UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação de Ideação Suicida em Postagens de Mídias Sociais

Hiparco Lins Vieira

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Hiparco Lins Vieira

Classificação de Ideação Suicida em Postagens de Mídias Sociais

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Diego Raphael Amancio

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Lins Vieira, Hiparco
L657c Classificação de Ideação Suicida em Postagens de Mídias Sociais / Hiparco Lins Vieira; orientador Diego Raphael Amancio. -- São Carlos, 2025.
64 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Ideação suicida. 2. Mídias sociais. 3. Processamento de linguagem natural. 4. Grandes modelos de linguagem. 5. Aprendizado de máquina. I. Amancio, Diego Raphael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Hiparco Lins Vieira

Classificação de Ideação Suicida em Postagens de Mídias Sociais

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Original version

São Carlos

2025

Este trabalho é dedicado à minha família, aos amigos, colegas e interessados nos assuntos de Inteligência Artificial e Processamento de Linguagem Natural.

AGRADECIMENTOS

Agradeço a Deus, à minha família e aos meus amigos pelos aprendizados que tive em minha vida.

"Nossas dúvidas são traidoras e nos fazem perder o bem que poderíamos conquistar se não fosse o medo de tentar."

O Menestrel - William Shakespeare

RESUMO

Vieira, H. L. **Classificação de Ideação Suicida em Postagens de Mídias Sociais**. 2025. 64 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Pensamentos suicidas representam um sério desafio para a saúde coletiva, exibindo taxas de óbito elevadas entre a população jovem e gerando um considerável efeito negativo na sociedade. Frente às limitações dos modelos comuns de análise médica e ao uso cada vez maior das redes sociais para manifestar emoções e atitudes ligadas ao suicídio, a análise automática de postagens na internet se mostra como uma opção interessante para auxiliar ações preventivas no combate ao suicídio. Neste sentido, este trabalho investiga a classificação automática de postagens em redes sociais com o objetivo de identificar manifestações de ideação suicida, explorando técnicas de processamento de linguagem natural, aprendizado de máquina, aprendizado profundo e grandes modelos de linguagem. Foram comparadas duas abordagens distintas: uma baseada na representação vetorial de postagens utilizando o Sentence-BERT (SBERT) combinada a regressão logística e outra na utilização de grandes modelos de linguagem (LLMs) considerando diferentes estratégias de *prompting*. Os experimentos demonstraram que o classificador com SBERT apresentou desempenho equilibrado entre precisão e revocação, atingindo acurácia superior a 93% nas três métricas. Equanto isto, a abordagem com LLM mostrou resultados com menor precisão, mas maior sensibilidade, alcançando revocação próxima a 98% em seu melhor cenário, o que evidencia a capacidade do modelo de identificar praticamente todos os casos positivos, ainda que ao custo de mais falsos positivos. Desta forma, os resultados indicam que as duas estratégias apresentam potenciais complementares: o SBERT garante maior robustez e estabilidade nos resultados, enquanto a LLM mostra um potencial maior de mitigação de risco, apontando para possibilidades de integração futura entre os métodos em sistemas de apoio à detecção precoce de ideação suicida em redes sociais.

Palavras-chave: Ideação suicida. Mídias sociais. Processamento de linguagem natural. Grandes modelos de linguagem. Aprendizado de máquina.

ABSTRACT

Vieira, H. L. **Classification of Suicidal Ideation in Social Media Posts**. 2025. 64 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Suicidal thoughts pose a serious public health challenge, with high mortality rates among young people and a particularly negative impact on society. Given the limitations of common medical analysis models and the increasing use of social media to express emotions and attitudes related to suicide, the automatic analysis of online posts presents an interesting option for supporting preventive measures in the fight against suicide. Therefore, this work investigates the automatic classification of social media messages to identify manifestations of suicidal ideation, exploring techniques from natural language processing, machine learning, deep learning, and large language models. Two distinct approaches were compared: one based on the vector representation of messages using Sentence-BERT (SBERT) combined with logistic regression, and the other using large language models (LLMs) considering different prompting strategies. The experiments presented, in which the SBERT classification system demonstrated balanced performance between solutions and recall, achieving accuracy above 93% in all three metrics. In this regard, the LLM approach showed results with lower precision but higher sensitivity, achieving recall close to 98% in its best-case scenario, demonstrating the model's ability to identify virtually all positive cases, albeit at the cost of more false positives. Thus, the results indicate that the two strategies have complementary potential: SBERT ensures greater robustness and stability in results, while LLM shows greater risk mitigation potential, pointing to possibilities for future integration between the methods in systems supporting the early detection of suicidal ideation on social media.

Keywords: Suicidal ideation. Social media. Natural language processing. Large language models. Machine learning.

LISTA DE FIGURAS

Figura 1 – Arquitetura do transformer. Fonte: adaptado de (Vaswani *et al.*, 2017) 36
Figura 2 – Arquitetura do SBERT para o cálculo de notas de similaridade. Fonte: Adaptada de (Reimers; Gurevych, 2019) 39
Figura 3 – Histograma da quantidade de tokens nas postagens da classe com ideação suicida. Ocorrências superiores a 500 tokens foram agrupados para melhor visualização. 44
Figura 4 – Histograma da quantidade de tokens nas postagens da classe sem ideação suicida. Ocorrências superiores a 500 tokens foram agrupados para melhor visualização. 44
Figura 5 – Boxplot mostrando a distribuição da contagem de tokens das duas classes. Os *outliers* não foram postados para facilitar a visualização. . . 45
Figura 6 – Gráfico de barras com as palavras mais frequentes na classe com ideação suicida. 46
Figura 7 – Gráfico de barras com as palavras mais frequentes na classe sem ideação suicida. 46
Figura 8 – Matriz de confusão obtida a partir da codificação utilizando SBERT e regressão logística. 50
Figura 9 – Matriz de confusão para o prompt A 53
Figura 10 – Matriz de confusão para o prompt B 53

