheimer no Japão (YAM, 2023; ECONOMIST, 2023; TILLY, 2023) durante os meses de estreia e exibição (julho e agosto de 2023) do filme Barbie da Warner Bros. Studios e de Oppenheimer, filme dirigido por Christopher Nolan produzido pela Universal Studios. O banimento das postagens e a suspensão de contas se deu principalmente pela repercussão da estreia do filme Oppenheimer em um período considerado triste e de extremo luto para o povo japonês, agosto, mês que nos dias 6 e 9 de

[4] Gaslighting é o termo usado para designar uma forma de abuso psicológico em que informações são manipuladas até que a vítima não consiga mais acreditar na própria percepção da realidade (FATO, 2022; ROGERS,).

[5] Processo HC/82424 - HABEAS CORPUS (caso Ellwanger)

agosto de 1945 o Japão sofreu o bombardeio de bombas nucleares (*Little Boy* e *Fat Man*) pelos Estados Unidos. O uso dos termos e *hashtags* contendo qualquer menção ao filme foi considerado extremamente desrespeitoso e ofensivo para com a dor do povo japonês.

O terceiro, não tão recente, foi o post textual de uma usuária do X, na época Twitter, sobre a Miss Piauí 2017, Monalysa Alcântara, onde dizia o seguinte: “Credooooo! A miss piaui tem cara de empregadinha, cara comum, não tem perfil de miss, não era pra ta ai. Sorry. #MissBrasil #MissBrasil2017”[6]. O post é um exemplo explícito de discurso tóxico e racismo estrutural relacionado aos padrões sociais de beleza. Os termos "comum"e "cara de empregadinha"refletem o racismo e a discriminação contra pessoas negras e nordestinas, as quais são frequentemente vistas como pertencentes a papéis subalternos na sociedade. Naquele mesmo ano, Monalysa foi coroada Miss Brasil 2017, sendo a terceira mulher negra a conquistar o título.

O último exemplo e mais dicotômico é sobre a *digital influencer* Thais Carla, que disparou em vídeo contra Gisele Bündchen o seguinte discurso em relação ao seu peso e corpo: “Gisele, que é uma tripa. É um papelzinho, ela vira assim, você não sabe o que é lado e o que é frente.”[7]. A *influencer*, modelo *plus size* e agora cantora tem como suas pautas a quebra dos padrões de beleza, contra a "ditadura da magreza"(BARDFORD, 2012), o respeito as pessoas gordas e sua luta contra a gordofobia dentro dos mais diversos nichos sociais. Thais Carla já foi e ainda é alvo de diversos ataques gordofóbicos de pessoas públicas e nas redes sociais de maneira geral. Seu vídeo é resgatado diversas vezes, principalmente quando a *influencer* se pronuncia de forma polêmica em relação a tamanhos de roupas, espaços de assentos em avião, largura de portas em hotéis e *Airbnbs*, levantando o questionamento em relação à romantização da obesidade e gordofobia reversa (ou magrofobia (ORTIZ, 2017), e também de outro fênomeno relacionado ao peso e aos padrões sociais de beleza, o *skinny shaming* (FABELLO, 2018).

1.1 Desafios, Soluções e Motivação

No mundo em rápida mudança, percebe-se que metodologias precisas de detecção de discurso de ódio são extremamente relevantes e cruciais no enfrentamento de todos os tipos de discurso e comportamento hostil em ambientes virtuais (KARAYIğIT; AKDAGLI; ACı, 2022; LUO; BOUAZIZI; OHTSUKI, 2021; MATHEW et al., 2021).

Considerando a possível sobreposição entre o discurso de ódio e a toxicidade online, mesmo que haja alguma distinção, a detecção automática desses fenômenos não pode ser considerada trivial e é geralmente bastante desafiadora (EVKOSKI et al., 2021; EVKOSKI et al., 2022; ARSHAD-AYAZ; NASEEM; HIZAOUI, 2022). Como exemplo desafiador

[6] O ódio contra negros nas redes sociais: denúncias de racismo crescem 36% no Brasil
[7] Tiktok - Corte do vídeo com fala problemática de Thais Carla postado por @vi_toriaxp.

pode-se mencionar novamente o caso do banimento das postagens promovendo o filme Oppenheimer no Japão. Neste exemplo o desafio está além da plataforma social, está na história do povo e nas tristes memórias do 6 e 9 de agosto de 1945. O discuro de ódio e a toxicidade têm efeitos amplos e profundos sobre os usuários, a sociedade e os governos. Uma abordagem adequada para resolver esses problemas requer uma estratégia multifacetada que envolva empresas de tecnologia, principalmente plataformas de mídia social, governos, sociedade civil e a comunidade científica para planejar e desenvolver ferramentas, diretrizes, campanhas e mecanismos para combater o discurso de ódio. O objetivo é imunizar os usuários e manter um ambiente online seguro e construtivo, ao mesmo tempo, em que salvaguarda a liberdade de expressão.

Um dos principais desafios dessa tarefa é o conjunto limitado de recursos textuais, bem como a variedade de tipos e estilos de linguagem usados no discurso de ódio (KARAYIğIT; AKDAGLI; ACı, 2022; LEV-ON, 2018). Esses tipos de discurso de ódio compartilham muitas características com outros problemas atuais, como desinformação, assédio, terrorismo, crimes de discriminação racial e agressão contra mulheres, pessoas LGBTQIAPN+, membros de minorias religiosas e outros grupos.

Os recursos textuais formam um conjunto limitado para esta tarefa e, dado que o conteúdo pode ou não ser considerado discurso de ódio para um grupo, com base principalmente nas definições acordadas por este grupo, bem como na sua história. O discurso de ódio também se expõe como parte da toxicidade online, que abrange um espectro mais amplo de comportamentos online prejudiciais, tendo como conceito gêmeo o Discurso Prejudicial. Ambos têm impactos sociais, psicológicos e emocionais significativos nos indivíduos e contribuem para um ambiente online hostil, que às vezes também pode afetar o ambiente externo, referindo-se ao conceito social de violência simbólica (BOURDIEU, 1979).

Dentre os diversos *Corpora* disponíveis para a tarefa de identificação de ódio e toxicidade em português, como *Hierarchically-Labeled-Portuguese-Hate-Speech-Dataset* (FORTUNA et al., 2019), ToLD-BR(LEITE et al., 2020), HateBR(VARGAS et al., 2022), OLID-BR (TRAJANO; BORDINI; VIEIRA, 2023) e TupPyE-Dataset(OLIVEIRA; REIS; EBECKEN, 2023), há discordâncias em aspectos que não se restringem apenas as definições básicas de ódio e toxicidade, mas também abrangem informações variadas de fontes diferentes, como a anotação, a concordância entre os anotadores e a revisão por especialistas ou indivíduos mais habilitados para identificar o fenômeno. Tais informações podem estar diretamente correlacionadas com a eficácia dos mecanismos de detecção automática, de tal modo que a informação inserida proveniente da etapa de anotação e revisão contribuem para inclusão ou exclusão de características, ou a simples desconsideração.

Motivado pelos desafios e exemplos mencionados acima, pela necessidade e desejo de garantir aos indivíduos espaços virtuais saudáveis para convivência, discussão e

desenvolvimento, este trabalho propõe uma investigação sobre Toxicidade e Discurso de Ódio irradiados nas redes sociais, especialmente. Tal investigação considera heurísticas linguísticas e computacionais (modelos de língua pré-treinados e modelos de *embeddings*) para orientar seu enquadramento correto e o desenvolvimento de uma solução de análise e detecção, com maior utilização de recursos semânticos presentes no discurso.

1.2 Objetivos

O objetivo deste trabalho é investigar, aplicar e discutir melhorias no modelo de AM profunda tabular, FT-Transformer, com o uso de *embeddings* textuais (estáticas ou contextuais) para as tarefas de detecção de ódio e toxicidade em comentários textuais (tweets) do X. Além do estabelecido de forma geral acima, espera-se também:

- Aplicar diferentes modelos para a representação de *features* textuais, como de Sentence-Transformers (REIMERS; GUREVYCH, 2019), BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), Albertina-PT (RODRIGUES et al., 2023) e alguns LLMs como *Text Embedders*;
- Testar e avaliar o desempenho da adaptação do modelo de aprendizado de máquina profunda desenvolvida para dados tabulares, FT-Transformer, proposto por Gorishniy et al. (2021);
- Discutir os resultados alcançados, uso de base de dados, quantidade de dados anotados e uso de recursos computacionais;

1.3 Contribuições

A relevância desta pesquisa, assim como suas contribuições, obtiveram reconhecimento pela comunidade acadêmica refletidos através da publicação e aceitação de trabalhos em conferências importantes.

- ***A Semi-Supervised Approach to Detect Toxic Comments* (SARAIVA et al., 2021):** Este trabalho propôs uma estratégia semi-supervisionada inovadora para a detecção de comentários tóxicos, utilizando uma abordagem baseada em grafos heterogêneos. O modelo proposto superou diversas abordagens baseadas em grafos e alcançou resultados competitivos em comparação com as arquiteturas Transformer, utilizando apenas 10% de dados rotulados. A pesquisa demonstrou a eficácia da modelagem dos dados em grafos heterogêneos e do classificador Gradient Boosting na detecção de comentários tóxicos, especialmente devido à capacidade de capturar o conhecimento morfológico, sintático e semântico das palavras mediante

embeddings estáticas, GloVe (PENNINGTON; SOCHER; MANNING, 2014). Apesar da abordagem implementada neste trabalho divergir da proposta desta dissertação, o estudo feito e os resultados obtidos corroboraram para a formação deste autor e também para o desenvolvimento desta pesquisa;

- ***A Transformer-based Tabular Approach to Detect Toxic Comments* (DAMAS et al., 2024):** Este trabalho apresenta uma abordagem inédita para a classificação binária de comentários tóxicos e discurso de ódio, utilizando um modelo de aprendizagem profunda para dados tabulares, FT-Transformer, e *embeddings* contextuais diversos, incluindo modelos específicos para o português brasileiro. A pesquisa explorou o uso de diferentes modelos de *embeddings*, assim como o uso de modelos de linguagem pré-treinados como geradores de *embeddings*, incluindo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) e Albertina PT-BR (RODRIGUES et al., 2023). A abordagem obteve resultados bastante interessantes principalmente com *embeddings* geradas pelo BERTimbau, SBERTimbau (versão Sentence Transformer do BERTimbau) e do modelo de *embedding* da OpenAI, o text-embedding-3-large. Com o text-embedding-3-large a abordagem atingiu taxa de acurácia de 76% e um F1-score de 75% no conjunto de teste. O estudo destaca a importância dos modelos de aprendizagem profunda para dados tabulares ou aprendizagem profunda tabular (APT), adaptado do inglês *Tabular Deep Learning* (TDL), e de modelos de *embedding* robustos ou de abordagens relativamente modernas de geração de *embeddings* a partir de modelos de linguagem pré-treinados. Este trabalho enfatiza a capacidade dessa abordagem e dos modelos de TDL de lidar com dados heterogêneos, multimodais e unimodais, além da eficiência no processamento e melhoria nos resultados ampliando sua aplicabilidade para tarefas de PLN como a detecção de discurso tóxico e de ódio nos comentários textuais de redes sociais.

1.4 Organização do texto

Excetuando este primeiro capítulo, que resumidamente introduz realidade e problemática acerca da Toxicidade e Ódio nas redes sociais, em especial o X, evidencia alguns desafios e ações para enfrentar a questão relacionada ao potencial nocivo no discurso em forma de comentários dos usuários, o presente trabalho segue organizado da seguinte forma:

- O Capítulo 2 apresenta uma breve revisão de literatura sobre métodos, modelos e trabalhos sobre Aprendizagem de Máquina (*Machine Learning*) e Processamento de Linguagem Natural (PLN) em adição aos trabalhos relacionados à análise do discurso e detecção de ódio e toxicidade em textos;
- O Capítulo 3 exibe o “estado da arte” e alguns trabalhos correlatos;

- O Capítulo 4 descreve a abordagem proposta;
- No Capítulo 5 descreve os experimentos realizados e os resultados relativos à pesquisa;
- E por fim, o Capítulo 6 apresenta as conclusões e trabalhos futuros.

2 Referencial Teórico

Esta seção tem como por objetivo principal e primordial apresentar uma breve, mas sucinta elucidação que permeia não só o Processamento de Linguagem Natural (PLN) como também o Aprendizado de Máquina, passando desde a abordagem clássica até os Grandes Modelos de Linguagem ou de Língua também conhecidos pelo seu termo em inglês *Large Language Models* (LLMs) e explorando também as formas de representação textuais.

2.1 Processamento de Linguagem Natural

A linguagem, elemento fundamental e idiossincrático do ser humano, é a chave para a comunicação e a construção da realidade. Abrangendo desde a fala e a escrita até gestos e imagens, ela nos permite interagir, compartilhar ideias e moldar o mundo ao nosso redor.

O Processamento de Linguagem Natural (PLN) e a Computação Linguística (CL), ou ainda Linguística Computacional, são subcampos e instersecções de estudo da Computação e da Linguística que empregam diversas técnicas para solucionar problemas relacionados à automatização da compreensão e geração de linguagens humanas naturais, assim como também envolvem os estudos teóricos dos aspectos linguísticos da comunicação humana em toda a sua diversidade e o desenvolvimento de novas abordagens computacionais para compreensão e representação desses dados e conhecimento (TSUJII, 2021). Conforme definido por Eisenstein (2019), o PLN abrange um conjunto de métodos cujo propósito é tornar a linguagem humana acessível às máquinas.

A linha do tempo, apresentada na Figura 1, pontua na história alguns dos acontecimentos, métodos e técnicas desenvolvidos de relevância para realização do propósito já citado do PLN. Um exemplo é o experimento IBM-Georgetown em 1954 (GARVIN, 1968), que envolveu a tradução completamente automática de mais de sessenta (60) frases russas para o inglês.

Outros marcos de extrema relevância para o PLN são as Representações de Vetores de Palavras, Word2Vec (MIKOLOV et al., 2013) em 2013 e GloVe (PENNINGTON; SOCHER; MANNING, 2014) em 2014, o modelo Transformer (VASWANI et al., 2017) de aprendizagem profunda introduzido em 2017, em 2019 os modelos BERT (DEVLIN et al., 2018) foram anunciados, e a partir dele diversos outros modelos de língua surgiram como GPT, T5, ALBERT, RoBERTa, ELECTRA, OPT, Cohere, AlexaTM, Codex, Sparrow, PaLM, Vicuna, Bard e LLaMA[1] (YANG et al., 2024).

[1] A Comprehensive Survey of Large Language Models (LLMs)

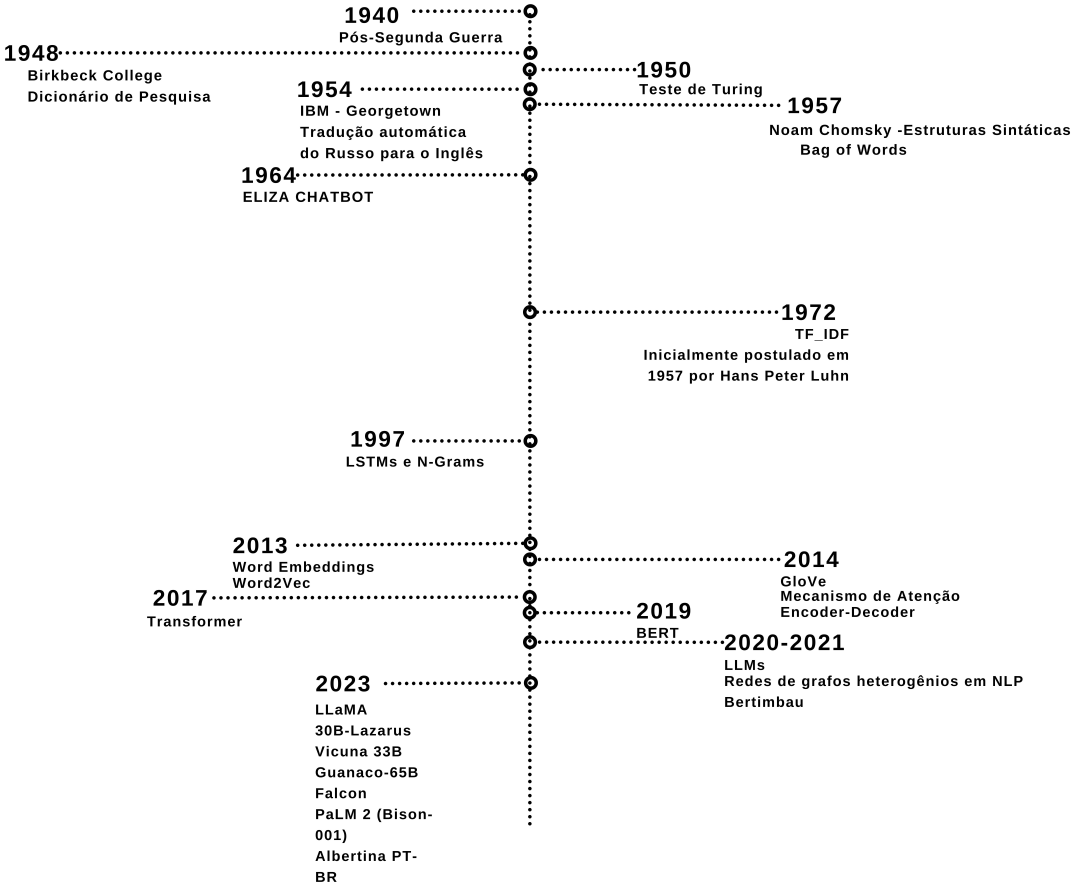


Figura 1 – Linha do tempo com os principais estudos e avanços do PLN (fonte: o autor).

O Processamento de Linguagem Natural estrutura-se resumidamente por um empilhamento de camadas não isoladas, assim como os estudos da linguagem humana, onde cada uma representa um aspecto da linguagem e o estudo linguístico, como representado na Figura 2.

Cada camada deste modelo representa um nível ou fase do estudo linguístico e por conseguinte também do Processamento de Linguagem Natural, onde o nível de complexidade cresce na direção da camada topo.

Em sua base, a Fonética e Fonologia estudam os sons (fonemas), sua organização e representação. Acima temos a Morfologia, onde acontece o que se chama de processamento morfológico do texto e a análise lexical (os caracteres e formação de palavras), na camada sobrejacente, a Sintaxe, estuda-se a formação de sintagmas e sentenças, acontece a análise sintática, incluindo além da gramática, as expressões, locuções e demais componentes atômicos e compostos de uma sentença. Na Semântica, se performam as análises semânticas, os estudos dos significados das palavras e das frases, a agregação de valor a cada componente, a lógica e o que implica cada componente de um texto. É nesta camada em que se concentra esta pesquisa e as abordagens empregadas para analisar o fenômeno em questão. Logo acima, tem-se a Pragmática, nesta camada analisam-se todas as nuances que envolvem a

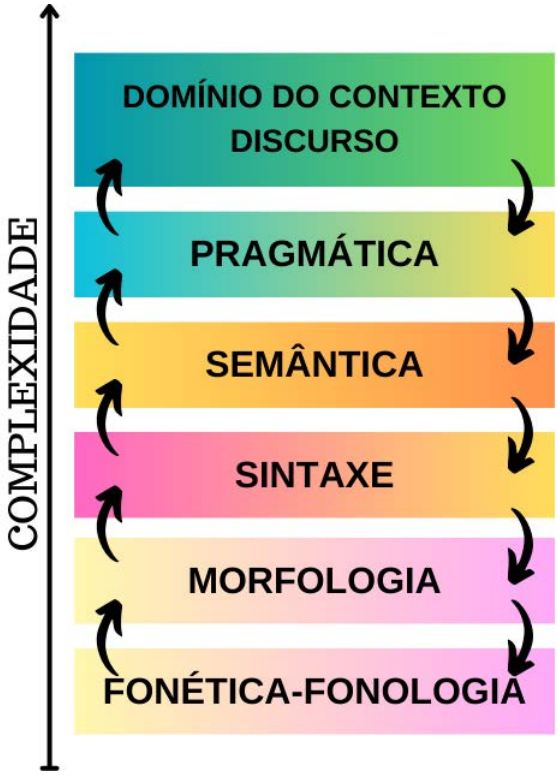


Figura 2 – Modelo de Camadas do Processamento de Linguagem Natural (Adaptado de Dale (2010), Schulz (2021), Caseli, Nunes e Pagano (2023)).

comunicação através da linguagem, aqui se incluem também as variações linguísticas, o coloquialismo, a ressignificação de termos, aqui se enquadra tudo o que pode influenciar o contexto comunicacional muito além da semântica e sintaxe. E por último, no topo da pilha temos o Discurso e o Domínio de Contexto, que envolve principalmente o raciocínio ou percepção contextual, a integração do discurso, a interpretação interlíngua ou intermídia (no contexto das mídias sociais) e a compreensão dos significados pretendidos pelo falante (DALE, 2010; GRūZīTIS, 2010; WIRIYATHAMMABHUM et al., 2016; WIJERATNE; SILVA, 2020; SCHULZ, 2021; ABERA; DECHASA, 2021; CASELI; NUNES; PAGANO, 2023).

Isto é, a linguagem é um sistema estruturado de comunicação que envolve combinações complexas de elementos constituintes, como caracteres, palavras, expressões, estruturas compostas e completas, sentenças e frases (VAJJALA et al., 2020). O PLN exige uma compreensão dos conceitos linguísticos que estruturam a linguagem humana, além de muitas outras nuances intrínsecas a própria dinâmica e evolução da linguagem. Um ponto crucial a se compreender é que os computadores têm uma aptidão inata para lidar eficientemente com linguagens de programação, mas enfrentam desafios significativos quando se trata da compreensão da linguagem natural. Isso acontece devido à diferença substancial entre linguagens de programação, que seguem regras gramaticais (sintáticas e semânticas) muito rígidas, tendo sido traduzidas em comandos de nível de circuitos digitais, e linguagens naturais, as quais são inerentemente intuitivas e não seguem uma

estrutura objetiva.