LISTA DE TABELAS

Tabela 1 – Top 10 palavras mais frequentes na classe com ideação suicida 47
Tabela 2 – Top 10 palavras mais frequentes na classe sem ideação suicida 47
Tabela 3 – Prompts testados para classificação de ideação suicida. 51
Tabela 4 – Resultados de classificação por prompt 52
Tabela 5 – Exemplos postagens retiradas dos dados de validação e suas respectivas classificações (S - suicida, NS - não suicida) a partir do prompt A (PA), prompt B (PB) e SBERT + classificador (SBERT). 55
Tabela 6 – Comparativo entre as técnicas propostas e modelos de referência. . . . 56

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
BERT	*Bidirectional Encoder Representations from Transformers*
BOW	*Bag of Words*
CBOW	*Continuous Bag of Words*
CNN	*Convolutional Neural Networks*
GBDT	*Gradient Boosted Decision Trees*
GPT	*Generative Pre-trained Transformer*
IA	Inteligência Artificial
KNN	*K-Nearest Neighbors*
LLM	*Large Language Model*
LSTM	*Long Short-Term Memory*
MLM	*Masked Language Model*
NLP	*Natural Language Processing*
NSP	*Next-Sentence Prediction*
OMS	Organização Mundial de Saúde
PLN	Processamento de Linguagem Natural
SBERT	*Sentence BERT*
SVM	*Support Vector Machine*
URL	*Uniform Resource Locator*

SUMÁRIO

1 INTRODUÇÃO 25
1.1 Objetivos 29
1.2 Contribuições 29
1.3 Divisão do Trabalho 29

2 ANÁLISE DE SENTIMENTOS 31
2.1 Pré-processamento de dados textuais 31
2.1.1 Padronização das letras 32
2.1.2 Tratamento da pontuação 32
2.1.3 Remoção de URLs 32
2.1.4 Tratamento de emojis 33
2.1.5 Remoção de *stopwords* 33
2.1.6 Remoção de números 33
2.1.7 Tokenização 33
2.2 Vetorização de textos 34
2.3 Arquitetura Transformer 34
2.4 Modelos baseados em Encoder 37
2.4.1 BERT 37
2.4.2 Sentence BERT 38
2.5 Modelos Generativos 40
2.5.1 Modelos baseados em decodificação 40
2.5.2 Configuração dos modelos 41

3 CONJUNTO DE DADOS 43
3.1 Tamanho das postagens 43
3.2 Frequência de palavras 45

4 EXPERIMENTOS E RESULTADOS 49
4.1 Classificação de Ideação Suicida Utilizando SBERT 49
4.1.1 Experimentos 49
4.1.2 Resultados - SBERT 49
4.2 Classificação de Ideação Suicida Utilizando LLM 51
4.2.1 Configuração da LLM 51
4.2.2 Resultados - LLM 52
4.3 Exemplos 54
4.4 Comparativo com modelos de referência 54

5 CONSIDERAÇÕES FINAIS 57

REFERÊNCIAS 59

1 INTRODUÇÃO

A depressão é um transtorno psiquiátrico de humor que leva a sentimentos de tristeza, desesperança, inutilidade, diminuição de energia, perda de interesse e mudanças no apetite (Balcı; Saraç, 2024; Benjachairat *et al.*, 2024). A depressão pode variar em intensidade, sendo usualmente categorizada como leve, moderada ou grave (Benjachairat *et al.*, 2024). Na intesidade moderada ou grave, a depressão pode levar à automutilação e a pensamentos suicidas.

Uma das ferramentas clínicas para a avaliação do risco de uma pessoa cometer um suicídio é a escala de ideação suicida (Allarakha; Uttekar, 2021). A ideação suicida é um termo abrangente e inclui diversos pensamentos e comportamentos, tais como pensamentos de tirar a própria vida, ameaças de suicídio e tentativas de suicídio.

A ideação suicida é uma das principais causas de morte em jovens entre 10 e 24 anos e se tornou uma questão de grande preocupação para professores, psicólogos e clínicos que lidam com adolescentes em risco de suicídio (Zhong *et al.*, 2023). A ideação resulta de fatores psicológicos e psicossociais. Fatores como bullying, dificuldades sociais, desajustes falimiares e experiências traumáticas durante a infância estão associados a um aumento do risco de suicídio (Polanco-Roman *et al.*, 2021), sendo a ideação e o pensamento suicida preditores do suicídio (Zhong *et al.*, 2023).

Em 2024, a Organização Mundial de Saúde (OMS) estimou que mais de 720 mil morrem vítimas de suicídio todos os anos (Organization, 2024). Estes números mostram a urgência para adoção de medidas de detecção e prevenção do suicídio. A OMS sugere medidas eficientes para prevenção e controle do suicídio, tais como: limitar o acesso aos meios de suicídio como armas, por exemplo; promover habilidades de vida socioemocionais em adolescentes; identificar, avaliar, gerenciar e acompanhar precocemente qualquer pessoa afetada por comportamentos suicidas e interagir com a mídia para reportar suicídio de forma responsável.

A prevenção do suicídio requer a identificação de indivíduos em risco e o fornecimento de cuidados clínicos e psicológicos adequados (Naseem *et al.*, 2024). Contudo, os sistemas tradicionais de assistência à saúde mental frequentemente enfrentam desafios, tais como elevados custos e escassez de recursos, dificultando a prestação de serviços de saúde mental (Omar *et al.*, 2024). Além disto, apesar da maior conscientização sobre o suicídio, os métodos para diagnosticar ideação suicida e medir o risco em pacientes continuam desafiadores (Franklin *et al.*, 2017). Estudos apontam que 60% das pessoas que morreram por suicídio negaram ter tentado suicídio a especialistas em saúde mental (McHugh *et al.*, 2019).

O desenvolvimento das mídias sociais tem trazido caminhos alternativos para a prevenção do suicídio. As mídias sociais têm sido um meio onde as pessoas compartilham seus pensamentos com frequência (Islam *et al.*, 2023). Indivíduos podem recorrer às redes sociais como uma alternativa para conversarem, expressarem seus sentimentos, serem ouvidos, receber atenção e ajuda (Benjachairat *et al.*, 2024).

Pesquisas sobre detecção de depressão em redes sociais indicam que usuários com inclinações à depressão mostram uma afinidade evidente em relação a motivações pessimistas (Vandana; Marriwala; Chaudhary, 2023). Além disto, 80% das pessoas que expressam ideação suicida o fazem por meio das mídias sociais (Robinson *et al.*, 2015) e mais de 20% dos indivíduos que tentam suicídio e 50% dos que o completam deixam uma mensagem em mídias sociais (DeJong; Overholser; Stockmeier, 2010; Ji *et al.*, 2021)

De forma recíproca, outros indivíduos podem consumir o conteúdo mencionado anteriormente e aprender sobre os sentimentos das outras pessoas. Este conteúdo, que inclui textos, mensagens, postagens, tweets, entre outros, é armazenado (Choudhury *et al.*, 2016) e pode ser explorado para o desenvolvimento de tecnologias assistivas de prevenção ao suicídio, sinalização de ideação suicida e na identificação e apoio de indivíduos vulneráveis (Benjachairat *et al.*, 2024; Sofia *et al.*, 2023). Entretanto, a análise conteúdo de ideação suicida em textos nas redes sociais é desafiadora, devido a fatores como a ausência de palavras suspeitas, sinais incomuns e informações detalhadas sobre um indivíduo. (Islam *et al.*, 2023)

Por outro lado, as pesquisas no campo de processamento de linguagem natural e grandes modelos de linguagem têm tido avanços extraordinários, fornecendo aos computadores uma ampla gama de técnicas para representar, compreender e analisar a linguagem humana (Kruthika *et al.*, 2024). Isto possibilita uma compreensão mais profunda dos textos ao permitir a extração de nuances e contexto, estimulando o desenvolvimento de soluções para diversos problemas, dentre os quais destacam-se aqueles relacionados a detecção, avaliação de risco e prevenção do suicídio (Elyoseph; Shoval; Levkovich, 2024).

Ji *et al.* (Ji *et al.*, 2018) analisa conteúdo online (Reddit SuicideWatch and Twitter) para detectar ideação suicida através de aprendizado supervisionado. Foram extraídos dos dados características estatísticas (número de palavras, *tokens* e caracteres), classes gramaticais, características linguísticas (utilizando *Linguistic Inquiry and Word Count*), frequência de palavras, *word embeddings* (word2vec) e características de tópicos utilizando latent Dirichlet allocation. Com isso, foram treinados seis classificadores: máquina de vetores suporte (SVM), *random forest*, *gradient boosted decision trees* (GBDT), XGBoost, redes neural *feedforward* e rede neural LSTM (Long Short-Term Memory). Os resultados mostraram o forte potencial das técnicas combinadas para a classificação de ideação suicida.

O'Dea *et al.* (O'Dea *et al.*, 2015) examinou a possibilidade de se determinar o

nível de preocupação sobre postagens relacionadas ao suicídio no Twitter usando apenas o conteúdo postado e aprendizado de máquina. Dois algoritmos foram explorados: SVM e regressão logística. A pesquisa demonstrou que é possível distinguir o nível de preocupação entre tweets relacionados a suicídio usando humanos e aprendizagem de máquina. Os autores destacam que o classificador desenvolvido replicou dos codificadores humanos participantes da pesquisa e confirmaram que o Twitter é usado por indivíduos para expressar tendências suicidas.

Tadesse *et al.* (Tadesse *et al.*, 2019) propôs uma solução usando aprendizado profundo para detecção de ideação suicida em mídias sociais (Reddit). A arquitetura apresentada possui uma camada de *word embedding* (word2vec), seguida por camadas LSTM e CNN. Os resultados obtidos mostraram um desempenho superior da arquitetura LSTM-CNN quando comparado com CNN, LSTM, SVM, random forest, XGBoost e Naïve Bayes.

Wang *et al.* (Wang *et al.*, 2021) apresenta uma arquitetura de aprendizado profundo para predição de suicídio dentro de trinta dias e seis meses usando dados de postagens em mídias sociais. Utilizou-se a técnica Doc2vec para a criação de embeddings de palavras e documentos. Além disto, também foram extraídas emoções e classes gramaticais. A rede neural utilizada foi uma rede reural *C-attention*. A rede neural foi comparada com outras técnicas de aprendizado de máquina, tais como SVM, regressão logística, random forest, K-nearest neighbors (KNN) e árvore de decisão. Os resultados mostraram que o modelo C-attention apresenta melhores resultados para predições de suicídio em tempos mais longos, enquanto o KNN e a SVM apresentam melhores resultados em tempos mais curtos.

Bayham e Benhiba (Bayram; Benhiba, 2021) exploraram técnicas de aprendizado de máquina (KNN, SVM, ensemble e regressão logística) para identificar pessoas que tentaram suicídio baseado nas postagens do Twitter ocorridas 30 e 182 dias antes do incidente. Os melhores resultados foram alcançados com o modelo de ensemble. Segundo a pesquisa, o desempenho do modelo foi consideravelmente melhor para predições de curto prazo, levando a crer que os sinais de ideação suicida nas postagens são mais fortes próximo ao incidente.

Chatterjee *et al.* (Chatterjee *et al.*, 2022) treinaram modelos de aprendizado de máquina a partir de um conjunto de dados de mensagens obtidas do Twitter e Reddit. Os dados analisados caracterizaram sintomas suicidas no conteúdo do tweet, estatísticas de emoticons e sentimento do conteúdo, perfil do usuário, traços de personalidade e características temporais das postagens. Os melhores resultados foram obtidos usando regressão logística, embora técnicas como random forest, SVM e XGBoost também tenham sido usadas.