Em decorrência dessa característica, as abordagens baseadas unicamente em regras estão fadadas a ignorar as matizes presentes em qualquer forma de comunicação mediante Linguagem Natural Caseli, Freitas e Viola (2022). Para tanto, ao longo dos anos outras abordagens foram desenvolvidas a fim de comportar melhor o que estava além da regularidade definida por uma regra. Além desta, chamada de Simbólica, outras duas abordagens foram desenvolvidas e estão ainda em uso, uma delas é a abordagem Estatística, principalmente baseada no reconhecimento de padrões através frequência e probabilidade, e por último a abordagem Neural. Esta é a responsável pelos grandes modelos de língua hoje como o GPT-4 e o Albertina-PT, modelos criados a partir do treinando de redes neurais profundas que além da quantidade massiva de dados também possuem outros mecanismos para intensificar e aperfeiçoar o aprendizado.

2.2 Aprendizagem de Máquina

O Aprendizado de Máquina (do inglês: *Machine Learning* — ML) é um ramo da Inteligência Artificial (IA) que se concentra na criação de modelos e algoritmos que permitem aos computadores aprender com os dados e fazer previsões ou julgamentos sem programação explícita. É um campo com muitos tipos e abordagens diferentes em constante evolução. O principal objetivo do Aprendizado de Máquina é fornecer às máquinas a capacidade de executar melhor tarefas, aprendendo a partir de padrões e dados, em vez de instruções explícitas (TOM, 2019).

Pode-se enumerar quatro (04) tipos principais de aprendizado de máquina, são eles: aprendizado supervisionado, aprendizado não supervisionado, aprendizado por reforço e aprendizado semisupervisionado.

Aprendizado Supervisionado: Neste tipo de aprendizado de máquina, o modelo é treinado em dados rotulados, onde são fornecidos os dados de entrada e suas correspondentes saídas corretas. O modelo aprende a fazer previsões ou classificar dados novos e invisíveis, identificando padrões e relacionamentos nos dados de treinamento. As aplicações comuns de aprendizagem supervisionada incluem reconhecimento de imagem, detecção de spam e tradução de idiomas (RASCHKA; MIRJALILI, 2019).

Aprendizagem não supervisionada: essa aprendizagem lida com dados não rotulados, onde o algoritmo deve identificar padrões ou estruturas nos dados por conta própria. É frequentemente usado para tarefas como *clustering* e redução de dimensionalidade. Os algoritmos de cluster agrupam pontos de dados semelhantes, enquanto as técnicas de redução de dimensionalidade visam simplificar dados complexos, reduzindo o número de recursos, mantendo informações essenciais (RIVAS, 2020).

Aprendizagem por Reforço: é um arquétipo diferente onde um agente interage com um ambiente e aprende a realizar ações para maximizar uma recompensa cumulativa. O agente recebe *feedback* na forma de recompensas ou penalidades com base em suas ações, permitindo-lhe aprender estratégias ideais ao longo do tempo. As aplicações da aprendizagem por reforço incluem jogos, robótica autônoma e carros autônomos (DHINAKARAN et al., 2023).

Aprendizado Semisupervisionado: representa um paradigma que fica entre a aprendizagem supervisionada e a aprendizagem não supervisionada. Na aprendizagem semissupervisionada, o modelo é treinado em um conjunto de dados que contém dados rotulados e não rotulados (ABNEY, 2007). Esta combinação de dados rotulados e não rotulados torna a aprendizagem semissupervisionada particularmente útil em cenários onde a aquisição de dados rotulados é cara ou demorada, mas ainda fornece informações valiosas para o treinamento. A ideia principal por trás do aprendizado semissupervisionado é aproveitar os dados rotulados limitados com o conjunto maior de dados não rotulados para melhorar o desempenho do modelo. Ao fazer isso, generaliza melhor para exemplos inéditos. Esta abordagem aproveita o pressuposto de que a distribuição dos dados rotulados é semelhante à dos dados não rotulados, permitindo que o modelo aprenda padrões mais robustos e precisos.

Cada tipo de aprendizado de máquina possui seu próprio conjunto de algoritmos e aplicações, e a escolha de qual tipo usar depende do problema específico e dos dados disponíveis. A aprendizagem automática continua a avançar e encontrou aplicações em vários domínios, desde cuidados de saúde e finanças até entretenimento e sistemas autônomos, desempenhando um papel fundamental na definição do futuro da tecnologia e da automação.

2.2.1 Aprendizagem Clássica

O aprendizado de máquina clássico refere-se a uma abordagem fundamental em Inteligência Artificial (IA) que abrange algoritmos e técnicas desenvolvidas antes do surgimento do aprendizado profundo, suas bases foram estabelecidas entre as décadas de 50 e 60. A aprendizagem clássica abrange uma variedade de algoritmos e técnicas anteriores à revolução do aprendizado profundo, com foco em modelos que aprendem padrões e relacionamentos em conjuntos de dados, permitindo-lhes fazer previsões sobre dados novos e não vistos. Uma das características mais notáveis do aprendizado de máquina clássico é sua dependência da engenharia de recursos, onde o conhecimento do domínio é usado para extrair manualmente recursos relevantes de dados brutos para melhorar o desempenho do modelo (MOHAMMED; KHAN; BASHIER, 2016; BONACCORSO, 2018; RAY, 2019; REBALA; RAVI; CHURIWALA, 2019; MAHESH, 2020; SEN; MEHTAB; ENGELBRECHT, 2021).

Alguns dos avanços e modelos mais notáveis no aprendizado de máquina clássico incluem:

Neural Networks: as Redes Neurais são de longe as ferramentas computacionais mais fundamentais para o processamento de linguagem e precursoras das Redes Neurais Recorrentes (RNNs) e do Transformer. Essas redes são baseadas no modelo de neurônio artificial proposto por McCulloch e Pitts em 1943 (MCCULLOCH; PITTS, 1990), com uma única camada oculta, estes neurônios são os blocos de construção de redes neurais mais complexas. Eles podem aprender relações lineares entre recursos e resultados, estabelecendo as bases para avanços futuros.

Support Vector Machines (SVMs): excelentes para tarefas de classificação, as SVMs criam um hiperplano que separa melhor os pontos de dados pertencentes a diferentes categorias. Embora não sejam estritamente redes neurais, algumas formulações de SVM utilizam funções de kernel que introduzem não linearidade, semelhante à forma como as redes neurais com múltiplas camadas ocultas conseguem isso (LORENA; CARVALHO, 2007; RAY, 2019; MAHESH, 2020).

Decision Trees: essas estruturas semelhantes a árvores onde objetivo é criar um modelo que preveja o valor de uma variável alvo, aprendendo regras de decisão simples inferidas a partir dos recursos dos dados[2]. Embora não sejam redes neurais em si, as árvores de decisão podem ser combinadas com módulos de redes neurais para criar modelos híbridos que aproveitem os pontos fortes de ambas as abordagens[3].

K-Nearest Neighbors (KNN): é um método de aprendizagem supervisionado não paramétrico proposto por Fix (1985), que usa proximidade para fazer classificações ou previsões sobre o agrupamento de um ponto de dados individual, ou seja, classifica os pontos de dados com base na votação majoritária de seus vizinhos mais próximos nos dados de treinamento. Não envolve redes neurais, mas serve como um conceito fundamental para métodos de aprendizagem que usam distâncias mais complexas[4]. É um dos algoritmos de classificação e regressão mais populares e simples usados no aprendizado de máquina atualmente[5].

Outra classe de modelos bastante notável são as *Gradient Boosting Machines*, esta assim como alguns de seus exemplos serão pontuados a seguir devido sua relevância e utilização nas tarefas de AM, como classificação de textos e classificação de dados tabulares.

[2] scikit-learn - Decision Trees
[3] IBM - What is a decision tree?
[4] What is the k-nearest neighbors (KNN) algorithm?
[5] GeekforGeeks - K-Nearest Neighbor(KNN) Algorithm

2.2.1.1 Gradient-Boosted Decision Trees

Gradient-Boosted Decision Trees-GBDTs (CHEN; GUESTRIN, 2016) ou Árvores de Decisão otimizadas por gradiente são um conjunto de técnicas de aprendizado de máquina comumente utilizadas para aprimorar o valor preditivo de um modelo por meio de etapas sucessivas no processo de aprendizado. GBDTs são notadamente referenciadas como poderosas técnicas de Aprendizado de Máquina e uma ferramenta utilizada para as mais diversas aplicações que vão da classificação de textos (WADUD et al., 2022) até o estudo das rochas na exploração de petróleo Dev e Eden (2019). GBDT usa como método base de aprendizado as árvores de decisão e soma a predição resultante de uma série de árvores, os quais são sempre treinadas de modo a minimizar a diferença residual entre os valores reais alvo e os valores preditos pelo modelo.

GBDTs são bastante populares devido a sua acurácia, eficiência, interpretabilidade e, até os dias atuais, ainda é a técnica mais usada para competições de ML e também para processar dados nos formatos tabulares (GORISHNIY et al., 2021; GRINSZTAJN; OYALLON; VAROQUAUX, 2022), sendo o grupo de técnicas de aprendizagem de máquina não profunda que entrega resultados competitivos, e muitas vezes que são o estado da arte[6]. Ainda assim, o estado da arte em problemas com dados tabulares por muito tempo se resumiram quase que exclusivamente ao uso das GBDTs. Pode-se citar inclusive boas bibliotecas, bem estabelecidas e com documentação ampla para o uso, desenvolvimento e aplicação de GBDTs como o XGBoost[7](CHEN; GUESTRIN, 2016), LightGBM[8](KE et al., 2017), o CatBoost[9](DOROGUSH; ERSHOV; GULIN, 2018) e a ThunderGBM[10](WEN et al., 2020).

2.2.1.1.1 Gradient Boost Classifier

A utilização muito comum de uma GBDT é como modelo de classificação, para tanto um conjunto de classificadores múltiplos ou comitê de classificadores, comumente conhecidos como *Ensemble Classifiers*, são utilizados para fins de classificação com melhora significativa do desempenho em comparação ao uso de classificadores individuais, aprimorando assim também sua precisão. Técnicas comuns, como florestas aleatórias (Random Forests), dependem de uma média direta dos modelos no conjunto. Em contraste, os métodos de reforço seguem uma estratégia distinta e construtiva para a formação de conjuntos. Já o *boosting*, introduzido por Schapire (1990), representa um meta-algoritmo que pode ser conceituado como uma técnica para cálculo da média de um modelo ou conjunto de modelos. É um dos métodos de conjunto mais amplamente empregados e está entre os conceitos

[6] Analytics Vidhya - Introduction to XGBoost Algorithm in Machine Learning
[7] dmlc XGBoost - eXtreme Gradient Boosting
[8] Microsoft LightGBM — Light Gradient Boosting Machine
[9] Yandex CatBoost
[10] ThunderGBM

de aprendizagem mais influentes introduzidos nas últimas duas décadas. Originalmente concebido para tarefas de classificação, o boosting também provou ser vantajoso quando estendido a problemas de regressão. A etapa inicial envolve a criação de um classificador "fraco", o que significa que sua precisão no conjunto de dados de treinamento só precisa ser marginalmente superior à suposição aleatória. Posteriormente, uma sequência de modelos é construída iterativamente, com cada modelo treinado em um conjunto de dados no qual os pontos mal classificados (ou, no caso de regressão, aqueles que são mal previstos) pelo modelo anterior recebem maior peso (FREUND; SCHAPIRE; ABE, 1999).

Para preencher a lacuna com a estrutura estatística, foi introduzida uma formulação de métodos de reforço baseada em gradiente descendente (FREUND; SCHAPIRE, 1997; FRIEDMAN; HASTIE; TIBSHIRANI, 2000; FRIEDMAN, 2001). Esta formulação, conhecida como máquinas de aumento de gradiente, ou GBMs, forneceu uma base teórica para os hiper parâmetros do modelo e serviu como metodologia para o desenvolvimento posterior de modelos de aumento de gradiente. Nas máquinas de aumento de gradiente, ou *Boosting*, o processo de aprendizagem envolve o ajuste sucessivo de novos modelos para melhorar a precisão da estimativa da variável de resposta. O conceito fundamental por trás deste algoritmo é projetar novos modelos-base que sejam altamente correlacionados com o gradiente negativo da função de perda associada a todo o conjunto, embora a escolha da função de perda possa ser arbitrária, e normalmente é Mayr et al. (2014), Dutta, Kim e Dominic (2020).

Dentre os GBDTs uma das escolhas mais frequentes em ML *hackathons* é o XGBoost, outro método de *boosting* com aprimoramento de desempenho, e por diversas razões ele se mantém como a primeira escolha, algumas delas são:

- Alta performance: foi otimizado para ser rápido e eficiente, sendo apropriado para grandes quantidades de dados e aplicações em tempo real;
- Métodos de Regularização: prevenir o overfitting e a aumento demasiado da capacidade generalização;
- Habilidade com dados incompletos: minimiza a necesidade de prepocessamento e imputação de dados.

Para melhor ilustrar a arquitetura de um Gradiente Boost Classifier, arquitetura essa que também pode ser utilizada como referência ilustrativa para o XGBoost Classifier, utilizar-se-á a Figura 3. Nela um conjunto de dados é submetido a um conjunto de modelos-base para treinamento e teste, de maneira sequencial. Em geral, as Árvores de Decisão são os exemplos de classificadores mais utilizados para compor este esquema, e que geralmente é apelidado de modelo base, fraco ou *weak learner*, não sendo este modelo o único e exclusivo ao implementar uma GBDT. Neste esquema, o primeiro modelo recebe

o conjunto original de dados e os modelos subsequentes aprendem a partir dos erros de seu antecessor imediato, resultando ao final na interseção de todos os erros e acertos dessa sequência.

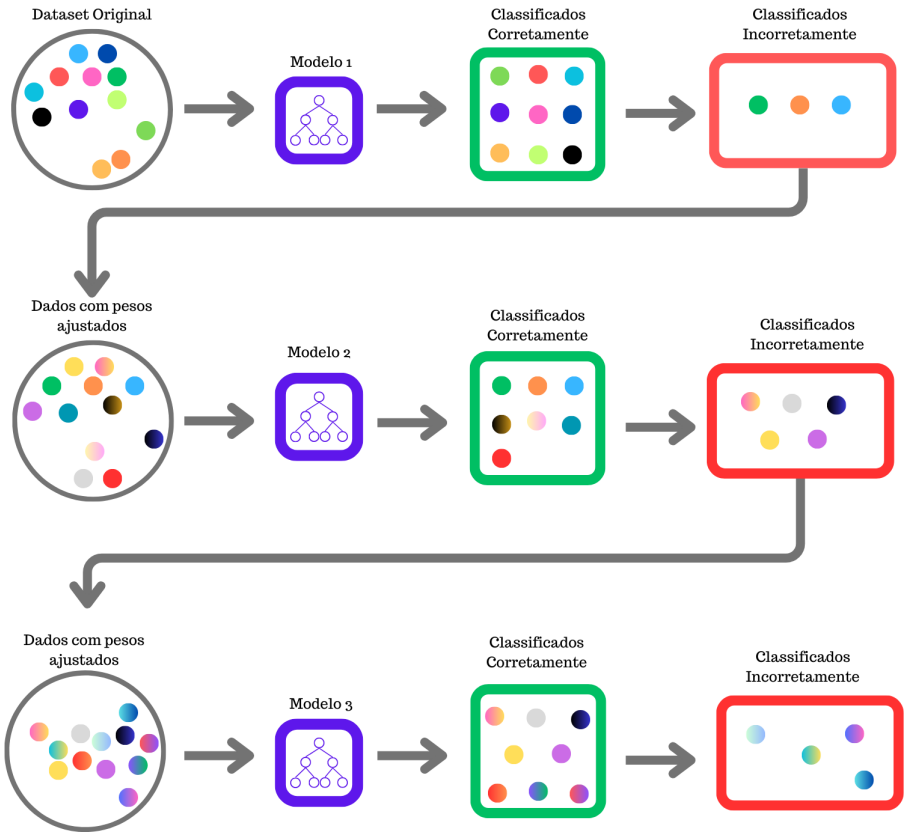


Figura 3 – Exemplo de Método Ensemble: Gradiente Boost Classifier (fonte: o autor).

Como na Figura 3, o XGBoost também é composto por um conjunto de métodos de aprendizagem base paralelos e a eles são empregadas técnicas de regularização com intuito de aumentar a capacidade de generalização do modelo com atenção para a prevenção do *overfitting*. Basicamente o XGBoost opera adicionando novos modelos-base ao comitê, onde a cada nova inclusão desses métodos-base estes focam-se em corrigir os erros computados pelos colegas de comitê anteriormente adicionados. Durante o treinamento desse comitê é aplicada uma técnica de otimização de gradiente descendente com objetivo de minimizar a função de perda predefinida (CHEN; GUESTRIN, 2016).

Após ser reconhecido pelo CERN (*Conseil Européen pour la Recherche Nucléaire* ou Organização Europeia para a Pesquisa Nuclear) como a melhor abordagem para classificar sinais do *Large Hadron Collider*-LHC Adam-Bourdarios et al. (2015), Abbas et al. (2020) o XGBoost emergiu como a mais proeminente, robusta, direta e útil técnica para classificação em diversos outros campos de aplicação, inclusive aqueles que exigiam dos modelos uma grande habilidade de lidar com relações complexas (dependências) nos dados, muito comum em dados textuais, grafos e dados tabulares, alta escalabilidade, processamento paralelo, alto desempenho preditivo e grande versatilidade entre diferentes bases de dados. Uma

dessas características, a capacidade de tratar dependências de dados, fez com que essa técnica também obtivesse grandes resultados para processar dados em formato tabular e se opusesse durante muito tempo com excelentes resultados as redes neurais, principalmente aos modelos de Aprendizagem Profunda.

2.2.2 Aprendizagem Profunda ou *Deep Learning*

A Aprendizagem Profunda — AP (do inglês: *Deep Learning*) é uma classe particular de métodos de Aprendizado de Máquina com a capacidade de alcançar poder e flexibilidade enquanto aprende a representar os dados como uma hierarquia aninhada de conceitos, com cada conceito definido em relação a conceitos mais simples e representações mais abstratas computadas em termos de conceitos menos abstratos (TIWARI; TIWARI; TIWARI, 2018).A modelagem de uma AP baseia-se na forma como o cérebro humano processa dados, no futuro, ao longo das etapas de aprendizagem essa técnica obterá mais sucesso devido à sua capacidade de aprender em vários níveis de representação com pouca engenharia manual.

A Aprendizagem Profunda é um subconjunto da Aprendizagem de Máquina e também um tipo de Rede Neural Artificial, ajudando a diferenciar principalmente essas abordagens, é o número de camadas intermediárias(ou ocultas). A Figura 4 apresenta dois exemplos Redes Neurais, uma considerada Rasa e outra considerada Profunda.

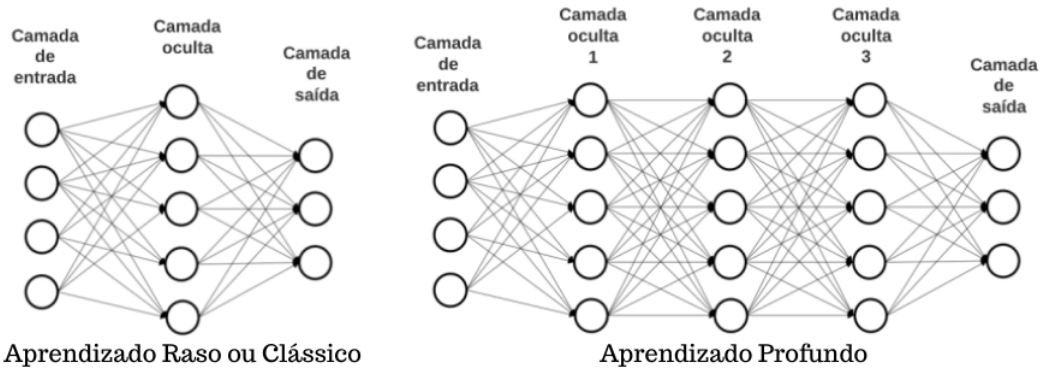


Figura 4 – Exemplo de Redes Neurais Artificais Rasa e Profunda (fonte: Packt - Why deep learning?).

A Aprendizagem Profunda é um nicho de métodos que vem sendo utilizado com bastante recorrência em diversas áreas e para diversos propósitos (CAO, 2022), desde a análise de imagens de CT Scan para detecção de lesões pulmonares causadas pelo COVID-19 (TER-SARKISOV, 2022), a melhoria nas estratégias de recomendação conforme as preferências do usuário (LIU; JIANG; ZHANG, 2022), até a detecção das mais variadas formas de discurso ofensivo, como o discurso de ódio (WANG, 2023; TOKTAROVA et al., 2023) e o discurso tóxico (MALIK; AGGRAWAL; VISHWAKARMA, 2021).

2.2.2.1 Transformer

É inegável que o advento das redes neurais Transformer (VASWANI et al., 2017) trouxeram mudanças significativas ao cenário da IA, do aprendizado de máquina e principalmente do PLN de modo geral nos últimos anos. Embora os modelos inspirados em Transformers tenham se destacado em diversas tarefas, os mecanismos subjacentes a essas conquistas nem sempre puderam ser bem compreendidos (BRASOVEANU; ANDONIE, 2020; GILLIOZ et al., 2020; SINGH; MAHMOOD, 2021).

Para compreender verdadeiramente um texto, em todas as suas nuances, é necessário compreender tanto o seu conteúdo (por exemplo, o significado das palavras, expressões, frases ou sentenças) quanto o contexto (por exemplo, onde, quando e por que o texto foi gerado e em que circunstâncias). Assim, a construção de redes que operem nestes diversos níveis representa um desafio devido à disponibilidade limitada de dados de formação Gillioz et al. (2020).