Aldhyani *et al.* (Aldhyani *et al.*, 2022) propôs um sistema de classificação de ideação suicida a fim de classificar postagens em mídias sociais como suicidas e não suicidas. Os

dados utilizados foram postagens do Reddit obtidos no Kaggle. Duas abordagens foram consideradas: CNN-BiLSTM e XGBoost. A partir de características textuais, os autores obtiveram uma acurácia de 95% com a CNN-BiLSTM, mostrando o potencial dos algoritmos de aprendizagem de máquina e aprendizado profundo na para a resolução do problema em questão.

Srinivasarao *et al.* (Srinivasarao *et al.*, 2024) propôs uma metodologia baseada em N-grams, *word2vec* e LSTMs para identificar inclinções suicidas em postagens no Twitter. Uma metodologia similar foi proposta por Singh *et al.* (Singh *et al.*, 2024), que utilizou uma arquitetura neural com camadas convolucionais e LSTMs para classificar as postagens. Allam *et al.* (Allam *et al.*, 2025) extraiu as características do texto utilizando técnicas de contagem, TF-IDF, tokenização, *stemming* e estas informações foram utilizadas no treinamento de uma *random forest*.

Setiawan *et al.* (Setiawan *et al.*, 2024) propôs uma metodologia para o treinamento de um classificador de tweets no idioma indonésio. Para isto, os autores treinaram oito modelos com quatro variantes do modelo IndoBERT através de transferência de aprendizado, combinados com redes LSTM e CNN, ambas ajustadas para classificar ideação suicida.

Técnicas baseadas em aprendizado de máquina e aprendizado profundo não são as únicas a serem utilizadas em análises de comportamento e ideação suicídas. Alguns autores realizam análise de texto relacionado ao suicídio utilizando redes complexas, visto que estas permitem a visualização e quantificação das relações complexas entre múltiplos sintomas psicossociais e psicológicos e sua influência na ideação suicida (Beurs, 2017).

Teixeira *et al.* (Teixeira *et al.*, 2021) utilizou ciência cognitiva de redes para identificar a estrutura semântica e emocional de cartas reais de suicídio. Esta abordagem apresentou um diferencial em relação aos modelos de caixa preta, possibilitando a extração de ideas e emoções-chave contidos no texto de forma transparente e facilitando a interpretação dos resultados.

Tayim *et al.* (Tayim *et al.*, 2025) utilizou a análise de redes para identificar as maneiras sutis pelas quais diferentes tipos de violência conjugal em mulheres árabes estão relacionados com a ideação suicida. Para isto, os autores realizaram uma análise transversal com rede não direcionada e não causal, além de medidas de centralidade para analisar dados coletados de entrevistas realizadas nas mídias sociais.

Desta forma, evidencia-se que uma das maiores dificuldades na prevenção ao suicídio é unificar diferentes fontes de informações em uma base de informações que pode ser explorada para identificação de ideação suicida.

1.1 Objetivos

O objetivo geral deste trabalho é o desenvolvimento de um classificador de postagens no Twitter para a detecção de ideação suicida. Este objetivo está dividido nos seguintes objetivos específicos:

1. analisar o conjuto de dados de ideação suicida do Kaggle (Komati, 2020);
2. realizar a limpeza dos dados para treinamento de classificadores;
3. realizar a extração de características do texto;
4. treinar um modelo de aprendizado profundo para classificar as postagens;
5. avaliar a qualidade dos modelos treinados.

1.2 Contribuições

As principais contribuições deste trabalho estão no desenvolvimento e análise de técnicas para a prevenção do suicídio em redes sociais.

1.3 Divisão do Trabalho

O trabalho está dividido da seguinte forma:

- o Capítulo 2 descreve os principais conceitos utilizados para o desenvolvimento do trabalho;
- o Capítulo 3 apresenta o conjunto de dados utilizado nos experimentos;
- o Capítulo 4 detalha os experimentos e resultados obtidos;
- o Capítulo 5 tece as considerações finais do trabalho.

2 ANÁLISE DE SENTIMENTOS

A análise de sentimentos é uma subárea do processamento de linguagem natural e tem como objetivo a classificação e categorização automática de emoções, opiniões ou atitudes em textos escritos. Esta área se tornou uma ferramenta essencial para compreender os sentimentos presentes em textos de diversos domínios (Tan; Lee; Lim, 2023).

O problema de classificação de textos envolve em atribuir uma classe ou rótulo a um texto. No problema de análise de sentimentos, essa classe é um sentimento, que pode ser "positivo", "negativo", "alegria", "raiva"etc. Para a realização desta tarefa, são treinados classificadores.

Conforme mencionado anteriormente, alguns classificadores utilizam técnicas clássicas de aprendizado de máquina, tais como máquinas de vetores suporte (Cristianini; Shawe-Taylor, 2000), Naïve Bayes (Raschka, 2014), regressão logística (James *et al.*, 2021) e árvores de decisão (Breiman *et al.*, 2017). Outros classificadores são baseados em aprendizado profundo e utilizam diversas estratégias e arquiteturas de redes neurais para resolver o problema, tais como as redes perceptron de múltiplas camadas, LSTMs, redes baseadas em transformadores e grandes modelos de linguagem (LLMs).

As metodologias de análise de sentimentos abrangem várias etapas essenciais, tais como: (1) pré-processamento, (2) vetorização de textos e (3) classificação. Na primeira etapa, os dados de texto passam por uma limpeza para fins de padronização e/ou removeção de informações irrelevantes. Nesta etapa, pode-se remover palavras que possuem pouca contribuição para o sentido do texto (conhecidas como *stopwords*), caracteres especiais, pontuação, números, emojis, links etc. A etapa de extração de características ocorre após o pré-processamento. Este etapa é responsável pos transformar o texto em um formato compreensível pelo computador, usualmente na forma de vetores numéricos. Dentre as principais técnicas destacam-se: *Bag of Words* (BOW), word2vec e embeddings baseados em *transformers*, tais como o BERT, que será apresentado ao longo deste capítulo.