A arquitetura original de uma rede Transformer foi introduzida em 2017. Vaswani et al. (2017) demonstrou a possibilidade de modelar redes que podem alcançar excelentes resultados em tarefas de Processamento de Linguagem Natural (PLN) através de uma série de camadas de atenção sequenciais. O modelo do Transformer por si só é muito simples, composto por codificador e um decodificador. O codificador normalmente compreende várias camadas de atenção com vários cabeçotes, camadas de adição e normalização e camadas de feedforward. O mecanismo de atenção *multi-head* equipa o modelo com a capacidade de coordenar informações de diferentes subespaços de representação (por exemplo, matrizes de pesos múltiplos) em várias posições (por exemplo, palavras diferentes em uma frase), ou seja, uma capacidade de atenção multidirecionada (VASWANI et al., 2017). O modelo arquitetural para o Transformer proposto por Vaswani et al. (2017) é apresentado na Figura 5.

Nesse modelo, suas saídas podem alimentar outros codificadores (em verde) ou decodificadores (em azul) presentes em na arquitetura. O número de codificadores e decodificadores pode variar de arquitetura para arquitetura, mas geralmente eles estão dispostos de maneira emparelhado (por exemplo, 10 codificadores e 10 decodificadores). Em outras arquiteturas, codificadores e decodificadores podem atuar em tarefas distintas (por exemplo, codificador para resposta a perguntas e decodificador para compreensão de texto) (MAITHRA; SCHMIDT, 2020). Embora o desenvolvimento inicial do modelo tenha sido para tarefas de tradução automática, desde então ele foi aplicado em vários domínios e provou ser eficaz.

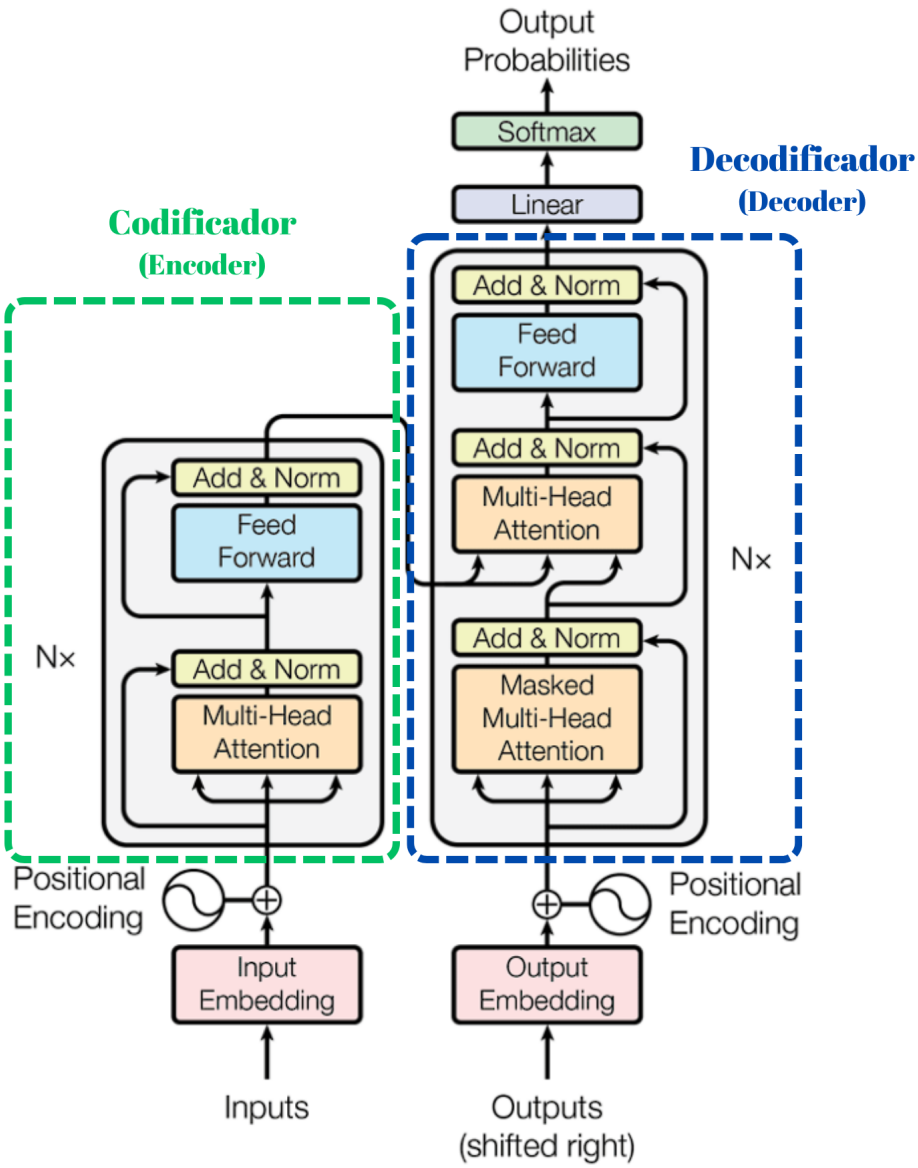


Figura 5 – Arquitetura original do modelo Transformer proposto por Vaswani et al. (2017).

2.2.2.2 BERT e outros LLMs

Um *Language Model* (LM) ou Modelo de Linguagem, ou também Modelo de Língua, é um modelo de aprendizado de máquina treinado para conduzir uma distribuição de probabilidade sobre palavras, de modo a aprender os padrões subjacentes da linguagem (PICOVOICE, 2023) e visa prever e gerar uma linguagem plausível (SMITH, 2023; DATABRICKS, 2024; BHAT, 2024; NAVEED et al., 2023). Exemplicando, um LM tenta prever a próxima palavra mais apropriada para preencher um espaço em branco em uma sentença, frase ou texto, com base no contexto do texto fornecido, tarefa conhecida como preenchimento automático. Basicamente, LMs funcionam estimando a probabilidade de um *token* (instância de uma sequência de caracteres em algum documento ou texto específico agrupados como uma unidade semântica útil para processamento) (SCHÜTZE;

MANNING; RAGHAVAN, 2008) ou sequência de *tokens* ocorrerem em uma sequência mais longa de *tokens* (GOOGLE, 2023).

Um *Large Language Model* (LLM) ou Grande Modelo de Linguagem (ou ainda modelo de linguagem de grande porte) são modelos de linguagem pré-treinados com escala de parâmetros significativa, estes parâmetros são os pesos (*weights*) que o modelo aprende ao longo do treinamento. O termo *Large* refere-se não só ao número de parâmetros e a quantidade de dados de treinamento, mas também ao seu poder de processamento e as suas capacidades relacionadas a sequências de texto mais longas. Assim como o LM, o LLM também é um modelo de aprendizado profundo proficiente, que utiliza em sua maioria arquiteturas Transformer e passam por intenso treinamento com extensos conjuntos de dados, o que lhe garante substancialidade e lhe permite reconhecer, traduzir, prever ou gerar texto e outras formas de conteúdo alcançando melhoria de desempenho, resultado e habilidades especiais não presentes em modelos de linguagem de pequena escala (ZHAO et al., 2023; MINAEE et al., 2024).

Um Modelo BERT ou GPT, são os exemplos de LLMs mais conhecidos, tendo sido desenvolvidos a partir do refinamento de modelos Transformer com quantidade realmente massiva de dados. Outro LLM muito conhecido é o ELMo (PETERS et al., 2018), diferentemente do BERT e GPT, este é puramente baseado em uma LSTM Li et al. (2020). Outra característica do BERT e ELMo bem peculiar é sua abordagem bidirecional de aprendizado e representação de informação, enquanto as duas primeiras versões do GPT eram unidirecionais Li et al. (2020), Peters et al. (2018). A bidirecionalidade do BERT e ELMo permitem com que estes modelos apossem-se de aspectos dependentes do contexto em relação ao significado de uma palavra (GHATI, 2020), considerando tanto o contexto à esquerda quanto à direita das palavras em uma sentença de forma simultânea, propiciando a incorporação de relações complexas e dependências entre as palavras (GEEKSFORGEEKS, 2024), maior captura de aspectos semânticos.

De 2017 até 2024, surgiram vários artigos introduzindo diversos modelos de linguagem inspirados em Transformers, incluindo nomes bem conhecidos como BERT (DEVLIN et al., 2018), RoBERTa (LIU et al., 2019), AlBERT (LAN et al., 2019), XLNet (YANG et al., 2019), DistilBERT (SANH et al., 2019), Reformer (KITAEV; KAISER; LEVSKAYA, 2020), LLaMA (TOUVRON et al., 2023), além de alguns modelos para Língua Portuguesa como o Bertimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), Albertina-PT (RODRIGUES et al., 2023)(que inclui PT-PT[11] e PT-BR[12]), o Sabiá (PIRES et al., 2023) e o Sabiá-2 (ALMEIDA et al., 2024).

De forma geral, esses modelos apresentam uma maior complexidade, principalmente ao que se refere às melhorias implementadas em relação aos primeiros modelos de

[11] PT-PT: Português de Portugal e outros países
[12] PT-BR: Português do Brasil

Transformer e BERT. A explicabilidade desses modelos através de seu fluxo e resultados do processamento de informações é uma tarefa bastante desafiadora (BRASOVEANU; ANDONIE, 2020), e para ajudar nesse processo algumas abordagens de visualização emergiram como recursos para auxiliar na interpretabilidade dos modelos e seus resultados, como VizBERT (AKEN et al., 2020), AttViz (SKRLJ et al., 2020), BERT Attention Analysis (CLARK et al., 2019) e Análise de Transformer baseado em Norma-Vetor (KOBAYASHI et al., 2020), sem esquecer de mencionar também SHAP (RIBEIRO; SINGH; GUESTRIN, 2016) e XAI (SAMEK; MÜLLER, 2019).

2.2.2.3 Aprendizagem Profunda Tabular

Dados tabulares são a forma mais comum de tipo de dados em diversas aplicações reais e na maiorias das competições de *Machine Learning* (HUANG et al., 2020; GORISHNIY et al., 2021) ao redor do mundo. Contudo, esse tipo de dado fora até pouco tempo atrás negligenciado nas pesquisas em *Deep Learning*. Muitos métodos de DL geralmente se sobressaem em diversos cenários onde ML é aplicada, incluindo medicina, mercado financeiro, clima, industria de bens e serviços, dentre vários outros campos cujos dados estão em bancos de dados relacionais (SHWARTZ-ZIV; ARMON, 2022).

Durante a última década, métodos de ML clássicos como *Gradient-Boosted Decision Trees* (GBDT) dominaram e mostravam desempenho superior a abordagens de DL quando aplicadas a problemas geralmente representados por dados tabulares, principalmente pelo tempo reduzido de treinamento requerido e robustez (SHWARTZ-ZIV; ARMON, 2022). Quando redes neurais profundas são aplicadas a dados tabulares, alguns desafios costumam emergir, como diversidade e mistura de diferentes tipos de características (numérico, ordinal e categórico), escassez de metadados, problemas relacionados a localidade (SCHNEIDER, 2021) e dados faltantes (SHWARTZ-ZIV; ARMON, 2022).

Algumas abordagens de AP para dados tabulares apresentadas nos últimos anos alcançaram desempenho compatível a modelos tradicionais como GBDT e XGBoost, e em alguns casos obtiveram desempenho promissor, oferecendo menor complexidade na otimização em comparação aos modelos de AP mais utilizados como CNN (CHEN et al., 2023b; CHEN et al., 2023c) e em alguns casos não requerendo pré-processamento de dados (ARIK; PFISTER, 2021). Dentre essas técnicas, pode-se listar algumas como TabTransformer (HUANG et al., 2020), Trompt (CHEN et al., 2023c), TabPFN (HOLLMANN et al., 2022), TabNet (ARIK; PFISTER, 2021), SAINT (SOMEPALLI et al., 2021), ExcelFormer (CHEN et al., 2023b), NODE (POPOV; MOROZOV; BABENKO, 2019), XTab (ZHU et al., 2023), GatedTabTransformer (CHOLAKOV; KOLEV, 2022), GANDALF (JOSEPH; RAJ, 2023), TabCaps (CHEN et al., 2022a), DANets (CHEN et al., 2022b), TabLLM (HEGSELMANN et al., 2023), ResNet e FT-Transformers (GORISHNIY et al., 2021).

O FT-Transformer (*Feature Tokenizer + Transformer*) ou FTT, uma técnica que

utiliza uma simples adaptação ao Transformer proposto por Vaswani et al. (2017) que será um pouco mais detalhada nos seguintes parágrafos em vista que este será a técnica principal a ser explorada nesta pesquisa. FT-Transformer é uma técnica relativamente recente desenvolvida para processamento e análise de dados no formato tabular, sendo similar a outro modelo chamado AutoInt (SONG et al., 2019) que também modela as características de entrada (numéricas e categóricas) e suas interações em um mesmo espaço dimensional. Diferentemente do AutoInt, o FT-Transformer tem como seu foco primário a conversão dos dados em *embeddings* para então serem processados por uma pilha de camadas de Transformers (SCHUMACHER, 2024), cada camada operando no nível de característica de um objeto. Mais detalhes sobre a arquitetura FT-Transformer, assim como exemplos ilustrativos de seus componentes, estão presentes na seção 4.2 do capítulo 4 mais adiante.

A primeira etapa compreendida nesta arquitetura envolve um *Feature Tokenizer* ou FT que transforma todas as características contidas na entrada em *embeddings*. As *embeddings* são então passadas e processadas pela pilha de Transformer. Nesta mesma etapa a *embedding* do token de saída ou token de classificação (CLS) é acrescentado a primeira e última camadas da pilha, onde a última representação do token de saída é utilizada para a predição.

O FT-Transformer assim como outros modelos desenvolvidos para dados tabulares são geralmente aplicados em conjuntos de dados compostos por tipos numéricos, categóricos e textuais, ainda podendo conter *embeddings* já pre-processadas em diferentes campos da ciência e indústria. Até o momento há poucos trabalhos cujo foco seja a classificação de textos e inexistentes para as mais variadas análises que envolvem dados oriundos das redes sociais, o que coloca essa técnica, assim como outras anteriormente citadas, em evidência para ser devidamente explorada e utilizada para as mais variadas tarefas que envolvem classificação de dados textuais.

2.3 Representação de Texto

Para o PLN, as palavras são geralmente consideradas as unidades primárias significativas para o processamento (FINATTO et al., 2023). Estruturas de alto nível em um idioma, como frases e sentenças, são ainda compostas por palavras. Para o ser humano poder compreender uma língua, é crucial compreender o significado das palavras. Portanto, é essencial representá-las com precisão para que então os modelos possam compreender, categorizar ou gerar melhor texto em tarefas de PLN. O processo de representação em si não é tão somente um processo de extração de características, mas sim de transformação do texto em dados compreensíveis ao processamento computacional.

Uma palavra pode ser naturalmente representada como uma sequência de vários

caracteres. Nas tarefas de PLN, principalmente aquelas que envolvem a classificação de um determinado texto, este deve ser convertido de forma que o computador possa entendê-lo e processá-lo, a esse processo se dá o nome genérico de Representação de Texto (*Text Representation*), pode-se aqui entender e estender esse conceito para comportar o conceito de Incorporação de Palavras, prioritariamente *Word embeddings* (WE), como também Representação de Conhecimento, outros termos muito utilizados para descrever essa necessidade de se transpor a linguagem natural para uma linguagem processável pela máquina (NASEEM et al., 2020; LIU; SUN, 2023).

A ineficácia em representar em caracteres brutos, em relação à representação numérica se dá principalmente pelos comprimentos variáveis das palavras e seu uso em métodos de aprendizado de máquina, que sempre recebem os dados de entrada em alguma forma de representação numérica.

Existe também uma representação chamada de *one-hot*, sendo outra abordagem natural para representar palavras, atribuindo um índice exclusivo a cada palavra (LIU; SUN, 2023). A representação única possui limitações para representar palavras. Esta representação*one-hot* não é capaz de capturar a relação semântica entre as palavras, é esparsa, de alta dimensão, o qual é muito ineficiente, havendo a necessidade de se utilizar técnicas de redução de dimensionalidade e ainda sua inflexibilidade para representar novas palavras, o que acaba por requerer a atribuição de novos índices para novas palavras e mudando a dimensões da representação, que pode levar a problemas em aplicações do PLN.

Com intuito de solucionar o problema da representação *one-hot*, abordagens de representação distribuída foram propostas. A hipótese distributiva (FIRTH, 1957; HARRIS, 1954) de que objetos linguísticos com distribuições semelhantes têm significados semelhantes é a base para o aprendizado da representação distribuída de palavras. Com base nessa hipótese, diversos modelos de representação de palavras, como CBOW (*Continuos Bag-of-Words*) e Skip-gram, foram propostos e aplicados em diferentes áreas (LIU; SUN, 2023).

A Figura 6 desenreda uma cronografia no desenvolvimento das formas e técnicas de representação de características textuais do Processamento de Linguagem Natural.

A ilustração apresenta desde formas de representação simbólica (em amarelo), como também distribuídas (em azul). Todas têm seu papel de destaque no PLN, algumas receberam mais atenção ao longo dos anos e conforme sua capacidade representativa correspondia positivamente na solução de tarefas que envolvem o processamento de linguagem, estas foram ganhando ainda mais evidência, servindo de base e evoluindo para formas mais robustas de representação, como o Word2Vec e os modelos de língua pré-treinados.

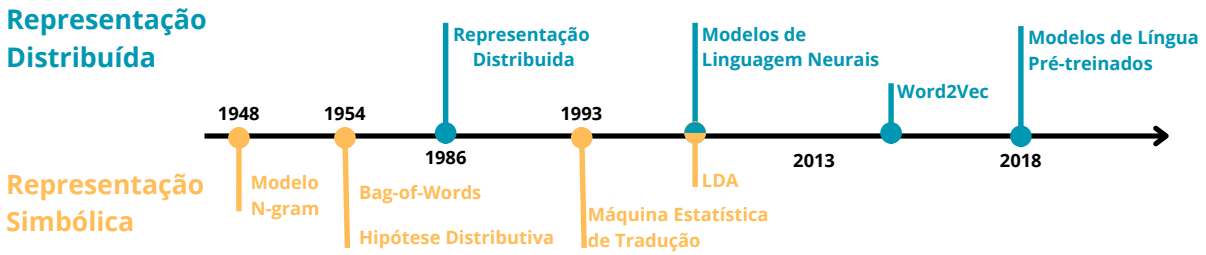


Figura 6 – Cronografia das Representações textuais distribuídas e simbólicas. (fonte: Liu e Sun (2023).

Word embeddings (WE) ou Vetor de Incorporação de Palavras é uma abordagem com a qual representamos documentos e palavras contidas num determinado *córpus*. Define-se como uma entrada vetorial numérica, que permite que palavras com significados semelhantes tenham a mesma representação, podendo aproximar o significado e representar uma palavra em um espaço dimensional reduzido (WADUD; D.; RAHMAN, 2022). Russell e Norvig (2013) asseveram que o uso de WE acaba tendo mais utilidade do que o uso de vetores binários em praticamente todas as aplicações de aprendizado profundo para tarefas de PLN, e em alguns casos de aprendizado clássico também. Podemos distinguir as *Word embeddings* em duas gerações, a primeira baseada frequência de vetores de palavras e uma segunda conhecida como uma representação livre de contexto (MARS, 2022).

Antes da ascensão das representações distribuídas (contextuais e não-contextuais) e da Aprendizagem Profunda, outras formas de representar um texto dominavam as aplicações em PLN e alguns ainda são muito utilizados hoje. Essas baseiam-se na frequência das palavras, principalmente, estes métodos transformam o texto em um vetor que contém o número de palavras contidas no texto(ou documento). Em suma, essa classe de métodos de representação pode ser dividida em Representação Categórica de Palavras (*Categorical Word Representation*) e na Representação de Palavras Ponderadas(*Weighted Word Representation*) (NASEEM et al., 2020).

Minuciando um pouco, a representação categórica de palavras é a abordagem mais direta e simples para representação de texto, nela as palavras são representadas simbolicamente como “1” ou “0”. Aqui, dois modelos se enquadram como métodos de representação de palavras categóricas: *One-Hot Encoding* e *Bag-of-Words* (BoW).

One-Hot Encoding: É a forma mais básica de representação de texto, a dimensionalidade é igual ao número de termos presentes no vocabulário. Cada termo do vocabulário é representado como uma variável binária, 0 ou 1. Especificamente, cada palavra é composta por zeros e uns, com o índice da palavra correspondente marcado como 1, enquanto todas as outras dimensões são marcadas como zero (0). Cada única palavra recebe uma dimensão única sendo representada por 1 nessa dimensão, com zeros em outros lugares.

Bag-of-Words (BoW): BoW é uma extensão da codificação one-hot. Envolve resumir

as representações únicas das palavras numa frase. BoW encontra aplicação em vários campos, incluindo Processamento de Linguagem Natural (PLN), visão computacional e recuperação de informação, entre outros. A matriz de palavras criada usando BoW não considera as relações semânticas entre as palavras e desconsidera a ordem das palavras e a gramática nas frases.

O BoW estende o *One-Hot Encoding* codificando *tokens* no vocabulário como vetores codificados 1-hot. No entanto, à medida que o tamanho do vocabulário aumenta, também aumenta o comprimento dos vetores. Isso pode resultar em uma matriz esparsa com muitos "0s", que não fornece informações sobre a ordem do texto ou a gramática usada nas frases (LIU et al., 2020).

Adiante, os métodos de representação de palavras ponderadas, que incluem *Term Frequency* (TF) e *Term Frequency-Inverse Document Frequency* (TF-IDF) estão relacionados à representação categórica de palavras, mas incorporam representações numéricas baseadas na frequência das palavras contidas no texto.

Frequência do termo (TF): TF é um método direto de extrair recursos de texto. Ele calcula a frequência com que uma palavra aparece em um documento. Uma palavra pode ocorrer com mais frequência em documentos maiores do que em documentos menores. Portanto, o TF é calculado dividindo a frequência de uma palavra pelo número total de palavras no documento.

Frequência de documento inversa de frequência de termo (TF-IDF): TF-IDF foi introduzido para mitigar o impacto de palavras comuns como "o" e "e" em um corpus. TF significa *Term Frequency*, conforme explicado anteriormente, e IDF significa *Inverse Document Frequency*, uma técnica usada em conjunto com TF para reduzir a influência de palavras comuns. O IDF atribui pesos maiores a palavras mais frequentes, mas que ocorrem em poucos documentos. A combinação TF-IDF é expressa matematicamente, onde TF é multiplicado pelo logaritmo do número total de documentos dividido pelo número de documentos contendo o termo. O TF-IDF é construído sobre o conceito do modelo BoW e, portanto, não captura ordem de palavras, semântica ou informações sintáticas. É adequado para uso como um recurso de nível léxico (WADUD; D.; RAHMAN, 2022).

Entretanto, estas não são as únicas arquiteturas para o treinamento das variadas Word Embeddinds existentes. Algumas delas são: Latent Semantic Analysis (LSA) (DUMAIS et al., 1988), Neural Network Language Model (NNLM) (BENGIO; DUCHARME; VINCENT, 2000), Max–Margin Loss (MML) (COLLOBERT; WESTON, 2008), e Log-Bilinear Language Model (LBL) (MNIH; HINTON, 2007), FastText (JOULIN et al., 2016; BOJANOWSKI et al., 2017), Global Vectors (GloVe) (PENNINGTON; SOCHER; MANNING, 2014), CoVe (Context Vectors model) (MCCANN et al., 2017), Virtual Model: Order (LAI et al., 2016), SENNA (COLLOBERT et al., 2011) e FLAIR (AKBIK; BLYTHE; VOLLGRAF, 2018).

Dentre as citadas anteriormente, o Word2Vec é o caso mais notório. Nele existem duas arquiteturas para aprendizado de representações de palavras, o CBOW (*Continous Bag-of-Words*), que tenta predizer uma palavra baseado no contexto a que está inserida, e o Skip-gram, que dada uma palavra, tenta-se predizer as palavras que estão à sua volta, esquerda e à direita (MIKOLOV et al., 2013). O Word2Vec é um embedding dito estático, significando que para cada palavra no vocabulário gerado, tem-se um único embedding, ou vetor que a represente. Por exemplo, um *Word Embedding* com 50 valores pode representar 50 recursos exclusivos.

A seguir, duas representações se destacam de maneira geral e se tornaram a base para muitas outras representações que surgiram depois (LI et al., 2020), são elas: o ELMo (*Embeddings from Language Models*) (PETERS et al., 2018) e o BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2018), também conhecidas como *contextual embeddings* (*embeddings contextuais*). Em vez de mapear uma palavra para um vetor fixo, como muitas das abordagens mencionadas anteriormente, ELMo e BERT usam redes neurais multicamadas para construir representações contextualizadas dinâmicas de cada palavra com base em seu contexto específico no texto, o que é especialmente útil para aquelas palavras que possuem diversos sentidos (significados) (LIU; SUN, 2023).

O ELMo emergiu para superar as limitações das *embeddings* de palavras anteriores, suas representações são uma função presente nas camadas internas do modelo de linguagem bidirecional (biLM). Mesmo a entrada sendo uma sequência de *tokens*, o modelo de linguagem aprende a prever a probabilidade do próximo token dado aos *tokens* anteriores (MARS, 2022). Essa sequência contém *tokens* antes do token alvo, enquanto em uma passagem direta. Na passagem anterior, o histórico contém palavras após o token em questão. A arquitetura BiLSTM ajuda o ELMo a aprender aspectos mais dependentes entre contexto e os significados das palavras nas camadas superiores, juntamente com aspectos de sintaxe nas camadas inferiores. Isso leva a melhores incorporações de palavras e diferentes representações de uma palavra, dependendo do contexto em que ela aparece. (PETERS et al., 2017) demonstraram que as representações baseadas em biLM superaram o CoVe (*Contextualized Word Vectors*)em todas as tarefas consideradas, principalmente porque o CoVe depende de dados multilíngues cuja quantidade é limitada (WANG et al., 2020b). ELMo combina representações de LSTMs progressivos e retrógrados sem considerar as interações entre os contextos esquerdo e direito. Por outro lado, GPT (RADFORD et al., 2018) e GPT2 (RADFORD et al., 2019) empregam um decodificador da esquerda para a direita, onde cada token só pode atender ao seu contexto anterior. Essas arquiteturas não são ideais para tarefas no nível de frase, como reconhecimento de entidade nomeada e análise de sentimento, onde é essencial incorporar informações de ambas as direções.

Já o BERT rompe um pouco com abordagens como ELMo e GPT, ele emprega um objetivo de modelagem de linguagem mascarada (MLM), onde certos *tokens* em uma

sequência de entrada são mascarados aleatoriamente, e a tarefa é prever essas posições mascaradas com base na sequência corrompida como entrada. O BERT emprega um codificador Transformer durante o pré-treinamento para considerar contextos bidirecionais. Além disso, o BERT utiliza um objetivo de previsão de próxima frase (NSP). O objetivo do NSP visa aprimorar tarefas que exigem raciocínio entre pares de sentenças, como resposta a perguntas e inferência em linguagem natural.

Assim como o GPT, o BERT emprega *tokens* especiais para criar uma única sequência contígua para cada entrada. Especificamente, o primeiro token é consistentemente um token de classificação especial [CLS], e os pares de frases são separados usando um token especial [SEP]. O BERT segue uma abordagem de pré-treinamento seguida de ajuste fino. O estado oculto final de [CLS] é usado para tarefas no nível da frase, enquanto o estado oculto final de cada token é utilizado para tarefas no nível do token. Contudo, como acontece com o GPT2, a melhoria do BERT em relação ao ELMo e GPT ainda demanda maior explicação e busca das razões. Isto pode estar relacionado ao uso de diferentes objetivos, conjuntos de dados e arquiteturas em comparação com métodos anteriores.

Em síntese, *embeddings* contextuais são normalmente derivados de modelos baseados em Transformer. Esses *embeddings* são gerados inserindo a frase inteira em um modelo pré-treinado. Neste ponto, vale ressaltar que embora um vocabulário de palavras tenha sido criado, este vocabulário não abrange incorporações contextuais. As *embeddings* geradas para cada palavra sofrem interferência na presença de outras palavras na mesma frase. É essa interação com outras palavras em uma determinada frase que chamamos de contexto. Os modelos baseados em Transformer operam principalmente com mecanismos de atenção ativos, onde a atenção serve como um mecanismo para compreender as relações entre uma palavra e suas palavras vizinhas. Consequentemente, ao considerar uma palavra específica, suas *embeddings* não são estáticas, mas geradas dinamicamente por meio de um modelo pré-treinado ou refinado.

3 Trabalhos Relacionados

Esta seção apresenta um panorama de abordagens e trabalhos relevantes na área de classificação de textos, com foco na detecção de discurso de ódio e toxicidade em redes sociais. Apesar da natureza dinâmica do campo, utilizou-se de ferramentas como Google Scholar, Portal de Teses da CAPES, biblioteca SOL da SBC, Elicit, Scite, Consensus, ResearchBuddy, Sourcely, e Scispace para selecionar trabalhos que demonstram a diversidade de abordagens propostas para a detecção de ódio e toxicidade online.

A grande maioria dos estudos recentes em PLN e análise de discurso (discurso de ódio, toxicidade, *bullying*, etc.) implementam uma ou mais técnicas de machine learning. As mais comuns e ainda consideradas dentro do "hype" do PLN, são: *Convolutional Neural Networks* (CNNs), *Recurrent Neural Networks* (RNNs), *Autoencoders*, *Encoder-decoder sequence-to-sequence* (Seq2Seq), Mecanismos de Atenção, Transformer, Modelos de Linguagem Pré-treinados (MLPTs) [1], *Compressive Transformers* e *Long-Range Transformer* (RAE et al., 2019; QIN; FENG; DURME, 2022), entretanto também pode-se acrescentar as Redes de Grafos Convolucionais (BÖLÜCÜ; CANBAY, 2021; PANDEY, 2023; MISHRA et al., 2019; DUONG; ZHANG; LU, 2022), outros métodos de aprendizagem de máquina baseada em grafos para a tarefa de detecção de ódio e toxicidade (JIANG et al., 2022; BEATTY, 2020; SARRACÉN; ROSSO, 2022; GHOSH et al., 2023), e métodos de *Tabular Deep Learning* (TDL) como o TabNet (YOUNUS; QURESHI, 2022; CHOPRA et al., 2023).

3.1 Ódio e Toxicidade — abordagens para detecção

Dentre as primeiras pesquisas acerca do discurso de ódio e discurso ofensivo os trabalhos de Schmidt e Wiegand (2017), Waseem e Hovy (2016), Burnap e Williams (2016), Nobata et al. (2016), Pelle e Moreira (2017) de certa forma encabeçam este tema e com a partir deles muitos outros trabalhos se iniciam com o intuito de definir e analisar as diversas formas de discurso ofensivo ou tóxico em diversas línguas.

Em seu trabalho Schmidt e Wiegand (2017) apresentam de maneira bastante concisa e fornecendo uma visão resumida de 22 estudos no domínio do discurso de ódio, onde se consideraram também outras formas de linguagem ofensiva, como o cyberbullying, como discurso de ódio — diferindo da abordagem muito singular de Fortuna e Nunes (2018), que diferenciaram o discurso de ódio dos demais tipos de ofensas.

Waseem e Hovy (2016)apresentaram um conjunto de dados composto por cerca

[1] A Complete Guide to Natural Language Processing

de 16.000 tweets, com 3.383 anotados como contendo conteúdo sexista e 1.972 como contendo conteúdo racista e uma abordaram a detecção de discurso de ódio em microtextos usando uma abordagem multiclasses, respectivamente. Waseem e Hovy (2016) conduziram experimentos usando regressão logística para classificar se o conteúdo do tweet continha discurso de ódio ou não. O classificador pode chegar ao seu melhor resultado, com medida F1 de 73,93%, ao combinar N-gramas de caracteres com o gênero do usuário que postou o tweet.

Burnap e Williams (2016) tiveram como objetivo identificar textos preconceituosos em relação à raça, orientação sexual e deficiência física. O conjunto de dados utilizado foi criado pelos autores e compreendem tweets relacionados a dois eventos específicos: a reeleição de Barack Obama e o anúncio público de Jason Collins como o primeiro atleta assumidamente gay em uma equipe esportiva profissional dos EUA, e geraram inicialmente conjuntos de 300k a 1.2M de comentários, onde dois mil comentários foram aleatoriamente selecionados para receber anotação manual. Os autores trabalharam com a ideia da existência de classes, alvo do discurso de ódio, que pode haver interseções entre classes, dependências. A ideia de usar dependências digitadas em tarefas de detecção de discurso de ódio originou-se de trabalhos anteriores dos autores. Neste estudo, o objetivo foi testar a eficácia desta abordagem no caso de múltiplas classes-alvo. Os resultados mostraram que o uso de n-gramas de palavras em combinação com n-gramas de dependência digitados e termos odiosos produziu resultados para as três classes estudadas, deficiência, raça e orientação sexual, de 51% a 73%, utilizando como métrica o F1.

Nobata et al. (2016) desenvolveram uma abordagem supervisionada para detecção de linguagem abusiva utlizando o modelo de regressão do Vowpal Wabbit[2], que engloba discurso de ódio, palavrões e linguagem depreciativa, utilizando técnicas de processamento de linguagem natural. Os autores usaram n-gramas, recursos linguísticos como o número de tokens em comentários e sinais de pontuação, recursos sintáticos, uso de *Tags*, relações de dependência e recursos semânticos como incorporação de palavras (Word2Vec). O modelo de regressão obteve desempenho de 78.3% utilizando todas as *features* para o *dataset* WWW2015 (DJURIC et al., 2015), e para o dataset criado para avaliar o modelo obteve 82.6%, também F1-score.

Vigna et al. (2017) trataram a detecção de discurso de ódio em textos de redes sociais do Facebook, que foram coletados a partir de 99 postagens reunindo cerca de 17.000 comentários. Dois classificadores foram utilizados na experimentação: um SVM e uma rede neural LSTM. Além de recursos léxicos para identificar a polaridade dos sentimentos nos textos, foram utilizadas como características o número de tokens, n-gramas ao nível de caracteres, palavras e lemas. No classificador LSTM, foram utilizados a polaridade de cada palavra e léxicos. Os experimentos realizados contemplam uma abordagem multi-classe

[2] Vowpal Wabbit

com níveis de discurso de ódio nos textos e outra, binária com SVM e LSTM, onde a acurária varia de 72.95% a 75.23% no SVM e de 79.81% a 80.60% para o LSTM. O que corresponde após avaliação com as demais métricas a 72% para o SVM e 79% para o LSTM (VIGNA et al., 2017).

Bourgonje et al. (2018a) abordaram o discurso de ódio sobre a ótica de mais de um idioma (inglês e alemão). Eles usaram três conjuntos de dados de texto disponíveis publicamente: dois do X (um em inglês e um em alemão) e um da Wikipedia (em inglês). O estudo comparou diversas técnicas de detecção de discurso de ódio, como Regressão Logística e Entropia Máxima. Para todos os conjuntos de dados, foram utilizados apenas *Bag-of-Words*, que os autores consideraram equivalentes a unigramas. Foram testados cinco classificadores diferentes, onde Naive-Bayes (NB) e Regressão Logística (RL) se destacaram, atingindo F1-score e acurácia no dataset de tweets em inglês 80.06% e 85.67% para NB, e 70.93% e 80.21% para RL, respectivamente.

Anagnostou, Mollas e Tsoumakas (2018) direcionaram sua pesquisa na detecção automática de discurso de ódio em comentários do YouTube. A abordagem dos autores baseou-se em um mecanismo de classificação binária supervisionada que utilizou o SVM para classificar comentários como odiosos ou não, com TF-IDF e Bag-of-Words como *features* obtendo F1-score de 65%.

Já Zhang, Robinson e Tepper (2018) trataram a detecção de discurso de ódio em microtextos no tocante à temática de refugiados na Europa. Os autores utilizaram duas redes neurais para realizar tal tarefa — uma CNN e uma GRU —, além de *embeddings* do corpus obtido do Google News. Foram utilizados diversos corpora disponibilizados por outros autores e foram feitas comparações com o uso de técnicas, como SVM e a técnica criada, demonstrando que a técnica desenvolvida pelos autores obteve uma medida F1 superior às demais.

A abordagem de Pitsilis, Ramampiaro e Langseth (2018) para detecção de discurso de ódio baseia-se no uso de múltiplas redes LSTM para fazer a classificação. No pré-processamento houve uma vetorizaram dos tweets utilizando tokenização, padding e indexação. Como características, foram usados o histórico de postagens dos usuários, chamado de tendência acerca de cada classe, TF-IDF e os vetores dos tweets. Os autores utilizaram a base de dados disponibilizada por Waseem e Hovy (2016).

O trabalho de Bispo (2018) propôs criar um classificador cross-lingual, ou seja, treinar em outro conjunto traduzido do inglês para português e verificar qual seria o desempenho desse classificador no conjunto de dados em português. Nessa abordagem, o classificador alcançou uma precisão de 72%. O classificador, baseado em LSTM, foi testado em diversos cenários, inclusive utilizando GBDTs para melhora de suas métricas e considerando como conjuntos de *features* TF-IDF, N-gramas, *Bag-of-Words* e GloVe 100 dimensões. A base de dados, coletada e criada por Bispo (2018) (CLHS-Bispo2018), foi

composta de diversas fontes (X, Facebook, YouTube, portal de notícias G1 e Stormfront).

Silva e Serapiao (2018) apresentam em sua pesquisa uma arquitetura CNN para identificar discurso de ódio, que utilizada o corpus OffComBR(PELLE; MOREIRA, 2017) e o HLPHSD(FORTUNA et al., 2019) como conjunto de dados e uso de *embeddings* pré-treinadas como Wang2Vec e GloVe cujas dimensões eram 50, 100 e 300. Dentre os experimentos realizados, o modelo obteve os seguintes resultados para *F-score* e Acurácia: 89% e 82,64% para OffComBr-2, otimizador Adam e Wang2Vec 100 dimensões; 96% e 92,74% para OffComBr-3, também utilizando Adam como otimizador e Wang2Vec 300 dimensões; e 96% e 92,74% HLPHSD, aplicando RMSprop como otimizador e Wang2Vec 100 dimensões.

Frenda et al. (2019) retratam o discurso de ódio direcionado a mulheres em sua pesquisa. O objetivo era investigar analogias e diferenças entre sexismo e misoginia a partir de uma perspectiva computacional e afirmaram ser os primeiros a detectar ambos os conceitos usando a mesma abordagem. Utilizando dados do IberEval 2018 e Evalita 2018 (ambos sobre misoginia) e dados de Waseem e Hovy (2016) sobre sexismo, os autores descobriram que os tweets rotulados como sexistas ou misóginos continham mais pronomes femininos do que os tweets rotulados como neutros. Para classificação do texto, os autores realizaram pré-processamento (remoção de emoticons, URLs, símbolos e lematização), usaram caracteres e palavras n-gramas, TF-IDF e empregaram SVM como classificador.

Chowdhury et al. (2019) referiram-se ao discurso de ódio em árabe, apesar de não especificar o alvo específico, os autores relataram que a abordagem foi voltada para intolerância religiosa. Aqui foram coletados cerca de 5.000 tweets, dos quais 42% foram rotulados como tendo discurso de ódio e os 58% restantes como neutros. Os autores alegam ser os primeiros a utilizarem vetores de palavras em conjunto com vetores de grafos (mais especificamente, de retweets e seguidores) dos usuários que postaram as mensagens odiosas. Foi utilizada uma combinação das redes LSTM + CNN + NODE2VEC como abordagem de classificação denominada ARHNet. Assim, os resultados superaram soluções existentes na língua árabe, que usavam GRU, LSTM, SVM e *self-attention* em 2%.

Fortuna, Soler-Company e Wanner (2021) desenvolveram uma abordagem para detecção de discurso de ódio utilizando diferentes corpora em inglês. Eles padronizaram as classes presentes nos corpora para ser possível a utilização de forma intercambiável. Para classificação, foram utilizados os modelos ALBERT, BERT, fastText e SVM, e para o SVM foi utilizado BoW como conjunto de *features*. Os resultados mostraram como uma abordagem *cross-dataset* pode ser usada para detectar semelhança entre corpora e categorias, e ajudar a criar uma categorização uniforme de corpora. Seus experimentos reuniram obtiveram 136 modelos performando com F1 acima de 70%.

Arco et al. (2021) utilizaram bases de dados em espanhol e compararam diversos modelos, tanto de Aprendizado de Máquina tradicional quanto de Aprendizado de Máquina

Profundo, quanto ao desempenho na tarefa de detecção de discurso de ódio em textos. As bases de dados utilizadas consistem em textos extraídos do X; uma delas contendo 6.000 tweets obtidos pelo HaterNet — um sistema inteligente utilizado pelo Escritório Nacional Espanhol contra Crimes de Ódio da Secretaria de Estado de Segurança da Espanha —, e a outra contendo 1.600 tweets obtidos do desafio HatEval 2019. Os autores concluíram que os modelos que utilizam aprendizado profundo alcançaram melhores resultados que os de aprendizado de máquina tradicional. Além disso, os que utilizam transferência de aprendizagem (os quais os autores colocaram como uma categoria à parte) se sobressaíram dentre os modelos que usam aprendizado profundo.

Karim et al. (2020) posicionaram sua abordagem para detectar o discurso de ódio em bengali utilizando Modelos de Linguagem Pré-treinados (MLPTs). Foram experimentadas diversas técnicas, desde Naive Bayes e Regressão Logística, até CNN e MLPTs. Os melhores resultados foram obtidos com uma combinação de MLPTs (BERT em Bengali, XLM-R e BERT multilíngue). Além de realizarem classificação binária, a abordagem desenvolvida pelos autores também se mostrou apta para realizar classificação multi-classe, conseguindo os melhores resultados no estado da arte para o idioma bengali.

Soto et al. (2022) experimentaram o uso de diferentes *embeddings*, com dimensões distintas e utilizaram uma rede CNN para realizar a classificação dos textos. Os autores usaram *embeddings* específicos gerados para o Hierarchically-Labeled-Portuguese-Hate-Speech-Dataset (HLPHSD) e *embeddings* obtidos do NILC. Foram testadas as representações de *embeddings* wang2vec, word2vec, fastText e Glove. Este estudo trouxe importantes *insights* para a tarefa de detecção de ódio, e por extensão a de toxicidade, usando como base de dados os dataset de Pelle e Moreira (2017) e Fortuna (2017). Nele exploram-se as vantagens do uso de pequenos conjuntos de dados para desenvolver melhores e mais eficientes *embeddings* contextuais. Neste trabalho o uso do GloVe se mostrou o melhor método de embedding do Núcleo Interinstitucional de Linguística Computacional (NILC) [3] e o fastText o melhor método para *embeddings* específicas de conjuntos de dados. Essas constatações possibilitaram que as *embeddings* específicas produzissem um melhor ajuste ao conjunto, o que promoveu um treinamento mais curto e capaz de obter melhores precisões. O melhor resultado obtido com a representação Glove com 300 dimensões, com os *embeddings* NILC. No entanto, concluiu que os resultados obtidos indicam que as *embeddings* do NILC levariam a um ajuste igualmente eficiente para a tarefa de detecção de ódio e uso de linguagem ofensiva.

[3] Repositório de Word Embeddings do NILC

Diferentemente das abordagens e trabalhos anteriores, o trabalho de Saraiva et al. (2021), introduz uma abordagem semi-supervisionada para detecção de comentários tóxicos, cuja base de dados fora o ToLD-Br Leite et al. (2020) em sua versão binária. A abordagem ataca o problema ao modelar uma Rede de Grafos Heterogêneos (*Heterogeneous Graph Network - HGN*), para incorporar as características do texto mediante *embeddings* estáticas (GloVe), para então submeter essas informações para um classificador binário, o Gradiente Boost Classifier, utilizando o LGC como método de transdução.