A terceira etapa consiste em utilizar as características para treinamento de modelos de aprendizado de máquina ou apendizado profundo. Este capítulo apresenta um resumo destas três etapas e as principais técnicas exploradas ao longo desta pesquisa.

2.1 Pré-processamento de dados textuais

Dados textuais usualmente necessitam de um pré-processamento antes de serem inseridos em algoritmos de aprendizado. O pré-processamento é importante para padronizar o texto, colocando-o em um formato que permita com que os algoritmos de aprendizado extraiam padrões da melhor forma possível. Dentre as técnicas de pré-processamento,

destacam-se: (1) a padronização para letras minúsculas, (2) o tratamento da pontuação, (3) a remoção de URLs, (4) o tratamento de emojis, (5) a remoção de *stopwords*, (6) remoção de números e (7) a tokenização. O leitor pode encontrar mais informações sobre técnicas de pré-processamento de textos para análise de sentimentos em (Duong; Nguyen-Thi, 2021).

2.1.1 Padronização das letras

Os computadores interpretam letra maiúsculas e minúsculas de forma distinta. Por exemplo, as palavras "Universidade", "universidade"e "UNIVERSIDADE"possuem a mesma semântica, mas podem gerar vocabulários distintos. Caso isso não seja ajustado, os algoritmos de aprendizado podem tratar as três palavras de forma separada, aumentando desnecessariamente o tamanho do vocabulário e dificultando o aprendizado e a convergência.

Desta forma, na etapa de pré-processamento dos dados pode ser importante padronizar a capitalização das letras para minúscula, evitando-se, assim, problemas de aprendizado e generalização.

2.1.2 Tratamento da pontuação

A pontuação também pode levar a problemas de vocabulário excessivo, aprendizado e convergência. Expressões como "vamos", "vamos!", "vamos?"e "vamos??"podem ser consideradas vocabulários distintos e, a depender dos dados, o modelo pode enfrentar dificuldades para diferenciar umas das outras. Com isto, a pontuação pode ser removida em problemas onde não contribua de forma significativa.

Vale destacar que em casos de análise de sentimento em redes sociais, a pontuação é comumente utilizada para aumentar a intensidade das emoções. Neste caso, expressões como 'Corra.' e 'Corra!!!', podem transmitir ideias semelhantes, mas diferentes em intensidade ou emoção.

Percebe-se que este a estratégia de tratamento da pontuação dependerá do problema e da metodologia adequada para trainemento dos modelos.

2.1.3 Remoção de URLs

A presença de links nas postagens de redes sociais é relativamente comum. É importante remover estes links do texto quando estes não apresentam contribuição semântica significativa para a tarefa, a fim de que não sejam tratados como parte do vocabulário e as palavras que o compõem venham a se tornar ruído nas frases.

Muitos links podem ter palavras únicas. Caso o link não seja tratado adequadamente, ele pode se tornar parte do texto sem agregar significado. Isso pode levar a um aumento do vocabulário e do custo computacional para o treino do modelo sem levar a resultados melhores.

2.1.4 Tratamento de emojis

Emojis são amplamentes utilizados em redes sociais como uma forma de expressão direta de sentimentos e emoções. No âmbito de análise de sentimentos, o tratamento de emojis adequado, como por exemplo substituir " : @" por "$emoji_raiva$", pode trazer ganhos de desempenho ao proporcionar aos algoritmos uma forma adicional de representar a informação.

Outra vantagem a que estes podes auxiliar a entender ironia e sarcasmo. Por exemplo, a expressão "Parabéns!"pode indicar um sentimento de alegria ou de ironia, a depender do emoji que a acompanha.

Portanto, o tratamento de emojis em problemas de análise de sentimentos pode elevar consideravelmente as taxas de sucesso do modelo treinado.

2.1.5 Remoção de *stopwords*

Stopwords são palavras comuns nas línguas e carregam pouco significado em si (Saif *et al.*, 2014). As *stopwords* variam de idioma para idioma. Cada idioma possui uma lista pré-definida de *stopwords* que é utilizada no tratamento de textos. Para tarefas onde o foco seja na análise de termos do texto, a remoção das *stopwords* auxilia na simplificação do problema, remoção de ruído e redução do tempo de treinamento. Contudo, a remoção de stopwords não é adequada para problemas que exijam análise semântica, visto que há perda de informação e contexto.

2.1.6 Remoção de números

Em problemas de análise de sentimentos, normalmente os números não carregam informações relevantes sobre sentimentos. Com isso, a fim de se reduzir a geração de vocabulário desnecessário a partir de números, remove-se os números do textos. Um ponto de atenção é que esta etapa de pré-processamento deve ser realizada após o tratamento de emojis, visto que alguns deles são construídos com números, como ":3"e "8|", por exemplo.

2.1.7 Tokenização

Um texto não pode ser utilizado diretamente no treinamento de modelos, visto que o modelo não entende diretamente a estrutura. Para que o texto se torne compreensível, é necessário realizar uma transformação deste texto para uma estrutura adequada. Uma das formas de se fazer isto é separando o texto em componentes menores, podendo estes serem palavras, subpalavras ou até mesmo caracteres. Estes componentes são chamados tokens.

Os tokens podem ser expressos através de números, possibilitando o aprendizado da relação entre tokens através de modelos algébricos e, através de técnicas de vetorização, serem codificados na forma de *embeddings*.

2.2 Vetorização de textos

Uma das técnicas mais simples vetorização de textos é a *Bag of Words*. Nesta técnica, gera-se uma lista de todas as palavras únicas do texto e associa-se cada palavra a contagem sua respectiva quantidade de ocorrências no texto (Ekman, 2021). Em caso de comparação de similaridaed entre documentos, pode-se utilizar o vetor de cada documento e comparar utilizando métricas de similaridade. No caso de classificação, o vetor pode ser utilizado como entrada para algoritmos de aprendizado de máquina ou até mesmo de aprendizado profundo.