A Tabela 1, lista e resume um pouco os trabalhos e abordagens mencionados até então. Alguns deles merecem maior atenção por suas abordagens investigarem o ódio e toxicidade nas redes sociais para o português brasileiro, são eles os trabalhos de (SILVA; SERAPIAO, 2018), Bispo (2018), Saraiva et al. (2021), (FORTUNA; SOLER-COMPANY; WANNER, 2021) e Soto et al. (2022).

Tabela 1 – Trabalhos relacionados à Detecção Discurso de Ódio e Toxicidade.
Legenda: F_1 : F-score, A: Acurácia e P: Precisão.

Trabalho	***Features***	**Classificador**	**Base de Dados**	**Métricas**
Waseem e Hovy (2016)	*N-gramas*	Regressão Logística	X	73.89% (F_1)
Burnap e Williams (2016)	Dependências tipadas, BoW e termos odiosos	SVM	X	73% (F_1)
Nobata et al. (2016)	*N-gramas, número de tokens, classificação gramatical, embeddings*	Regressão de Vowpal Wabbit	X	82.6% (F_1)
Vigna et al. (2017)	*Léxicos + Word2Vec*	SVM e LSTM	Facebook	72% e 79% (A)
Silva e Serapiao (2018)	*WE pré-treinadas*	CNN	OffComBr e HLPHSD	96% (F_1) e 92% (A)
Bispo (2018)	*TF-IDF, N-gramas, BoW e GloVe*	LSTM	CLHS-Bispo2018	72% (P)
Bourgonje et al. (2018b)	BoW	NB e RL	X e Wikipédia	85.6% e 80.2% (A)
Anagnostou, Mollas e Tsoumakas (2018)	*BoW e TF-IDF*	SVM	YouTube	65% (F_1)
Zhang, Robinson e Tepper (2018)	*Embeddings*	CNN + GRU	X	94% (F_1)
Pitsilis, Ramampiaro e Langseth (2018)	*TF-IDF*	*Ensemble LSTM*	X	93% (F_1)
Frenda et al. (2019)	*N-gramas, BoW, TF-IDF e lexicons*	SVM	X	73.9% (F_1)
Chowdhury et al. (2019)	*Embeddings de palavras e grafos*	ARHNet	X	78% (F_1)
Saraiva et al. (2021)	*GloVe*	HGN + XGBoost	ToLD-Br	73% (F_1)
Fortuna, Soler-Company e Wanner (2021)	*Embeddings contextuais e BoW*	MLPTs e SVM	Variada	$\geq$ 70% (F_1)
Arco et al. (2021)	*TF-IDF e WE*	LSTM, CNN, BiLSTM, SVM, RL, BERT, BETO e XLM	X	77% (F_1)
Soto et al. (2022)	*WE*	CNN	X	92% (F_1) e 86% (A)
Younus e Qureshi (2022)	*ByT5*	TabNet	Gab + X	46% (F_1)
Oliveira, Reis e Ebecken (2023)	*BERT Embeddings*	BERTimbau	X + Instagram	89% (F_1)

Outros trabalhos publicados nos últimos três anos merecem, para mais da menção, seu espaço neste capítulo e também na Tabela 1, justamente por introduzirem e implementarem modelagens para o problema de detecção das formas de toxicidade e ódio nas redes sociais de maneira disruptiva, e com a aplicação e aprimoramento de modelos de aprendizagem profunda, que pode ser dita aqui como não convencional, a aprendizagem profunda tabular ou com dados tabulares (*Tabular Deep Learning* — TDL). São eles, os trabalhos de Younus e Qureshi (2022) e Chopra et al. (2023)[4] que desenvolvem *frameworks* para detecção de sexismo e ódio, utilizando como modelo principal para classificação e detecção o TabNet, modelo de TDL, além do uso de modelos pré-treinados de *embeddings*, como BERT e o ByT5.

Referenciando também os trabalhos publicados recentemente, mas que não podem compor a Tabela 1, merecem também bastante destaque e a atenção dos leitores por oferecerem grande conhecimento sobre o uso e as capacidades dos LLMs na tarefa de detecção de ódio e toxicidade como em Oliveira et al. (2024), Assis et al. (2024), Junqueira et al. (2023a), Oliveira et al. (2023), de outras temáticas relativas ao NLP, no desenvolvimento de ferramentas mais robustas de análise, detecção e combate ao ódio como o recente trabalho de Vargas et al. (2023), e por fim a apresentação de novos LLMs para Língua Portuguesa como o Sabiá (PIRES et al., 2023), o Glória (LOPES; MAGALHÃES; SEMEDO, 2024) e o Sabiá-2 (ALMEIDA et al., 2024).

3.1.1 Datasets

Infelizmente o Português-BR não é uma linguagem onde há uma vasta disponibilidade de recursos (abordagens, datasets, modelos, etc.) para desenvolver as mais variadas pesquisas no campo da PLN e Aprendizado de Máquina sobre a problemática do Ódio e Toxicidade Online, como no inglês de maneira geral, podendo até em alguns aspectos ser considerada uma *low-resource language*, contudo essa tem se transformado gradualmente nos últimos anos. Alguns trabalhos, mencionados mais adiante, tomaram a iniciativa para dirimir ou suavizar esse problema e oferecem grande ferramental para o campo de estudo.

Uma das grandes necessidades de recursos para o Português-BR é a criação e disponibilização de datasets curados. Essa necessidade é só uma ponta do problema geral dos recursos textuais relativos a conjuntos de dados, o qual é a curadoria de dados, cabendo mencionar que a anotação e revisão de dados é outro ponto bastante crítico que afeta não só a comunidade científica na totalidade, mas interfere diretamente na qualidade dos modelos desenvolvidos e nas análises feitas sobre os mais variados fenômenos.

Atualmente, a Língua Portuguese dispõe de alguns conjuntos de dados considerados adequados e bem curados, o que ainda não a desqualifica como uma *low-resource language*,

[4] Trabalho Chopra et al. (2023) não incluído na Tabela 1 pois o artigo completo não está disponível gratuitamente para leitura na biblioteca digital da ACM.

já que a diversidade e quantidade de recursos é restrita. Para a tarefa de detecção de discurso de ódio e toxicidade, uma checagem adicional na qualidade dos dados e da anotação em si se faz necessária, havendo a ocorrência de anotações equivocadas em relação ao comentário pelo uso de termos específicos e outrora considerados ofensivos ou inadequados. Deve-se lembrar que na maioria dos modelos o aprendizado e a generalização do conhecimento acontece partir do que lhes é apresentado como informações e características. Nesse sentido pode-se elencar como bons e excelentes conjuntos de dados: OffComBR[5] (PELLE; MOREIRA, 2017) Hierarchically-Labeled-Portuguese-Hate-Speech-Dataset (HLPHSD)[6] (FORTUNA et al., 2019), ToLD-Br (LEITE et al., 2020)[7], HateBR [8] (VARGAS et al., 2022), OLID_BR [9] (TRAJANO; BORDINI; VIEIRA, 2023) e o TuPyE-Dataset[10] (OLIVEIRA; REIS; EBECKEN, 2023).

A seguir, na Tabela 2, são apresentados alguns conjuntos de dados existentes até o momento para as tarefas de discurso de ódio e discurso ofensivo ou tóxico em Português-BR.

Tabela 2 – Discurso Tóxico e de Ódio: corpora disponíveis em Português-BR.

Autor	Corpus	Tamanho	Fonte
Pelle e Moreira (2017)	OffComBR	10k	Portal de notícias
Fortuna et al. (2019)	HLPHSD	5.5k	X
Leite et al. (2020)	ToLD-Br	21k	X
Vargas et al. (2022)	HateBR	7k	Instagram
Trajano, Bordini e Vieira (2023)	OLID_BR	7.9k (13.5k)	YouTube, X e outros datasets.
Oliveira, Reis e Ebecken (2023)	TuPyE-Dataset	43k	X e Instagram
Lima, Pagano e Silva (2024)	redditbrtoxicitydataset	2.5k	Reddit
Miranda et al. (2024)	SoccerFansFB	276k	Facebook

A tabela estrutura-se da seguinte forma: Autor, quem propôs-criou o dataset; Corpus, o nome dado ao dataset pelo próprio autor ou nome usado neste trabalho no caso dos datasets do Reddit criado por Lima, Pagano e Silva (2024) e do Facebook criado por Miranda et al. (2024) para facilitar sua referência; Tamanho, que descreve o tamanho do dataset em número de linhas (exemplo: 1k = 1000 linhas); e Fonte, que explicita a fonte dos dados, se portal, blog, vídeo blog ou rede social.

[5] OffComBR
[6] Hierarchically-Labeled-Portuguese-Hate-Speech-Dataset
[7] Toxic Language Dataset for Brazilian Portuguese (TolD-Br)
[8] HateBR - Offensive Language and Hate Speech Dataset in Brazilian Portuguese
[9] OLID-BR
[10] Silly-Machine | TuPyE-Dataset: Brazilian Hate Speech Dataset (TuPy)

4 LLMeFT-Transformer

Este capítulo apresenta a proposta metodológica de Aprendizagem Profunda Tabular (APL) ou *Tabular Deep Learning* (TDL) para a classificação binária de comentários tóxicos e de ódio em redes sociais, denominada LLMeFT-Transformer (*Large Language Model embeddings with FT-Transformer classifier*). A abordagem desenvolvida aplica o modelo FT-Transformer de Gorishniy et al. (2021) e o estende utilizando uma arquitetura para aprendizagem profunda tabular (HU et al., 2024) e aplicando diversos modelos de linguagem (incluindo LLMs) como geradores de *features* textuais para a classificação de textos.

A Figura 7 apresenta o *pipeline* implementado para desenvolvimento da abordagem supervisionada LLMeFT-Transformer e obtenção dos objetivos desta pesquisa.

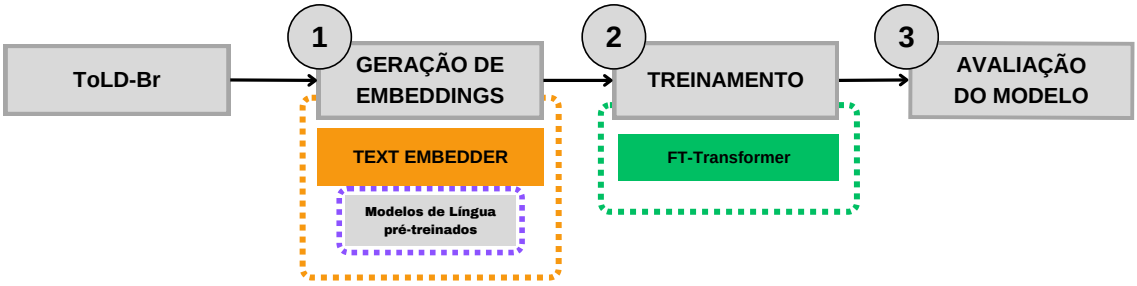


Figura 7 – Pipeline da abordagem LLMeFT-Transformer (fonte: o autor).

O desenvolvimento do LLMeFT-Transformer compreende majoritariamente três etapas: o processamento textual e a geração de *features* textuais, que convertem e transformam o texto em *embeddings* aplicando o modelo pré-treinado selecionado; o treinamento do modelo e por último a avaliação dos resultados do modelo.

As seções seguintes apresentam e exploram os seguintes detalhes da abordagem: i) a base de dados utilizada para treinamento e avaliação do modelo; ii) uma breve apresentação do modelo base FT-Transformer, incluindo uma explanação dos seus componentes o *Feature Tokenizer*, o *Transformer* e o LLMeFT-transformer. Aqui também explora-se a integração desses componentes com as *embeddings* geradas; iii) os experimentos conduzidos e os detalhes de implementação; e iv) algumas considerações gerais.

4.1 Base de Dados

A base de dados (córpus) utilizada para validar a abordagem proposta, LLMeFT-Transformer, foi o ToLD-Br (Toxic Language Dataset for Brazilian Portuguese) (LEITE et al., 2020) que é um conjunto de comentários textuais extraídos do X, antigo Twitter,

para análise e detecção de linguagem tóxica reunindo cerca de 21 mil tweets anotados manualmente em sete categorias: não-tóxico, LGBTQ+fobia, obsceno, insulto, racismo, misoginia e xenofobia.

Esta base de dados foi escolhida por reunir comentários textuais em português e ser anotada considerando diferentes aspectos acerca da toxicidade e ódio. Os comentários foram anotados por três avaliadores diferentes e para que o tweet fosse considerado tóxico, pelo menos um anotador teria que classificá-lo como tal. E segundo esse critério de toxicidade o ToLD-Br fica caracterizado conforme as Tabelas 3 e 4 a seguir.

Tabela 3 – Características do ToLD-Br - Versão Binária

Categoria	**Nº de Tweets**
Tóxico	1.490
Não-Tóxico	19.510

Tabela 4 – Características do ToLD-Br - Versão Multiclasse

Categoria	**Nº de Tweets**
LGBTQ+fobia	344
Obsceno	6.652
Insulto	4.385
Racismo	138
Misoginia	463
Xenofobia	151

O ToLD-Br é disponibilizado em duas versões permitindo sua utilidade e reprodutibilidade em diferentes abordagens de pesquisa conforme as tabelas anteriores, onde:

- **Versão Binária:** Um tweet é rotulado como "tóxico"se pelo menos um anotador o marcou como tal, caso contrário, é "não tóxico";
- **Versão Multiclasse:** Considera cada uma das sete categorias de forma independente (não tóxico, LGBTQ+fobia, obsceno, insulto, racismo, misoginia e xenofobia).

O ToLD-Br é liberado com todas as três anotações para cada tweet, o que possibilita trabalhar com diferentes níveis de concordância entre anotadores. Para garantir consistência e facilitar a comparação direta tanto com os experimentos desta abordagem quanto de outros trabalhos de detecção de ódio e toxicidade foram utilizados os *splits* pré-definidos por (LEITE et al., 2020) em que 80% dos dados sejam utilizados para treinamento, 10% para validação e 10% para teste, o que se alinha também com o experimento original de validação do próprio ToLD-Br.

Além dos *splits* pré-definidos foi adotado o cenário de concordância entre anotadores menos rigoroso. Neste cenário, o conjunto de dados contém 11.745 comentários não tóxicos e 9.255 comentários tóxicos, com proporção na distribuição de classes de 1:1,2, o que faz com que o ToLD-Br ainda possa ser considerado balanceado. Este cenário foi mantido para execução dos experimentos e avaliação dos resultados da abordagem proposta e oferecer uma comparação justa entre as demais abordagens de detecção de ódio e toxicidade que utilizam o ToLD-Br como base de dados.

4.2 FT-Transformer

O FT-Transformer é um modelo de APT apresentado em 2021 por Gorishniy et al. (2021) onde as características categóricas e numéricas presentes num *dataset* tabular são transformadas em *embeddings* por um componente chamado *Feature Tokenizer*, que conforme a especificação prévia de cada tipo de dado em cada coluna antes de aplicá-los a uma pilha de camadas Transformer, que operam no nível de características de cada objeto presente no conjunto de *embeddings* gerados. O FT-Transformer base consiste em dois componentes principais: o Feature Tokenizer e o Transformer.

4.2.1 Feature Tokenizer

O *Feature Tokenizer* desempenha um papel crucial na preparação dos dados para o Transformer, convertendo características heterogêneas, como numéricas e categóricas, em representações numéricas densas (*embeddings*). Isso é essencial para que o Transformer, originalmente projetado para dados sequenciais, possa processar de maneira eficaz dados tabulares, capturando as interações entre diferentes tipos de características.

Com base em Luthy (2023), Gorishniy et al. (2021), o *Feature Tokenizer* transforma as características numéricas por meio de uma multiplicação matricial entre os valores de *features* e uma matriz de pesos de aprendizagem $W^{(num)}$. Essa operação produz *embeddings* numéricas que capturam informações semânticas das características numéricas. Essas *embeddings* são então combinadas com um bias $b^{(num)}$, obtendo-se a resultante T que representa a transformação final dos *embeddings*. A Figura 8 ilustra esse processo.

4.2.2 Transformer

O modelo base do FT-Transformer emprega um mecanismo de atenção popular conhecido como atenção *multihead*[1] introduzido por Vaswani et al. (2017) e que demonstrou forte desempenho em diversas aplicações. Este mecanismo é uma extensão do mecanismo de autoatenção, que calcula uma soma ponderada de incorporações de entrada para criar uma

[1] Consultar Figura 5 da seção 2.2, capítulo 2

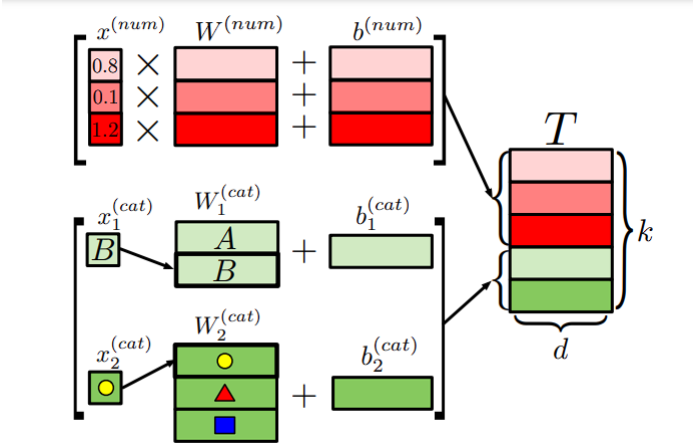


Figura 8 – Feature Tokenizer (FT) porposto por Gorishniy et al. (2021).

representação que é sensível ao contexto, ou seja, maior sensibilidade as nuances contidas nos dados. Neste mecanismo de autoatenção, são empregadas três camadas lineares, a *query* que serve como entrada que coloca uma questão ou busca informações específicas, a *key* que proporciona o contexto ou informações relevantes para auxílio na compreensão da busca, atuando como referência para comparação e estabelecendo conexões entre diferentes elementos na entrada, e o *value* que detém as informações ou o conteúdo real que o modelo se concentra com base na *query* e na *key*, compreendendo os detalhes que são atendidos e complementando para a representação final resultante (VASWANI et al., 2017; USZKOREIT, 2017; SHAW; USZKOREIT; VASWANI, 2018; WANG et al., 2020a).

O componente de *Multi-Head Self-Attention* permitiu a codificação de vários relacionamentos e nuances para cada palavra, elemento ou sentença completa na sequência de entrada, tornando um componente bastante útil para o modelo. Tal utilidade estende-se e permite aprimorar a capacidade de capturar padrões complexos, extração de informações significativas e geração de representações delicadas ao contexto. O paralelismo faz com que o mecanismo de atenção se torne mais eficiente, mantendo suas capacidades de atenção mais poderosas e duradouras. Para que houvesse um melhor desempenho Gorishniy et al. (2021) propôs a utilização de uma normalização variante da PreNorm proposta por Wang et al. (2019) anterior aos mecanismos de *Multi-Head Self-Attention* e *Feed Forward* para que houvesse uma otimização facilitada, e removendo a primeira normalização do primeiro Transformer.

4.2.3 LLMeFT-Transformer

Já no LLMeFT-Transformer, o FT opera especificamente durante uma etapa de Materialização (*Materialization*), sendo utilizado quando há colunas de texto que serão posteriormente processadas por algum modelo de linguagem para gerar sua representação. O princípio central do *Feature Tokenizer* é mapear cada valor de característica para uma representação numérica significativa (*embedding*), atuando como um passo crucial de

pré-processamento que vai além da tokenização convencional. A Figura 9, adaptada de, Hu et al. (2024) ilustra a arquitetura para aprendizado profundo em dados tabulares utilizada paro o LLMeFT-Transformer, destacando o estágio de operação do *Feature Tokenizer*.

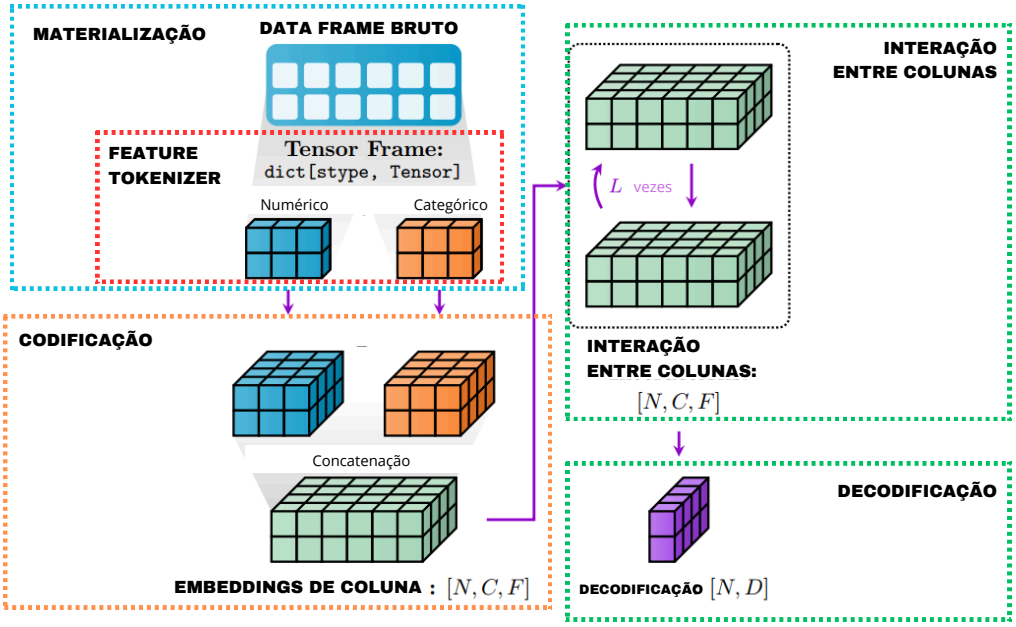


Figura 9 – Arquitetura PyTorch Frame adaptado de Hu et al. (2024).

A Figura 9 também indica os estágios principais do processo de aprendizagem profunda implementado nessa abordagem que antecedem a classificação, executada pelo Transformer. Esses estágios incluem Materialização (*Materialization*) , Codificação (*Column Embeddings*), Interação entre colunas (*Column-wise Interaction*) e Decodificação, cada um com suas especificidades descritas no texto original.

O uso desta arquitetura para aprendizagem profunda tabular, e consequentemente do Feature Tokenizer, oferece várias vantagens em relação a outras abordagens, como destacado por Luthy (2023), Hu et al. (2024). Essas vantagens incluem representação aprimorada das características, processamento eficiente de dados complexos, captura de dependências complexas, suporte a dados multimodais, integração com modelos pré-treinados, modularidade e flexibilidade, eficiência computacional e integração de ponta a ponta com modelos de aprendizado profundo *downstream*.

4.2.4 *Embeddings* Textuais

No Processamento de Linguagem Natural (PLN), a representação eficaz de texto é crucial. *Embeddings* textuais mapeiam palavras, frases ou documentos inteiros em vetores numéricos densos, capturando seu significado semântico e relações contextuais. Essa transformação permite que modelos como o FT-Transformer processem e compreendam informações textuais, habilitando-o para uma variedade de tarefas de PLN, incluindo

classificação, recuperação de informações, análise de similaridade semântica e avaliação de geração de texto.

Para avaliar o desempenho da abordagem em uma tarefa de classificação binária, diferentes métodos de *embeddings* textuais foram investigados, cada um com suas particularidades. Esses métodos incluem *embeddings* de sentença (*Sentence Embeddings*), LLM *Embeddings* (modelos de embeddings modernos como o, OpenAI *text-embedding-3-large* e o VoyageAI *voyage-multilingual-2*) e o uso direto de outros modelos de línguagem (BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) e Albertina PT-BR (RODRIGUES et al., 2023)) como geradores de *embeddings*. A escolha da técnica de *embedding* textual impacta na praticabilidade e no tamanho da representação numérica, afetando potencialmente o tempo de execução da abordagem.

O processo de geração de *embeddings* textuais tipicamente envolve tokenização, conversão em *embeddings* e, em alguns casos, refinamento (*fine-tuning*) com um conjunto de dados específico da tarefa. Dada a importância de avaliar não apenas a eficácia, mas também a eficiência em termos de tempo, usabilidade e facilidade de otimização, a comparação de várias técnicas de *embeddings* textuais é fundamental.

4.2.4.1 *Text Embedder*: A Interface entre Texto e o Modelo

O *Text Embedder* (TE) é responsável por integrar os *embeddings* textuais ao FT-Transformer, atuando como um codificador que converte texto em vetores numéricos utilizando modelos de *embeddings* ou modelos de linguagem pré-treinados. No LLMeFT-Transformer, o Text Embedder opera entre os estágios de Materialização e Codificação ilustrados na Figura 9, não há qualquer processamento textual ou geração de *embeddings* na versão base do FT-Transformer, assim como de seu *Feature Tokenizer*.

O TE, também conhecido como Incorporador Textual ou Incorporador de Características Textuais, é um mecanismo que anexa texto pré-processado a um vetor de características de alta dimensão, representando seu significado semântico e dependências. No contexto desta pesquisa, o TE permite o uso de modelos de *embeddings* textuais pré-existentes, modelos de linguagem pré-treinados (como BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) e Albertina PT-BR (RODRIGUES et al., 2023)) e LLMs como *Language Embedders* [2].

A escolha do método de embedding ideal é crítica, uma vez que as *embeddings* resultantes podem influenciar significativamente o resultado geral da abordagem. Fatores influenciadores incluem a tarefa e o domínio, o tamanho do modelo e eficiência computacional, e a disponibilidade de dados para treinamento e *fine-tuning*, com base nisso para essa abordagem os modelos listados na Tabela 5 foram selecionados para compor o TE.

[2] Para mais detalhes sobre a implementação e utilização de Text Embedders acessar o repositório no GitHub - Tabular-Transformer-Toxic2024.

Tabela 5 – Lista de Modelos de Embeddings.

Modelo	**Suporte**	**Tipo de Modelo**	**Dimensão de saída**
BERTimbau	Monolingual	BERT-based	1024
AlbertinaPTBR	Monolingual	BERT-based	1536
SBERTimbau	Monolingual	SBERT	1024
ME5Large	Multilingual	SBERT	1024
DeBERTaV2XL	Multilingual	BERT-based	1536
VoyageLarge2	Multilingual	LLMem	1536
OpenAI-TE3-large	Multilingual	LLMem	1536

4.3 Considerações Finais

A proposta metodológica apresentada neste capítulo contribui para a área de processamento de linguagem natural e análise de dados de redes sociais, ao abordar a classificação de comentários tóxicos com uma nova perspectiva de Aprendizagem Profunda Tabular, LLMeFT-Transformer.

Para que a comparação com outras abordagens se fizesse minimamente adequada o cenário de anotação descrito nas demais abordagens, no experimento original do ToLD-Br e o uso dos conjuntos pré-definidos foram mantidos.

Deve-se acrescentar que esses parâmetros de comparação (cenário de anotação e uso dos conjuntos pré-definidos) não permitiu uma avaliação mais ética da abordagem. Há pontos que merecem um escrutínio detalhado e ulterior em relação ao ToLD-Br e a possível criação de novas versões que observem e não priorizem a permanência do viés humano, tanto na anotação quanto na descrição e definição das classes Tóxico e Não-Tóxico. No tocante ao conjunto de dados, os resultados dos experimentos assim como a acurácia sugerem também uma necessidade de dados mais representativos no conjunto, o que pode indicar a necessidade do uso de técnicas de *Data Augmentation* e de sintetização de dados com LLMs. Outro ponto que pode oferecer um suporte na preservação dos dados mais representativos, principalmente na classe de interesse (Tóxico) seria o uso de Algoritmos Genéticos juntamente com as técnicas de geração de *embeddings* utilizadas.

Mesmo sem considerar essas hipóteses e possíveis alternativas, a escolha de um novo modelo pré-treinado multilíngua ou para o português (monolíngua) como *encoder* pode contribuir para a obtenção de menores taxas de perda e taxas de acurácia mais consistentes e interessantes para o uso deste tipo de abordagem em tarefas de classificação de textos que envolvem dados tabulares na temática de toxicidade online e discurso de ódio.

5 Experimentos e Resultados

Este capítulo descreve os experimentos e resultados para o LLMeFT-Transformer, abordagem de aprendizagem profunda proposta e implementada para classificação de textos deste trabalho.

5.1 Experimentos

O modelo LLMeFT-Transformer estende o FT-Transformer base alterando seu número de camadas de 4 para 12, a utilização de uma arquitetura para aprendizado profundo tabular (HU et al., 2024) e como processo adicional a modificação da classe base do Text Embedder (Gerador de *Embeddings*), e extensão desta para o processamento e geração de *features* textuais a partir da utilização de outros modelos de linguagem pré-treinados como *Language Embedders*, além das opções como SBERT, Open AI e Voyage AI já disponíveis para uso nos *frameworks* de aprendizagem profunda tabular como o PyTorch Framne (HU et al., 2024).

O processamento de características textuais por meio de diversos modelos de *embeddings*, e em especial modelos específicos para a Língua Portuguesa, garante maior flexibilidade e liberdade de escolha no processamento textual principalmente ao permitir a geração de *embeddings* a partir de modelos refinados para uma língua como o Português e para tarefas que requerem grande especificidade e preservação do dinamismo linguístico como análise do discurso. Outro detalhe implementado foi um *learning rate scheduler*, o que contribuiu para a abordagem obter uma convergência gradual, uma maior estabilidade inicial durante o treinamento e uma redução mais acelerada da perda.

Para verificar se o LLMeFT-Transformer incorpora adequadamente os atributos de entrada da mesma maneira que o FT-Transformer de Gorishniy et al. (2021) e o TEFT-Transformer de Luthy (2023) um experimento prévio foi realizado utilizando os mesmos parâmetros-base da implementação original do FT-Transformer e o dataset utilizado nesta mesma abordagem. Esse experimento averiguou se a nova abordagem pode tratar as características numéricas e categóricas de entrada da mesma forma que o FT-Transformer base e oferecer melhora em seus resultados de acurácia e perda.

Seguindo adiante, o experimento principal e foco deste trabalho pôde ser performado e então avaliar o desempenho do LLMeFT-Transformer, na tarefa de classificação binária de comentários textuais de redes sociais, utilizando como dataset principal o ToLD-Br, particionado em conjuntos de treinamento, teste e validação nas respectivas proporções 80%, 10% e 10%, já mencionado anteriormente. A etapa compreendida pelo treinamento e

avaliação do modelo, é direta, não envolve muitos detalhes sobre a seleção de parâmetros, configuração, definição do modelo e sua instanciação. Contudo, deve-se lembrar que para os tipos textuais a entrada pode ser interpretada como tokens, *embeddings* ou texto sem pre-processamento, cabendo a cada um deles um codificador específico. Nesta abordagem, os parâmetros definidos para o modelo em relação aos seus exemplos disponibilizados[1] sofreram pequenas alterações. Os parâmetros seguem especificados na Tabela 6.

Tabela 6 – Parâmetros do Modelo

Parâmetro	**Valor**
channels	256
out_channels	2
layers	12
learning_rate	0.0001
batch_size	512
loss function	'Cross Entropy'
epochs	100
optimizer	'AdamW'

5.2 Resultados

Inicialmente, a utilização de *embeddings* de sentenças se mostrou intrigante e significativa, principalmente em relação às métricas observadas paralelamente nos experimentos, excluindo as *embeddings* de LLMs. Os resultados obtidos também se mostraram promissores em relação à modelos mais tradicionais (XGBoost, LightGBM e GBDTs) e que até poucos anos atrás dominavam as tarefas de classificação e regressão de dados em tabelas, e em alguns casos as tarefas de PLN.

A Tabela 7 apresenta os resultados obtidos do LLMeFT-Transformer com diversos modelos de *embeddings* utilizados no *Text Embedder* para a base de dados do ToLD-Br. A tabela apresenta o F1-score para as classes Tóxico e Não-Tóxico, assim como a acurária do LLMeFT-Transformer com o respectivo *Text Embedder*. Esses resultados revelam que quando o TE utiliza o OpenAI-TE3-large o LLMeFT-Transformer atinge maior acurácia e pontuação F1 para a classe tóxica e não tóxica, tornando-o a escolha mais robusta dentre as apresentadas.

Os resultados ainda revelam que os modelos monolíngua, BERTimbau e SBERTIMBAU, mostraram forte desempenho entre os modelos monolíngues, com BERTimbau sendo mais assertivo para a classe não tóxica e SBERTimbau com melhor desempenho

[1] PyTorch Frame — Exemplos

Tabela 7 – Resultados com diferentes Text Embedders.

Suporte	Text Embedder	Tóxico	Não-Tóxico	Acurácia
Monolíngue	BERTimbau	0.7101	0.7531	0.7333
	AlbertinaPTBR	0.6826	0.7306	0.6906
	SBERTimbau	0.7158	0.7429	0.7300
Multilíngue	ME5Large	0.7378	0.7258	0.7319
	DeBERTaV2XL	0.6637	0.6974	0.6814
	VoyageLarge2	0.7143	0.7327	0.7238
	OpenAI-TE3-large	**0.7398**	**0.7740**	**0.7580**

As métricas F1-score estão representadas em Tóxico e Não-Tóxico.

para a classe tóxica. De maneira geral, os modelos de *embeddings* multilíngue como OpenAI-TE3-large, ME5Large e VoyageLarge2 apresentam excelente desempenho.

A Tabela 8 a seguir apresenta um comparativo entre diferentes abordagens para detecção de ódio e toxicidade desenvolvidas usando o ToLD-BR como conjunto de dados e compara a pontuação geral de F1-score a acurária (quando disponível) de cada abordagem.

Tabela 8 – Detecção de Ódio e Toxicidade: comparativo entre abordagens com o ToLD-Br.

Work	Abordagem	F1-score	Accurácia
Leite et al. (2020)	M-BERT + *transfer learning*	0.76	-
Saraiva et al. (2021)	GloVe + HGN + XGBoost	0.73	-
Oliveira et al. (2023)	Prompt GPT 3.5	0.73	-
Junqueira et al. (2023b)	Albertina PT-Br Base	0.74	0.78
Junqueira et al. (2023b)	BERTimbau Base	0.88	0.88
Junqueira et al. (2023b)	BERTimbau Large	0.89	0.89
Oliveira et al. (2024)	Prompt Sabiá + 10 few-shots	0.73	-
Oliveira et al. (2024)	Prompt GPT 3.5-turbo + zero-shot	0.74	-
LLMeFT-Transformer	**OpenAI-TE3-large + FTT**	**0.75**	**0.76**
	BERTimbau + FTT	0.73	0.73
	ME5Large + FTT	0.73	0.73

Os resultados mostram que as abordagens com melhor desempenho para detectar ódio e toxicidade envolvem extenso refinamento (fine-tuning), transferência de aprendizagem, ou zero-shot, diretamente utilizando como classificadores o BERTimbau e M-BERT (Multilingual BERT), ou por técnicas de prompt com GPT 3.5-turbo.

No entanto, métodos que utilizaram outras técnicas de representação de dados (grafos heterogêneos), *embeddings* estáticas e modelos de *embeddings* modernos também alcançaram resultados significativos e bastante competitivos mesmo sem ajustes adicionais, permitindo um caminho para melhorias futuras ou implementações mais complexas e híbridas, e isso foi mostrado com o LLMeFT-Transformer, em sua configuração com OpenAI-TE3-large + FTT.

Como recurso subsequente de avaliação para a abordagem proposta nesta pesquisa, apresenta-se a matriz de confusão na Figura 10 abaixo.



Figura 10 – Matriz de confusão para as classes tóxica e não tóxica.

A matriz de confusão revela que o LLMeFT-Transformer apresenta uma capacidade robusta de identificar com precisão comentários não tóxicos, conforme indicado pelo alto número de verdadeiros-negativos. Além disso, o modelo demonstra uma capacidade moderada para detectar comentários tóxicos, com 778 verdadeiros-positivos. No entanto, os resultados também indicam que notáveis erros de classificação, com 301 falsos-positivos e 194 falsos-negativos, sugerindo espaço para melhorias na redução das taxas de classificação incorretas.

5.3 Considerações Finais

Os resultados alcançados até o momento, assim como os erros gerados durante a experimentação e as contribuições potenciais desbravaram um caminho bastante claro e promissor para esta pesquisa na totalidade, porém aos olhos deste autor a abordagem merece um pouco mais de atenção, experimentação, testes e modificações quanto a sua estrutura base e quanto ao uso de outras técnicas e abordagens para gerar, extrair e melhorar *embeddings* como RAG (Retrieval-Augmented Generation) que tende a melhorar a relevância e a qualidade das *embeddings* geradas, expande a compreensão do LLM em relação ao contexto e gera *embeddings* mais informativas e contextualmente mais apropriadas.

6 Conclusões, Limitações e Trabalhos Futuros

Este capítulo apresenta as conclusões deste trabalho, as dificuldades encontradas, os desafios da temática, os resultados obtidos até então, além de uma breve discussão sobre a continuidade da pesquisa.

6.1 Conclusões

Este trabalho apresentou e avaliou a aplicação da abordagem LLMeFT-Transformer para a classificação de textos, com foco em dados estruturados em formato tabular. A pesquisa investigou o potencial dessa abordagem para lidar com o desafio da detecção de discurso de ódio e toxicidade em comentários online.

A abordagem foi implementada utilizando diversos modelos de linguagem como *Text Embedders*, dentre eles Sentence-Transformers, modelos BERT e modelos de embeddings comercias como o OpenAI *text-embedding-3-large* e o VoyageAI *voyage-large-2*. O desempenho do modelo foi avaliado em conjuntos de dados de teste e validação, sem a aplicação de *fine-tuning* ou *transfer-learning*.

Os resultados da abordagem frente a classificação binária de comentários textuais foram promissores, obtendo uma acurácia de 76% e um *F1-score* de 75% no conjunto teste ao aplicar o modelo de embedding *text-embedding-3-large* da OpenAI como *Text Embedder*.

Modelos de Tabular Deep Learning, como FT-Transformer, se destacam em tarefas de aprendizagem multimodal e processamento estruturado de dados heterogêneos com ajustes mínimos. Apesar da necessidade de utilização de GPU em algum momento do processo, o treinamento do modelo mostrou-se eficiente, sendo realizado em poucos minutos em ambientes do Google Colab com GPU gratuita.

O LLMeFT-Transformer demonstrou a possibilidade e capacidade de adequação dos modelos de Tabular Deep Learning (TDL), para lidar com dados textuais no formato tabular naturalmente, sem a necessidade de grandes ajustes ou transformações, enfatizando a importância da adaptação e aplicação de modelos TDL em tarefas de classificação textual. Com essa abordagem e estudo há um indicativo ou abertura para implementar e aperfeiçoar outros modelos de TDL, assim como investigar o uso de modelos como BGE M3 (FlagEmbeddings) (CHEN et al., 2023a) e de técnicas como LLM2Vec (BEHNAMGHADER et al., 2024), que é uma abordagem não supervisionada para transformar qualquer LLM baseado apenas em decodificador em um *Text Embedder*.

6.2 Limitações

Um desafio ainda presente na detecção de ódio e toxicidade online perpassa principalmente pela Inteligência Artificial, onde as técnicas de Aprendizagem de Máquina (AM) clássica e profunda requerem e dependem de conjuntos de dados adequados para serem treinadas e se tornarem mais eficazes. No caso do português e do discurso ofensivo e/ou de ódio no Brasil, existem apenas um pouco mais de quatro conjuntos para a tarefa, todos desbalanceados, requerendo o uso técnicas para o tratamento do desbalanceamento das classes.

Há também outro desafio intrínseco não superado que persiste nas etapas de elaboração e processamento dos dados, estendendo-se para a modelagem e treinamento de modelos de AM. Esse desafio é o viés humano incorporado durante a aquisição, anotação e tratamento de dados. Viés que interfere diretamente na qualidade dos resultados alcançados por abordagens de identificação de toxicidade e ódio que utilizem modelos de AM. Tanto o viés nos dados quanto no algoritmo ou modelagem podem fazer com que o modelo aprenda ou interprete erroneamente os sinais por ignorarem todas as informações contidas nos dados. Portanto, o modelo pode ser direcionado a não incorporar relações relevantes entre as entradas (características) e as saídas (predições) [1].

Do ponto de vista da Inteligência Artificial Ética ou IA Ética, o viés inserido durante e em qualquer estágio do *pipeline* de desenvolvimento de modelos contribui diretamente para uma injusta e desproporcional predição de um dado (SUTARIA, 2022; LIANG et al., 2022), podendo prejudicar grupos específicos ou indivíduos a partir de suas características e de como elas foram representadas e processadas. O Viés e a Equidade na IA e na AM são faces da mesma moeda, significando que a não observância de aspectos relacionados as diversas formas de viés afeta e não garante que o modelo alcance um estado de generalização (aprendizado) que seja imparcial, ou *fairness*, mesmo que seus resultados representados via acurácia, precisão e outras métricas se apresentem como bons, ou seja, um alto valor.

6.3 Trabalhos Futuros

A presente pesquisa demonstrou o potencial da abordagem LLMeFT-Transformer com modelos de TDL para a classificação de textos em formato tabular, especialmente no contexto da detecção de discurso de ódio e toxicidade. No entanto, reconhecemos que este é apenas o ponto de partida para uma exploração mais ampla e profunda deste campo de pesquisa. As seguintes direções de pesquisa são propostas para aprofundar o entendimento e aprimorar as soluções para o problema do discurso de ódio e toxicidade online:

1. Curadoria, Geração de Dados e Anotação Ética: coleta e processamento de novos

[1] Understanding Bias in Machine Learning Models

dados, criação de *datasets*, anotação assistida por agentes de IA e geração de dados sintéticos com LLMs;

2. Modelagem de *Embeddings* e técnicas de representação textual: investigar e aplicar técnicas de *embeddings* assim como buscar melhorias e alternativas de hibridização para melhor obtenção de *embeddings* mais representativas, incluem-se nessa investigação *Sentence Embeddings* e *Graph Embeddings*;

3. Explicabilidade e Interpretabilidade de modelos de aprendizagem profunda com foco para *Tabular Deep Learning* e *Geometric Deep Learning* permitindo principalmente a integração de ferramentas de interpretabilidade como abordado por Sahakyan, Aung e Rahwan (2021);

4. Criação e adaptação de modelos de AP: Investigar, adaptar e criar modelos de aprendizagem profunda tabular e geométrica (uni e multimodal) para classificação de textos como o TabTransformer, ExcelFormer, GraphSAGE, MeshCNN, Hyperbolic Neural Networks, GATs (*Graph Attention Networks*) e GCNs (*Graph Convolutional Networks*);

5. Mitigação de viés: Investigar proativamente estratégias de detecção de viéses como a abordagem de Nascimento, Cavalcanti e Costa-Abreu (2022), o *Fair Representation Learning* (FRL), *Adversarial Debiasing* e *Fairness Constraints (Regularization)*;

6. Impacto socio-psicológico e a criação de aplicações com IA: Adaptar o modelo para aplicações em tempo real, como a moderação de conteúdo em plataformas de mídia social, garantindo respostas rápidas e eficientes para conter a disseminação de discurso de ódio assim como o relatório de postagens e comentários nocivos mais descritivos.

Essas direções podem contribuir para um aprofundamento do impacto social e técnico desta temática, fortalecendo a qualidade dos modelos e a precisão na detecção de diferentes formas de discurso tóxico e de ódio. O que permite a expansão do escopo da pesquisa para incluir o reconhecimento de padrões mais complexos no discurso, como no *gaslighting*, na desinformação e no circunlóquio (padrão de discurso presente principalmente nos debates políticos e ideológicos).

Referências

ABBAS, M. et al. Extracting signals of higgs boson from background noise using deep neural networks. *arXiv preprint arXiv:2010.08201*, 2020. Citado na página 19.

ABERA, B.; DECHASA, S. A review of natural language processing techniques: Application to afan oromo. *International Journal of Computer Applications Technology and Research*, Volume 10, p. 51–54, 03 2021. Citado na página 13.

ABNEY, S. *Semisupervised learning for computational linguistics*. [S.l.]: CRC Press, 2007. Citado na página 15.