Um dos pontos negativos do BOW é que este desconsidera a gramática e a ordem com que as palavras aparecem nos no texto (Abubakar; Umar, 2022), levando a uma perda de informação e dificultando a sua aplicação em problemas onde seja necessária a extração de características semânticas do texto.

Outra técnica que proporcionou avanços em relação ao BOW é o Word2Vec (Mikolov *et al.*, 2013), um algoritmo para transformar palavras em vectores densos e contínuos (embeddings). O Word2Vec foi treinado utilizando-se duas tasks: o *Continuous Bag of Words* (CBOW) e o *Continuous Skip-Gram*. A primeira tarefa é a de prever uma palavra utilizando as palavras a sua volta, como, por exemplo, prever a palavra "aluno"a partir do contexto "o ? estuda muito". No caso do modelo skip-gram, busca-se prever tem-se como entrada "aluno"e deve-se prever as palavras ao redor "o", "estuda"e "muito".

Esses modelos levaram ao treino de uma camada de *embedding* com diversos ganhos em relação ao BOW, tais como alguma capacidade de entender as relações semânticas e sintáticas entre as palavras. Neste caso, palavras com significados próximos possuem embeddings também próximos (Ekman, 2021). Estes embeddings podem ser utilizados para a resolução de outros problemas, dispensando, neste caso, o treinamento de embeddings do zero.

Uma das desvantagens do Word2Vec é que ele representa diferentes significados de uma mesma palavra com o mesmo vetor. Outro ponto é que o algoritmo apenas visualiza o contexto, independentemente da ordem com que este contexto se apresente. Além disto, o Word2Vec não consegue inferir um embedding para palavras fora do vocabulário nem representar bem o significado de palavras raras no seu corpus de treinamento (Kumar, 2025).

2.3 Arquitetura Transformer

Antes de 2017, as principais soluções para processamento de linguagem natural eram baseadas na arquitetura encoder-decoder utilizando redes recorrentes (Cho *et al.*, 2014; Sutskever; Vinyals; Le, 2014; DataCamp, 2024). Os modelos recorrentes possuem a desvantagem de processar os estados de forma sequencial (token a token), impedindo a

paralelização do treino e da inferência, reduzindo a eficiência.

Porposto em 2017, o *transformer* é uma técnica baseada em atenção que apresentou ganhos significativos em relação aos modelos recorrentes (Ekman, 2021; Vaswani *et al.*, 2017), apesar de manter a característica encoder-decoder. A ideia por trás do mecanismo de atenção é deixar que a rede neural ou parte dela decida em qual/quais parte(s) dos dados de entrada a sua atenção deve ser focada.

O *transformer* foi inicialmente proposto para tarefas de tradução automática, onde insere-se um texto em inglês, por exemplo, e obtém-se a tradução deste texto para o francês. A arquitetura, exibida na Figura 1 é dividida em dois blocos principais: encoder e decoder. A entrada (texto a ser traduzido) é inserida no encoder que transmite a informação codificada ao decoder que gera a saída da rede (texto traduzido). É comum que se utilizem vários encoders e decoders em série, a fim de proporcionar uma melhor capacidade de aprendizado ao modelo.

O encoder visa transformar os tokens de entrada em representações que levam em consideração a semântica do texto. O encoder da arquitetura Transformer é constituído por três componentes principais. Inicialmente, os tokens de entrada são convertidos em representações vetoriais densas por meio da camada de input embeddings. Contudo, como o mecanismo de atenção não possui informação sobre a ordem dos tokens, é necessário incorporar informações posicionais. Para isto, os embeddings são somados a uma codificação posicional, utilizando funções senoidais e permitindo que o transformer distingua a posição relativa de cada token recebido.

Após essa etapa, a soma dos embeddings com informação posicional é processada por uma pilha de camadas de encoder. Cada camada segue a mesma estrutura, composta por dois subblocos organizados sequencialmente.

O primeiro subbloco é o mecanismo de atenção multi-cabeças. Este permite que os modelos relacionem cada palavra da entrada com outras palavras, escolhendo a qual token a sua atenção deve ser direcionada. Ao final deste subbloco, é adicionada uma camada de soma residual e normalização.

O segundo subbloco é uma rede neural totalmente conectada, que visa refinar a informação recebida do primeiro bloco. Esta rede é composta por duas camadas lineares separadas por uma função de ativação ReLU. O resultado é adicionado com uma conexão residual e normalizado a fim de garantir melhores gradiantes e estabilidade durante o treinamento.

O *decoder* tem como função a criação de sequências de texto (tokens) e opera de forma regressiva, iniciando o seu processo com um token de início e utilizando as saídas anteriores como entrada. Com isto, percebe-se que o mesmo utiliza tanto informações do *encoder* quanto de saídas passadas. Este processo se repete até que o decoder gere o token

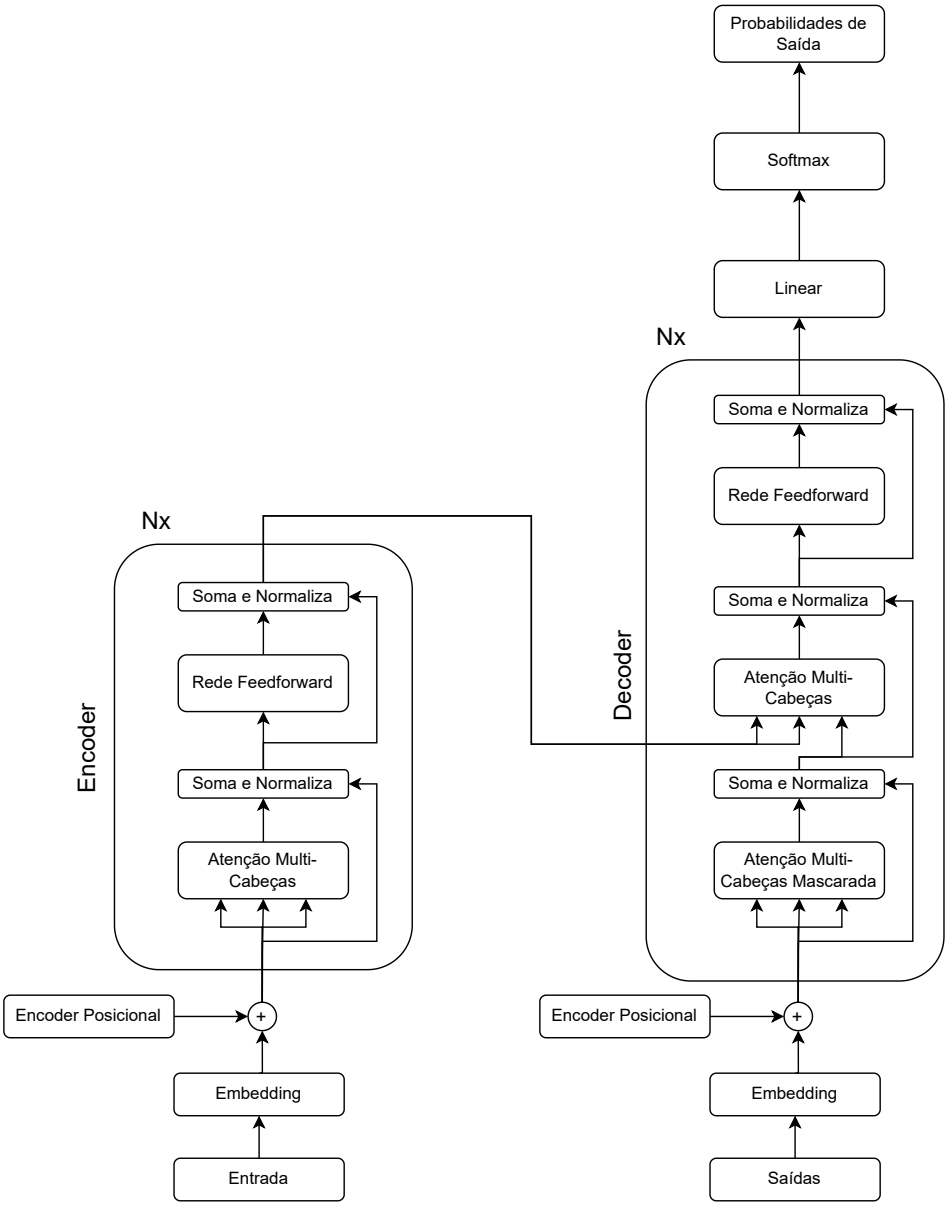

Figura 1 – Arquitetura do transformer. Fonte: adaptado de (Vaswani *et al.*, 2017)

de fim de geração de texto.

A a estrutura do *decoder* é similar ao do *encoder*, possuíndo dois subblocos de atenção com múltiplas cabeças, sendo uma delas mascarada, e um subbloco de rede neural totalmente conectada. Assim como no encoder, no final de cada subbloco há uma camada de soma residual e normalização.

Conforme mencionado anteriormente, o *decoder* recebe as saídas como entradas. Estas entradas passam por um processo de codificação, cujo resultado é somado a uma codificação posicional. Em seguida, esta soma é inserida no primeiro subbloco de atenção mascarada de multiplas cabeças. Este subbloco é semelhante ao mecanismo atenção do codificador, exceto pelo fato de que ele impede que cada palavra na sequência seja influenciada por tokens futuros. Em seguida, soma-se o resutlado a uma conexão residual e normaliza-o. A saída deste subbloco é conectada à entrada do segundo subbloco, de atenção de múltiplas cabeças.

No segundo subbloco, além da entrada do subbloco anterior, insere-se como entrada a saída do *encoder*. Isso permite que o *decoder* dê processe as entradas provenientes do *encoder* e esolha onde deve ficar a sua atenção. A saída deste, após adição residual e normalização, é inserida no terceiro subbloco, de rede neural totalmente conectada. Estes dois subblocos funcionam da forma quase idêntica ao *encoder*. A saída do terceiro subbloco é inserida em um classificador linear e, em seguida, uma camada de softmax. Isso permite que o transformer selecione uma das palavras do vocabulário como saída, que é posteriormente reinserida no *decoder* até que se obtenha o token de finalização.

2.4 Modelos baseados em Encoder

Esta seção descreve os seguintes modelos de linguagem baseados em *encoding*: BERT, DistilBERT e Sentence BERT. Estes modelos são treinados com grandes quantidades de dados, mas já estão muito distantes dos modelos de linguagem modernos. Na época em que foram treinados, eram considerados grandes. Contudo, com o avanço dos modelos, eles passaram a ser consideravelmente menores que os grandes modelos de linguagem (*Large Language Models* - LLMs) modernos. Com isto, passaram a ser chamados Modelos de linguagem pré-treinados (Zhao *et al.*, 2023).

2.4.1 BERT

O Bidirectional Encoder Representations from Transformers (BERT) é baseado na arquitetura do encoder dos Transformadores (Devlin *et al.*, 2018). Esta é uma forma de embedding treinada na realização simultânea de duas tarefas: *Masked Language Model* (MLM) e *Next-Sentence Prediction* (NSP).