ADAM-BOURDARIOS, C. et al. The higgs machine learning challenge. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2015. v. 664, n. 7, p. 072015. Citado na página 19.

AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. Contextual string embeddings for sequence labeling. In: *Proceedings of the 27th international conference on computational linguistics*. [S.l.: s.n.], 2018. p. 1638–1649. Citado na página 28.

AKEN, B. van et al. Visbert: Hidden-state visualizations for transformers. In: . [S.l.: s.n.], 2020. p. 207–211. Citado na página 24.

ALMEIDA, T. S. et al. Sabi\'a-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*, 2024. Citado 2 vezes nas páginas 23 e 38.

ALSHAREF, A. et al. An automated toxicity classification on social media using lstm and word embedding. *Computational Intelligence and Neuroscience*, Hindawi, v. 2022, 2022. Citado 2 vezes nas páginas 4 e 5.

ALTMAN, A. Liberalism and campus hate speech: A philosophical examination. *Ethics*, v. 103, p. 302 – 317, 1993. Citado na página 3.

ALVAREZ, A.; WINTER, F. Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, v. 34, p. 223–237, 2018. Citado na página 3.

ANAGNOSTOU, A.; MOLLAS, I.; TSOUMAKAS, G. Hatebusters: A web application for actively reporting youtube hate speech. In: *IJCAI*. [S.l.: s.n.], 2018. p. 5796–5798. Citado 2 vezes nas páginas 33 e 37.

ANDERSON, L.; BARNES, M. Hate speech. 2022. Citado na página 5.

ARCO, F. M. Plaza-del et al. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, Elsevier, v. 166, p. 114120, 2021. Citado 2 vezes nas páginas 34 e 37.

ARIK, S. Ö.; PFISTER, T. Tabnet: Attentive interpretable tabular learning. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2021. v. 35, n. 8, p. 6679–6687. Citado na página 24.

ARSHAD-AYAZ, A.; NASEEM, M. A.; HIZAOUI, H. Perspectives of muslim and minority canadian youth on hate speech and social media. 07 2022. Disponível em: <https://scite.ai/reports/10.20355/jcie29489>. Citado na página 6.

ASSIS, G. et al. Exploring portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. [S.l.: s.n.], 2024. p. 301–311. Citado na página 38.

BANKO, M.; MACKEEN, B.; RAY, L. A unified taxonomy of harmful content. In: *Proceedings of the fourth workshop on online abuse and harms*. [S.l.: s.n.], 2020. p. 125–137. Citado na página 3.

BARDFORD, V. *Será que chegou a hora de pararmos de chamar as pessoas de 'gordas'?* 2012. BBC News Magazine. Disponível em: <https://www.bbc.com/portuguese/noticias/2012/06/120531_gordo_linguagem_mv>. Citado na página 6.

BASTANI, P.; BAHRAMI, M. Covid-19 related misinformation on social media: A qualitative study from iran. *Journal of medical Internet research*, 2020. Citado na página 3.

BEATTY, M. Graph-based methods to detect hate speech diffusion on twitter. In: IEEE. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.], 2020. p. 502–506. Citado na página 31.

BEHNAMGHADER, P. et al. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. Citado na página 55.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. *Advances in neural information processing systems*, v. 13, 2000. Citado na página 28.

BERKMAN KLEIN CENTER. *Harmful Speech Online QA*. 2016. Disponível em: <https://cyber.harvard.edu/node/99714>. Citado na página 4.

BHAT, R. *Unlocking the Transition: LM to LLM Evolution*. 2024. LinkedIn. Disponível em: <https://www.linkedin.com/pulse/unlocking-transition-lm-llm-evolution-raviraja-bhat-idplc/>. Citado na página 22.

BISPO, T. D. Arquitetura lstm para classificação de discursos de ódio cross-lingual inglês-ptbr. Pós-Graduação em Ciência da Computação, 2018. Citado 3 vezes nas páginas 33, 36 e 37.

BOJANOWSKI, P. et al. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, MIT Press - Journals, v. 5, p. 135–146, dec 2017. ISSN 2307-387X. Citado na página 28.

BÖLÜCÜ, N.; CANBAY, P. Hate speech and offensive content identification with graph convolutional networks. In: *Forum for information retrieval evaluation (working notes)(FIRE), CEUR-WS. org*. [S.l.: s.n.], 2021. Citado na página 31.

BONACCORSO, G. *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. [S.l.]: Packt Publishing Ltd, 2018. Citado na página 15.

BOURDIEU, P. Symbolic power. *Critique of anthropology*, Sage Publications Sage CA: Thousand Oaks, CA, v. 4, n. 13-14, p. 77–85, 1979. Citado na página 7.

BOURGONJE, P. et al. Automatic classification of abusive language and personal attacks in various forms of online communication. In: SPRINGER INTERNATIONAL PUBLISHING. *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings 27*. [S.l.], 2018. p. 180–191. Citado na página 33.

BOURGONJE, P. et al. Automatic classification of abusive language and personal attacks in various forms of online communication. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.l.: s.n.], 2018. v. 10713 LNAI, p. 180–191. ISBN 9783319737058. ISSN 16113349. Citado na página 37.

BRASOVEANU, A.; ANDONIE, R. Visualizing transformers for nlp: A brief survey. *2020 24th International Conference Information Visualisation (IV)*, p. 270–279, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:232267493>. Citado 2 vezes nas páginas 21 e 24.

BUARQUE, B.; CRETTON, M. *Mapa do Ódio no Brasil - PERCEPÇÕES E RECOMENDAÇÕES PARA POLÍTICAS PÚBLICAS*. [S.l.], 2021. 70 p. Disponível em: <https://www.wordshealtheworld.com/wp-content/uploads/2021/01/HATE-MAP-IN-PORTUGUESE-2019-final.pdf>. Citado na página 2.

BURNAP, P.; WILLIAMS, M. L. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, SpringerOpen, v. 5, n. 1, p. 11, dec 2016. ISSN 21931127. Disponível em: <http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0072-6>. Citado 3 vezes nas páginas 31, 32 e 37.

CANDAL LUDMILA, V. T. *Especialistas divergem sobre eventuais crimes em falas sobre nazismo no Flow*. 2022. Disponível em: <https://www.cnnbrasil.com.br/nacional/especialistas-divergem-sobre-eventuais-crimes-em-falas-sobre-nazismo-no-flow/>. Citado na página 5.

CAO, L. Deep learning applications. *IEEE Intelligent Systems*, v. 37, n. 3, p. 3–5, 2022. Citado na página 20.

CAPOZZI, A. T. et al. A data viz platform as a support to study, analyze and understand the hate speech phenomenon. In: *Proceedings of the 2nd International Conference on Web Studies*. [S.l.: s.n.], 2018. p. 28–35. Citado na página 4.

CASELI, H.; FREITAS, C.; VIOLA, R. Processamento de linguagem natural. *Sociedade Brasileira de Computação*, 2022. Citado na página 14.

CASELI, H. d. M.; NUNES, M. d. G. V.; PAGANO, A. S. O que é pln? *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*, 2023. Citado 2 vezes nas páginas 15 e 13.

CETIC.BR. *92 milhões de brasileiros acessam a Internet apenas pelo telefone celular, aponta TIC Domicílios 2022*. 2023. Comitê Gestor da Internet no Brasil (CGI.br). Disponível em: <https://cetic.br/pt/noticia/

92-milhoes-de-brasileiros-acessam-a-internet-apenas-pelo-telefone-celular-aponta-tic-domicilios-2022/>. Citado na página 1.

CHEN, J. et al. Tabcaps: A capsule neural network for tabular data classification with bow routing. In: *The Eleventh International Conference on Learning Representations*. [S.l.: s.n.], 2022. Citado na página 24.

CHEN, J. et al. Danets: Deep abstract networks for tabular data classification and regression. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2022. v. 36, n. 4, p. 3930–3938. Citado na página 24.

CHEN, J. et al. *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 2023. Citado na página 55.

CHEN, J. et al. Excelformer: A neural network surpassing gbdts on tabular data. *arXiv preprint arXiv:2301.02819*, 2023. Citado na página 24.

CHEN, K.-Y. et al. Trompt: Towards a better deep neural network for tabular data. *arXiv preprint arXiv:2305.18446*, 2023. Citado na página 24.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado 2 vezes nas páginas 17 e 19.

CHOLAKOV, R.; KOLEV, T. The gatedtabtransformer. an enhanced deep learning architecture for tabular modeling. *arXiv preprint arXiv:2201.00199*, 2022. Citado na página 24.

CHOPRA, A. et al. A framework for online hate speech detection on code-mixed hindi-english text and hindi text in devanagari. *ACM Transactions on Asian and Low-Resource Language Information Processing*, ACM New York, NY, v. 22, n. 5, p. 1–21, 2023. Citado 2 vezes nas páginas 31 e 38.

CHOWDHURY, A. G. et al. Arhnet-leveraging community interaction for detection of religious hate speech in arabic. In: *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*. [S.l.: s.n.], 2019. p. 273–280. Citado 2 vezes nas páginas 34 e 37.

CINELLI, M. et al. Dynamics of online hate and misinformation. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 22083, 2021. Citado na página 4.

CJADAMS JEFFREY SORENSEN, J. E. L. D. M. M. n. W. C. *Toxic Comment Classification Challenge*. Kaggle, 2017. Disponível em: <https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge>. Citado na página 5.

CLARK, K. et al. What does BERT look at? an analysis of BERT's attention. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, 2019. p. 276–286. Disponível em: <https://aclanthology.org/W19-4828>. Citado na página 24.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th international conference on Machine learning*. [S.l.: s.n.], 2008. p. 160–167. Citado na página 28.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. *Journal of machine learning research*, v. 12, n. ARTICLE, p. 2493–2537, 2011. Citado na página 28.

DALE, R. Classical approaches to natural language processing. In: *Handbook of natural language processing, second edition*. [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7. Citado 2 vezes nas páginas 15 e 13.

DAMAS, G. et al. A transformer-based tabular approach to detect toxic comments. In: *To appear in the 34th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2024. Citado na página 9.

DATABRICKS. *Large Language Models (LLMs)*. 2024. What are Large Language Models (LLMs)? Disponível em: <https://www.databricks.com/glossary/large-language-models-llm>. Citado na página 22.

DAVIDSON, T.; BHATTACHARYA, D.; WEBER, I. *Racial bias in hate speech and abusive language detection datasets*. arXiv, 2019. Disponível em: <http://arxiv.org/abs/1905.12516>. Citado na página 4.

DEV, V. A.; EDEN, M. R. Gradient boosted decision trees for lithology classification. In: *Computer aided chemical engineering*. [S.l.]: Elsevier, 2019. v. 47, p. 113–118. Citado na página 17.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>. Citado 3 vezes nas páginas 11, 23 e 29.

DHINAKARAN, M. et al. *Machine Learning For Absolute Beginners*. [S.l.]: Booksclinic Publishing, 2023. Citado na página 15.

DJURIC, N. et al. Hate speech detection with comment embeddings. In: *Proceedings of the 24th international conference on world wide web*. [S.l.: s.n.], 2015. p. 29–30. Citado na página 32.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018. Citado na página 17.

DUMAIS, S. T. et al. Using latent semantic analysis to improve access to textual information. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.: s.n.], 1988. p. 281–285. Citado na página 28.

DUONG, C.; ZHANG, L.; LU, C.-T. Hatenet: A graph convolutional network approach to hate speech detection. In: IEEE. *2022 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2022. p. 5698–5707. Citado na página 31.

DUTTA, J.; KIM, Y. W.; DOMINIC, D. Comparison of gradient boosting and extreme boosting ensemble methods for webpage classification. In: *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. [S.l.: s.n.], 2020. p. 77–82. Citado na página 18.

ECONOMIST, T. *In Japan "Oppenheimer" is causing consternation*. 2023. Disponível em: <https://www.economist.com/culture/2023/08/10/in-japan-oppenheimer-is-causing-consternation>. Citado na página 5.

EISENSTEIN, J. *Introduction to natural language processing*. [S.l.]: MIT press, 2019. Citado na página 11.

EVKOSKI, B. et al. Evolution of topics and hate speech in retweet network communities. 12 2021. Disponível em: <https://scite.ai/reports/10.1007/s41109-021-00439-7>. Citado na página 6.

EVKOSKI, B. et al. Retweet communities reveal the main sources of hate speech. 03 2022. Disponível em: <https://scite.ai/reports/10.1371/journal.pone.0265602>. Citado na página 6.

FABELLO, M. A. *Skinny Shaming Is Not the Same as Fat Phobia*. 2018. SELF - Health. Disponível em: <https://www.self.com/story/skinny-shaming-is-not-the-same-as-fat-phobia>. Citado na página 6.

FATO, B. de. *O que É gaslighting: Você Não Está Ficando Maluca, nem É Coisa da Sua Cabeça*. 2022. Disponível em: <https://www.brasildefato.com.br/2022/03/23/o-que-e-gaslighting-voce-nao-esta-ficando-maluca-nem-e-coisa-da-sua-cabeca#:~:text=O%20termo%20%C3%A9%20usado%20para,na%20pr%C3%B3pria%20percep%C3%A7%C3%A3o%20da%20realidade.> Citado na página 5.

FINATTO, M. J. B. et al. Sequência de caracteres e palavras. *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*, 2023. Citado na página 25.

FIRMINO, A. A. et al. *Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas*. Tese (Doutorado), 2022. Citado na página 4.

FIRTH, J. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, p. 10–32, 1957. Citado na página 26.

FIX, E. *Discriminatory analysis: nonparametric discrimination, consistency properties*. [S.l.]: USAF school of Aviation Medicine, 1985. v. 1. Citado na página 16.

FORTUNA, P. *Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes*. 109 p. Tese (Doutorado) — Universidade do Porto, aug 2017. Disponível em: <https://repositorio-aberto.up.pt/handle/10216/106028>. Citado na página 35.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, ACM, v. 51, n. 4, p. 1–30, jul 2018. ISSN 15577341. Disponível em: <http://dl.acm.org/citation.cfm?doid=3236632.3232676>. Citado 2 vezes nas páginas 5 e 31.

FORTUNA, P. et al. A hierarchically-labeled Portuguese hate speech dataset. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, 2019. p. 94–104. Disponível em: <https://aclanthology.org/W19-3510>. Citado 3 vezes nas páginas 7, 34 e 39.

FORTUNA, P.; SOLER-COMPANY, J.; WANNER, L. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, Elsevier, v. 58, n. 3, p. 102524, 2021. Citado 3 vezes nas páginas 34, 36 e 37.

FRENDA, S. et al. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of intelligent & fuzzy systems*, IOS Press, v. 36, n. 5, p. 4743–4752, 2019. Citado 2 vezes nas páginas 34 e 37.

FREUD, S. *Group Psychology and the Analysis of the Ego*. [S.l.]: Read Books Ltd, 2014. Citado na página 3.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999. Citado na página 18.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997. Citado na página 18.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. Citado na página 18.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 18.

GABARRON, E.; OYEYEMI, S. O.; WYNN, R. Covid-19-related misinformation on social media: a systematic review. *Bulletin of the World Health Organization*, v. 99, p. 455 – 463A, 2021. Citado na página 2.

GARVIN, P. L. *The Georgetown-IBM experiment of 1954: an evaluation in retrospect*. [S.l.]: Mouton, 1968. Citado na página 11.

GEEKSFORGEEKS. *What are some key strengths of BERT over ELMO/ULMFiT?* 2024. Disponível em: <https://www.geeksforgeeks.org/what-are-some-key-strengths-of-bert-over-elmo-ulmfit/>. Citado na página 23.

GELBER, K. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, v. 24, p. 393 – 414, 2019. Citado na página 3.

GHATI, G. *Comparison between BERT, GPT-2 and ELMo*. 2020. MEDIUM. Disponível em: <https://medium.com/@gauravghati/comparison-between-bert-gpt-2-and-elmo-9ad140cd1cda>. Citado na página 23.

GHOSH, S. et al. *CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network*. 2023. Citado na página 31.

GILLIOZ, A. et al. Overview of the transformer-based models for nlp tasks. In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. [S.l.: s.n.], 2020. p. 179–183. Citado na página 21.

GOOGLE. *Introduction to Large Language Models*. 2023. Google for Developers - Machine Learning. Disponível em: <https://developers.google.com/machine-learning/resources/intro-llms>. Citado na página 23.

GORISHNIY, Y. et al. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, v. 34, p. 18932–18943, 2021. Citado 8 vezes nas páginas 15, 8, 17, 24, 41, 43, 44 e 49.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, v. 35, p. 507–520, 2022. Citado na página 17.

GRūZīTIS, N. *Formal Grammar and Semantics of Controlled Latvian Language*. Tese (Doutorado), 12 2010. Citado na página 13.

GUIMARÃES, S. S. et al. Characterizing toxicity on facebook comments in brazil. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. [S.l.: s.n.], 2020. p. 253–260. Citado 2 vezes nas páginas 4 e 5.

HANSEN, R. W. You've never been welcome here: exploring the relationship between exclusivity and incivility in online forums. *Journal of Information Technology & Politics*, Taylor & Francis, v. 20, n. 2, p. 139–153, 2023. Citado na página 4.

HARRIS, Z. S. Distributional structure. *Word*, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. Citado na página 26.

HEGSELMANN, S. et al. Tabllm: Few-shot classification of tabular data with large language models. In: PMLR. *International Conference on Artificial Intelligence and Statistics*. [S.l.], 2023. p. 5549–5581. Citado na página 24.

HOLLMANN, N. et al. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022. Citado na página 24.

HU, W. et al. Pytorch frame: A modular framework for multi-modal tabular learning. *arXiv preprint arXiv:2404.00776*, 2024. Citado 4 vezes nas páginas 15, 41, 45 e 49.

HUANG, X. et al. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020. Citado na página 24.

JIANG, Y. et al. Domain-adaptive graph based on post-hoc explanation for cross-domain hate speech detection. In: IEEE. *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. [S.l.], 2022. p. 1271–1276. Citado na página 31.

JOSEPH, M.; RAJ, H. *GANDALF: Gated Adaptive Network for Deep Automated Learning of Features*. 2023. Citado na página 24.

JOULIN, A. et al. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. Citado na página 28.

JUNQUEIRA, J. et al. Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2023. p. 146–155. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25446>. Citado na página 38.

JUNQUEIRA, J. da R. et al. Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In: SBC. *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.], 2023. p. 146–155. Citado na página 51.

KAPANTAI, E. et al. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, SAGE Publications Sage UK: London, England, v. 23, n. 5, p. 1301–1326, 2021. Citado na página 3.

KARAYIğIT, H.; AKDAGLI, A.; ACı Çiğdem İnan. Homophobic and hate speech detection using multilingual-bert model on turkish social media. 06 2022. Disponível em: <https://scite.ai/reports/10.5755/j01.itc.51.2.29988>. Citado 2 vezes nas páginas 6 e 7.

KARIM, M. R. et al. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. *arXiv e-prints*, p. arXiv–2012, 2020. Citado na página 35.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 17.

KITAEV, N.; KAISER, L.; LEVSKAYA, A. Reformer: The efficient transformer. *CoRR*, abs/2001.04451, 2020. Disponível em: <https://arxiv.org/abs/2001.04451>. Citado na página 23.

KOBAYASHI, G. et al. Attention module is not only a weight: Analyzing transformers with vector norms. *CoRR*, abs/2004.10102, 2020. Disponível em: <https://arxiv.org/abs/2004.10102>. Citado na página 24.

LAI, S. et al. How to generate a good word embedding. *IEEE Intelligent Systems*, IEEE, v. 31, n. 6, p. 5–14, 2016. Citado na página 28.

LAN, Z. et al. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. Disponível em: <http://arxiv.org/abs/1909.11942>. Citado na página 23.

LANIUS, C. L.; WEBER, R.; MACKENZIE, W. I. Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey. *Social Network Analysis and Mining*, v. 11, 2021. Citado na página 3.

LEITE, J. A. et al. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2020. p. 914–924. Disponível em: <https://aclanthology.org/2020.aacl-main.91>. Citado 6 vezes nas páginas 7, 36, 39, 41, 42 e 51.

LEV-ON, A. The anti-social network? framing social media in wartime. 07 2018. Disponível em: <https://scite.ai/reports/10.1177/2056305118800311>. Citado na página 7.

LI, Q. et al. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020. Citado 2 vezes nas páginas 23 e 29.

LIANG, W. et al. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, Nature Publishing Group UK London, v. 4, n. 8, p. 669–677, 2022. Citado na página 56.

LIMA, L. H. Q.; PAGANO, A. S.; SILVA, A. P. C. da. Toxic content detection in online social networks: a new dataset from brazilian reddit communities. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. [S.l.: s.n.], 2024. p. 472–482. Citado na página 39.

LIU, R.; JIANG, D.; ZHANG, X. A stable deep reinforcement learning framework for recommendation. *IEEE Intelligent Systems*, v. 37, n. 3, p. 76–84, 2022. Citado na página 20.

LIU, Y. et al. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. Disponível em: <http://arxiv.org/abs/1907.11692>. Citado na página 23.

LIU, Z. et al. Word representation. *Representation Learning for Natural Language Processing*, Springer, p. 13–41, 2020. Citado na página 28.

LIU, Z.; SUN, M. Representation learning and nlp. In: *Representation Learning for Natural Language Processing*. [S.l.]: Springer, 2023. p. 1–27. Citado 4 vezes nas páginas 15, 26, 27 e 29.

LOPES, R.; MAGALHÃES, J.; SEMEDO, D. Gl\'oria-a generative and open large language model for portuguese. *arXiv preprint arXiv:2402.12969*, 2024. Citado na página 38.

LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 16.

LUO, J.; BOUAZIZI, M.; OHTSUKI, T. Data augmentation for sentiment analysis using sentence compression-based seqgan with data screening. 01 2021. Disponível em: <https://scite.ai/reports/10.1109/access.2021.3094023>. Citado na página 6.

LUTHY, T. *TEFT-Transformer: A Novel Approach for Fake Review Detection*. 2023. Disponível em: <http://hdl.handle.net/2105/67638>. Citado 3 vezes nas páginas 43, 45 e 49.

MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020. Citado 2 vezes nas páginas 15 e 16.

MAITHRA; SCHMIDT, E. A survey of deep learning for scientific discovery. *CoRR*, abs/2003.11755, 2020. Disponível em: <https://arxiv.org/abs/2003.11755>. Citado na página 21.

MALIK, P.; AGGRAWAL, A.; VISHWAKARMA, D. K. Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, p. 1254–1259, 2021. Disponível em: <https://api.semanticscholar.org/CorpusID:233990897>. Citado na página 20.

MARS, M. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, MDPI, v. 12, n. 17, p. 8805, 2022. Citado 2 vezes nas páginas 27 e 29.

MATHEW, B. et al. Hatexplain: A benchmark dataset for explainable hate speech detection. 05 2021. Disponível em: <https://scite.ai/reports/10.1609/aaai.v35i17.17745>. Citado na página 6.

MAYR, A. et al. The evolution of boosting algorithms. *Methods of information in medicine*, Schattauer GmbH, v. 53, n. 06, p. 419–427, 2014. Citado na página 18.

MCCANN, B. et al. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 28.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, Springer, v. 52, p. 99–115, 1990. Citado na página 16.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, v. 26, 2013. Citado 2 vezes nas páginas 11 e 29.

MINAEE, S. et al. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024. Citado na página 23.

MIRANDA, S. et al. Hate speech on social media: behaviour of portuguese football fans on facebook. *Soccer & Society*, Taylor & Francis, p. 1–16, 2023. Citado na página 4.

MIRANDA, S. et al. Hate speech on social media: behaviour of portuguese football fans on facebook. *Soccer & Society*, Taylor & Francis, v. 25, n. 1, p. 76–91, 2024. Citado na página 39.

MISHRA, P. et al. Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*, 2019. Citado na página 31.

MNIH, A.; HINTON, G. Three new graphical models for statistical language modelling. In: *Proceedings of the 24th international conference on Machine learning*. [S.l.: s.n.], 2007. p. 641–648. Citado na página 28.

MOHAMMED, M.; KHAN, M. B.; BASHIER, E. B. M. *Machine learning: algorithms and applications*. [S.l.]: Crc Press, 2016. Citado na página 15.

MÜLLER, A. *Política do ódio no Brasil*. [S.l.]: Viseu, 2019. Citado na página 3.

MüLLER, K.; SCHWARZ, C. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 2020. Citado na página 3.

NASCIMENTO, F. R.; CAVALCANTI, G. D.; COSTA-ABREU, M. D. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, Elsevier, v. 201, 2022. Citado na página 57.

NASEEM, U. et al. *A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models*. 2020. Citado 2 vezes nas páginas 26 e 27.

NAVEED, H. et al. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. Citado na página 22.

NEMER, D. The three types of whatsapp users getting brazil's jair bolsonaro elected. *The Guardian*, v. 25, p. 2018, 2018. Citado na página 3.

NGUYEN, L. T.; NGUYEN, K. V.; NGUYEN, N. L.-T. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In: SPRINGER. *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*. [S.l.], 2021. p. 572–583. Citado 2 vezes nas páginas 4 e 5.

NIELSEN, L. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social Issues*, v. 58, p. 265–280, 2002. Citado na página 3.

NOBATA, C. et al. Abusive language detection in online user content. In: *Proceedings of the 25th international conference on world wide web*. [S.l.: s.n.], 2016. p. 145–153. Citado 4 vezes nas páginas 4, 31, 32 e 37.

OCAMPO, N. et al. An in-depth analysis of implicit and subtle hate speech messages. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. p. 1997–2013. Disponível em: <https://aclanthology.org/2023.eacl-main.147>. Citado 2 vezes nas páginas 4 e 5.

OLIVEIRA, A. et al. How good is chatgpt for detecting hate speech in portuguese? In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2023. p. 94–103. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25441>. Citado 2 vezes nas páginas 38 e 51.

OLIVEIRA, A. da S. et al. Toxic speech detection in portuguese: A comparative study of large language models. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. [S.l.: s.n.], 2024. p. 108–116. Citado 2 vezes nas páginas 38 e 51.

OLIVEIRA, F.; REIS, V.; EBECKEN, N. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*, 2023. Citado 3 vezes nas páginas 7, 37 e 39.

ORTIZ, V. *Programa aborda 'magrofobia' e cria polêmica; esse preconceito existe?* 2017. UOL - UNIVERSA. Disponível em: <https://www.uol.com.br/universa/noticias/redacao/2017/02/08/programa-aborda-magrofobia-e-cria-polemica-esse-preconceito-existe.htm?cmpid=copiaecola>. Citado na página 6.

OZ, M.; NURUMOV, B. Intolerant versus uncivil: Examining types, directions and deliberative attributes of incivility on facebook versus twitter. *First Monday*, 2022. Citado na página 4.

PANDEY, S. S. *A Comparative study of BERT-CNN and GCN for Hate Speech Detection*. Dissertação (B.S. thesis) — University of Twente, 2023. Citado na página 31.

PAPCUNOVÁ, J. et al. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems*, Springer, v. 9, n. 3, p. 2827–2842, 2023. Citado na página 5.

PELLE, R. P. D.; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. In: SBC. *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2017. Citado 4 vezes nas páginas 31, 34, 35 e 39.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado 3 vezes nas páginas 9, 11 e 28.

PÉREZ, J. M. et al. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, IEEE, v. 11, p. 30575–30590, 2023. Citado na página 5.

PETERS, M. E. et al. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017. Citado na página 29.

PETERS, M. E. et al. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. Disponível em: <http://arxiv.org/abs/1802.05365>. Citado 2 vezes nas páginas 23 e 29.

PIAUí, S. d. S. P. *Boletim de Dados de Violência Contra a Pessoa LGBTQIAPN+*. Teresina, Piauí, Brasil, 2024. Coordenação de Proteção aos LGBTQIAPN+ e Núcleo de Pesquisa Sobre Crianças, Adolescentes e Jovens (Nupec) da Universidade Federal do Piauí (UFPI). Citado na página 1.

PICOVOICE. *What is a Large Language Model?* 2023. Disponível em: <https://picovoice.ai/blog/what-is-a-large-language-model/>. Citado na página 22.

PIRES, R. et al. Sabiá: Portuguese large language models. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2023. p. 226–240. Citado 2 vezes nas páginas 23 e 38.

PITSILIS, G. K.; RAMAMPIARO, H.; LANGSETH, H. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, Springer, v. 48, p. 4730–4742, 2018. Citado 2 vezes nas páginas 33 e 37.

POLETTO, F. et al. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, v. 55, p. 477 – 523, 2020. Citado na página 3.

POPOV, S.; MOROZOV, S.; BABENKO, A. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019. Citado na página 24.

QIN, G.; FENG, Y.; DURME, B. V. The nlp task effectiveness of long-range transformers. *arXiv preprint arXiv:2202.07856*, 2022. Citado na página 31.

RADFORD, A. et al. Improving language understanding by generative pre-training. OpenAI, 2018. Citado na página 29.

RADFORD, A. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019. Citado na página 29.

RAE, J. W. et al. Compressive transformers for long-range sequence modelling. *CoRR*, abs/1911.05507, 2019. Disponível em: <http://arxiv.org/abs/1911.05507>. Citado na página 31.

RASCHKA, S.; MIRJALILI, V. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. [S.l.]: Packt Publishing Ltd, 2019. Citado na página 14.

RAY, S. A quick review of machine learning algorithms. In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. [S.l.: s.n.], 2019. p. 35–39. Citado 2 vezes nas páginas 15 e 16.

REBALA, G.; RAVI, A.; CHURIWALA, S. *An introduction to machine learning*. [S.l.]: Springer, 2019. Citado na página 15.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. Citado na página 8.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. Disponível em: <http://arxiv.org/abs/1602.04938>. Citado na página 24.

RIVAS, P. *Deep Learning for Beginners: A beginner's guide to getting up and running with deep learning from scratch using Python*. [S.l.]: Packt Publishing Ltd, 2020. Citado na página 14.

RODRIGUES, J. et al. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*, 2023. Citado 4 vezes nas páginas 8, 9, 23 e 46.

ROGERS, K. *"gaslighting": Veja o que realmente significa a Palavra da Moda*. Disponível em: <https://www.cnnbrasil.com.br/saude/gaslighting-veja-o-que-realmente-significa-a-palavra-da-moda/>. Citado na página 5.

ROOZENBEEK, J. et al. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, v. 8, 2022. Citado na página 3.

RUSSELL, S. J.; NORVIG, P. *Inteligência artificial*. [S.l.]: Elsevier, 2013. 988 p. ISBN 978-85352-3701-6. Citado na página 27.

SAHA, K.; CHANDRASEKHARAN, E.; CHOUDHURY, M. Prevalence and psychological effects of hateful speech in online college communities. *Proceedings of the 10th ACM Conference on Web Science*, 2019. Citado na página 3.

SAHAKYAN, M.; AUNG, Z.; RAHWAN, T. Explainable artificial intelligence for tabular data: A survey. *IEEE Access*, v. 9, p. 135392–135422, 2021. Citado na página 57.

SAIF, M. A. et al. Classification of online toxic comments using the logistic regression and neural networks models. In: AIP PUBLISHING. *AIP conference proceedings*. [S.l.], 2018. v. 2048, n. 1. Citado 2 vezes nas páginas 4 e 5.

SAKSESI, A. S.; NASRUN, M.; SETIANINGSIH, C. Analysis Text of Hate Speech Detection Using Recurrent Neural Network. In: *Proceedings - 2018 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2018. p. 242–248. ISBN 9781728119939. Citado na página 4.

SALEEM, H. M. et al. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *arXiv*, 2017. ISSN 23318422. Disponível em: <http://www.ta-cos.org/sites/ta-cos.org/files/tacos2016_SaleemDillionBeneschRuths.pdfhttp://arxiv.org/abs/1709.10159>. Citado na página 4.

SAMEK, W.; MÜLLER, K.-R. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, p. 5–22, 2019. Citado na página 24.

SANH, V. et al. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. Disponível em: <http://arxiv.org/abs/1910.01108>. Citado na página 23.

SARAIVA, G. D. et al. A semi-supervised approach to detect toxic comments. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., 2021. p. 1261–1267. Disponível em: <https://aclanthology.org/2021.ranlp-1.142>. Citado 4 vezes nas páginas 8, 36, 37 e 51.

SARRACÉN, G. L. De la P.; ROSSO, P. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. [S.l.: s.n.], 2022. p. 2196–2204. Citado na página 31.

SCHAPIRE, R. E. The strength of weak learnability. *Machine learning*, Springer, v. 5, p. 197–227, 1990. Citado na página 17.

SCHMIDT, A.; WIEGAND, M. A Survey on Hate Speech Detection using Natural Language Processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. p. 1–10. Disponível em: <http://www.aclweb.org/anthology/W/W17/W17-1101.pdfhttp://aclweb.org/anthology/W17-1101http://www.aclweb.org/anthology/W17-1101>. Citado 3 vezes nas páginas 4, 5 e 31.

SCHNEIDER, J. Locality-promoting representation learning. In: IEEE. *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2021. p. 8061–8068. Citado na página 24.

SCHOFIELD, A.; DAVIDSON, T. Identifying hate speech in social media. *XRDS: Crossroads, The ACM Magazine for Students*, ACM, v. 24, n. 2, p. 56–59, dec 2017. ISSN 1528-4972. Disponível em: <http://dl.acm.org/citation.cfm?doid=3175661.3155212>. Citado na página 4.

SCHULZ, K. Natural language processing for teaching ancient languages. *Teaching Classics in the Digital Age*, p. 37–48, 2021. Citado 2 vezes nas páginas 15 e 13.

SCHUMACHER, D. *FT-Transformer*. 2024. SERP AI - Deep Tabular Learning. Disponível em: <https://serp.ai/ft-transformer/>. Citado na página 25.

SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. *Introduction to information retrieval*. [S.l.]: Cambridge University Press Cambridge, 2008. v. 39. Citado na página 23.

SCHWEPPE, J.; PERRY, B. A continuum of hate: delimiting the field of hate studies. *Crime, Law and Social Change*, Springer, v. 77, n. 5, p. 503–528, 2022. Citado na página 4.

SELLARS, A. Defining Hate Speech. *SSRN Electronic Journal*, dec 2017. ISSN 1556-5068. Disponível em: <https://www.ssrn.com/abstract=2882244https://ssrn.com/abstract=2882244>. Citado na página 4.

SEN, J.; MEHTAB, S.; ENGELBRECHT, A. *Machine learning: algorithms, models and applications*. [S.l.]: BoD–Books on Demand, 2021. Citado na página 15.

SHAW, P.; USZKOREIT, J.; VASWANI, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. Citado na página 44.

SHWARTZ-ZIV, R.; ARMON, A. Tabular data: Deep learning is not all you need. *Information Fusion*, Elsevier, v. 81, p. 84–90, 2022. Citado na página 24.

SIEGEL, A.; BADAAN, V. no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, v. 114, p. 837 – 855, 2020. Citado na página 3.

SILVA, F. L. V. d.; FREITAS, L. A. d. Utilizando bertimbau para a classificação de discurso de ódio em português. UFPel, 2022. Citado na página 4.

SILVA, L. A. et al. Analyzing the targets of hate in online social media. *CoRR*, abs/1603.07709, 2016. Disponível em: <http://arxiv.org/abs/1603.07709>. Citado na página 1.

SILVA, S.; SERAPIAO, A. Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In: *Proceedings of KDMILE 2018, Symposium on Knowledge Discovery, Mining and Learning, São Paulo, SP, Brazil*. [S.l.: s.n.], 2018. Citado 3 vezes nas páginas 34, 36 e 37.

SINGH, S.; MAHMOOD, A. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, v. 9, p. 68675–68702, 2021. Citado na página 21.

SKRLJ, B. et al. Attviz: Online exploration of self-attention for transparent neural language modeling. *CoRR*, abs/2005.05716, 2020. Disponível em: <https://arxiv.org/abs/2005.05716>. Citado na página 24.

SMITH, K. *Large Language Models: Reasoning Capabilities and Limitations*. 2023. MEDIUM. Disponível em: <https://medium.com/@glovguy/large-language-models-reasoning-capabilities-and-limitations-951cee0ac642>. Citado na página 22.

SOMEPALLI, G. et al. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021. Citado na página 24.

SONG, W. et al. Autoint: Automatic feature interaction learning via self-attentive neural networks. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. [S.l.: s.n.], 2019. p. 1161–1170. Citado na página 25.

SOROKOWSKI, P. et al. Are online haters psychopaths? psychological predictors of online hating behavior. *Frontiers in Psychology*, v. 11, 2020. Citado na página 3.

SOTO, C. P. et al. Application-specific word embeddings for hate and offensive language detection. *Multimedia Tools and Applications*, Springer, v. 81, n. 19, p. 27111–27136, 2022. Citado 3 vezes nas páginas 35, 36 e 37.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado 4 vezes nas páginas 8, 9, 23 e 46.

SRIVASTAVA, S.; KHURANA, P. Detecting aggression and toxicity using a multi dimension capsule network. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, 2019. p. 157–162. Disponível em: <https://aclanthology.org/W19-3517>. Citado 2 vezes nas páginas 4 e 5.

SUTARIA, N. Bias and ethical concerns in machine learning. *ISACA J.*, v. 4, p. 1–4, 2022. Citado na página 56.

TALBOT, N. L. *User Manual for glossaries.sty*. [S.l.], 2012. Disponível em: <http://mirrors.ctan.org/macros/latex/contrib/glossaries/glossaries-user.pdf>. Acesso em: 11 mar. 2013. Citado na página 3.

TER-SARKISOV, A. One shot model for covid-19 classification and lesions segmentation in chest ct scans using long short-term memory network with attention mechanism. *IEEE Intelligent Systems*, v. 37, n. 3, p. 54–64, 2022. Citado na página 20.

TILLY, C. *Did Japan ban Oppenheimer?* 2023. Dexerto. Disponível em: <https://www.dexerto.com/tv-movies/oppenheimer-japan-ban-2211586/>. Citado na página 5.

TINGLEY, D.; WAGNER, G. Solar geoengineering and the chemtrails conspiracy on social media. *Palgrave Communications*, v. 3, p. 1–7, 2017. Citado na página 3.

TIWARI, T.; TIWARI, T.; TIWARI, S. How artificial intelligence, machine learning and deep learning are radically different? v. 8, p. 1–9, 2018. Citado na página 20.

TOKTAROVA, A. et al. Hate speech detection in social networks using machine learning and deep learning methods. *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, v. 14, n. 5, 2023. Citado na página 20.

TOM, T. Artificial intelligence basics: A non-technical introduction. *Monrovia, CA, USA: Appres*, 2019. Citado na página 14.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Citado na página 23.

TRAJANO, D.; BORDINI, R. H.; VIEIRA, R. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, Springer, p. 1–27, 2023. Citado 2 vezes nas páginas 7 e 39.

TRAJANO, D. d. O. *Detecção de linguagem tóxica aplicada a textos em português*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2023. Citado na página 4.

TSUJII, J. Natural language processing and computational linguistics. *Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 47, n. 4, p. 707–727, 2021. Citado na página 11.

USZKOREIT, J. *Transformer: A novel neural network architecture for language understanding. Google AI Blog*. 2017. Citado na página 44.

VAJJALA, S. et al. *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. [S.l.]: O'Reilly Media, 2020. Citado na página 13.

VARGAS, F. et al. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022. p. 7174–7183. Disponível em: <https://aclanthology.org/2022.lrec-1.777>. Citado 2 vezes nas páginas 7 e 39.

VARGAS, F. et al. NoHateBrazil: A Brazilian Portuguese text offensiveness analysis system. In: MITKOV, R.; ANGELOVA, G. (Ed.). *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023. p. 1180–1186. Disponível em: <https://aclanthology.org/2023.ranlp-1.125>. Citado na página 38.

VASWANI, A. et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>. Citado 7 vezes nas páginas 15, 11, 21, 22, 25, 43 e 44.

VICARIO, M. D. et al. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, National Acad Sciences, v. 113, n. 3, p. 554–559, 2016. Citado na página 3.

VIGNA, F. D. et al. Hate me, hate me not: Hate speech detection on facebook. In: *Italian Conference on Cybersecurity*. [s.n.], 2017. Disponível em: <https://api.semanticscholar.org/CorpusID:8293149>. Citado 3 vezes nas páginas 32, 33 e 37.

WADUD, M. A.; D., M. P.; RAHMAN, M. M. Word embedding methods for word representation in deep learning for natural language processing. *Iraqi Journal of Science*, p. 1349–1361, 03 2022. Citado 2 vezes nas páginas 27 e 28.

WADUD, M. A. H. et al. How can we manage offensive text in social media-a text classification approach using lstm-boost. *International Journal of Information Management Data Insights*, Elsevier, v. 2, n. 2, p. 100095, 2022. Citado na página 17.

WANG, Q. Advanced deep learning approaches for hate speech detection. *Highlights in Science, Engineering and Technology*, 2023. Citado na página 20.

WANG, Q. et al. Learning deep transformer models for machine translation. In: KORHONEN, A.; TRAUM, D.; MÀRQUEZ, L. (Ed.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 1810–1822. Disponível em: <https://aclanthology.org/P19-1176>. Citado na página 44.

WANG, S. et al. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. Citado na página 44.

WANG, Y. et al. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, Springer, v. 11, p. 1611–1630, 2020. Citado na página 29.

WASEEM, Z.; HOVY, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: *Proceedings of the NAACL student research workshop*. [S.l.: s.n.], 2016. p. 88–93. Citado 5 vezes nas páginas 31, 32, 33, 34 e 37.

WEITZEL, L. et al. Investigating deep learning approaches for hate speech detection in social media : Portuguese-br tweets. In: *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2023. p. 1–5. Citado na página 4.

WEN, Z. et al. Thundergbm: Fast gbdts and random forests on gpus. *Journal of Machine Learning Research*, v. 21, n. 108, p. 1–5, 2020. Citado na página 17.

WIJERATNE, Y.; SILVA, N. de. Natural language processing for government: Problems and potential. In: . [s.n.], 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:234797778>. Citado na página 13.

WIRIYATHAMMABHUM, P. et al. Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Computing Surveys*, v. 49, p. 1–44, 12 2016. Citado na página 13.

YAM, K. *'Oppenheimer' draws debate over the absence of Japanese bombing victims in the film*. 2023. NBC News - Asian America. Disponível em: <https://www.nbcnews.com/news/asian-america/oppenheimer-draws-debate-absence-japanese-bombing-victims-film-rcna96279>. Citado na página 5.

YANG, J. et al. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, ACM New York, NY, v. 18, n. 6, p. 1–32, 2024. Citado na página 11.

YANG, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. Disponível em: <http://arxiv.org/abs/1906.08237>. Citado na página 23.

YOUNUS, A.; QURESHI, M. A. A framework for sexism detection on social media via byt5 and tabnet. Technological University Dublin, 2022. Citado 3 vezes nas páginas 31, 37 e 38.

ZAMPIERI, M. et al. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 75–86. Disponível em: <https://aclanthology.org/S19-2010>. Citado 2 vezes nas páginas 4 e 5.

ZHANG, Z.; ROBINSON, D.; TEPPER, J. Detecting hate speech on twitter using a convolution-gru based deep neural network. In: SPRINGER. *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. [S.l.], 2018. p. 745–760. Citado 2 vezes nas páginas 33 e 37.

ZHAO, W. X. et al. A survey of large language models. *ArXiv*, abs/2303.18223, 2023. Citado na página 23.

ZHAO, Z.; ZHANG, Z.; HOPFGARTNER, F. A comparative study of using pre-trained language models for toxic comment classification. In: *Companion Proceedings of the Web Conference 2021*. [S.l.: s.n.], 2021. p. 500–507. Citado 2 vezes nas páginas 4 e 5.

ZHU, B. et al. Xtab: Cross-table pretraining for tabular transformers. *arXiv preprint arXiv:2305.06090*, 2023. Citado na página 24.