Semelhante ao treino do Word2Vec, a primeira tarefa busca ensinar ao modelo a

estrutura das frases. Esta tarefa é realizada com a previsão de palavras em uma frase utilizando palavras passadas e futuras. Para isto, as palavras a serem previstas são mascaradas aleatoriamente e as demais são mantidas para serem usadas pelo modelo. Os exemplos de entrada de treino foram formados da seguinte maneira (Ekman, 2021):

- 15% das palavras nas frases de entrada são selecionadas aleatoriamente para serem mascaradas;
- 80% dentre as palavras selecionadas é substituída por um embedding de uma máscara especial;
- 10% dos *embeddings* das máscaras selecionadas são substituidas por um *embedding* máscara de uma palavra selecionada aleatoriamente;
- nos outros 10%, os *embeddings* das palavras não são substituídos. Ao invés disso, usa-se o *embedding* da palavra correta.

Com isso, o modelo busca prever as palavras em uma frase, incluíndo as não mascaradas. O desempenho do treinamento é avaliado nos 15% de palavras mascaradas.

Na tarefa de NSP, busca-se ensinar ao modelo as relações sobre as frases através de um problema de classificação. No treinamento, apresenta-se ao modelo duas frases e o modelo deve determinar se a segunda frase é a sequência lógica da primeira. Para o treinamento desta task, em 50% dos casos apresentados ao modelo as duas frases são uma sequência lógica, enquanto nos demais casos as frases não são uma sequência lógica.

Um ponto importante do BERT é que este usa *wordpieces* (partes de palavras) como tokens ao invés de palavras inteiras. A técnica *wordpieces* foi proposta por (Schuster; Nakajima, 2012). Esta técnica permite ao BERT lidar melhor com lalavras fora do vocabulário, uma vez que uma palavra que não existe no vocabulário é substituída por uma sequência de subpalavras, que podem ser conhecidas, permitindo ao modelo generalizar o conhecimento adquirido com outras palavras semelhantes.

Outro ponto importante é que como o BERT analisa o contexto das palavras, o texto inserido deve ser minimamente preparado, sem remoção de *stopwords*, por exemplo. Isto ocorre porque o BERT consegue aprender e ponderar o peso de cada palavra dentro do contexto e do fluxo lógico de ideias, mesmo para palavras frequentes.

2.4.2 Sentence BERT

O BERT não é adequado para buscas por similaridade semântica e tarefas de agrupamento não-supervisionado devido ao elevado custo computacional. Nesta tarefa, possui-se duas frases e deseja-se determinar o quão similares semanticamente as duas frases são através de alguma métrica de similaridade. Para a realização desta tarefa, o BERT

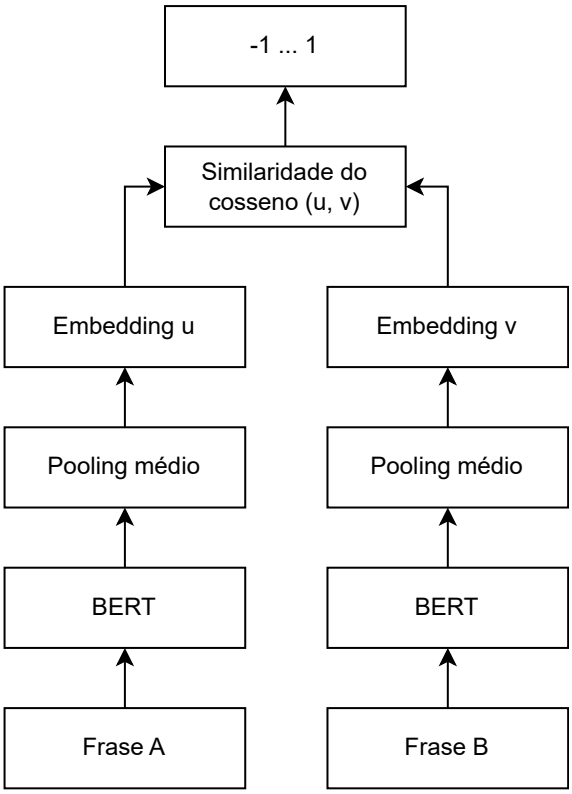

Figura 2 – Arquitetura do SBERT para o cálculo de notas de similaridade. Fonte: Adaptada de (Reimers; Gurevych, 2019)

deve receber as duas frases e gerar os embeddings, que seriam posteriormente enviados para uma camada *feedforward* a fim de se gerar uma nota de similaridade. Contudo, para um conjuto grande, seria necessário rodar o BERT para cada comparação, o que seria inviável.

Outra opção seria gerar *embeddings* utilizando o BERT. Neste caso, o BERT receberia uma frase e geraria *embeddings* de cada palavra na frase. Para que seja feito um único embedding para a frase, o resultado deve ser agrupado. Uma das formas com que isso pode ser feito é utilizando o *mean pooling*, o que leva a resultados fracos. Isso acontece porque o BERT não foi treinado para que esses embeddings fossem utilizados de forma adequada em tarefas de similaridade semântica.

Esse problema levou ao desenvolvimento do Sentence BERT (SBERT ou Sentence Transformer) (Reimers; Gurevych, 2019), cuja arquitetura é apresentada na Figura 2. O SBERT é um fine-tuning do BERT que utiliza estruturas de rede siamesas (ou *triplets*) para gerar embeddings de frases considerando a semântica. Estes *embeddings* podem ser comparados usando métricas de similaridade (como a do cosseno) com um custo computacional muito inferior ao BERT.

Uma das desvantagens que o *sentence transformer* pode enfrentar é como inserir

informação numérica dentro dos embeddings, característica herdada do BERT, que realiza tokenização por subpalavras e baixo desempenho em *embedding* com números (Wallace *et al.*, 2019).

Um ponto de atenção na utilização no SBERT é na codificação de várias frases em conjunto. O SBERT gera um único *embedding* por entrada. Com isto, caso o input contenha múltiplas frases, o *embedding* resultante será um meio termo das frases inseridas. Isso pode levar a uma represntação distorcida em alguns casos e impactar no desempenho de classificadores.

2.5 Modelos Generativos

Esta seção apresenta uma breve apresentação sobre alguns modelos importantes generativos e o Mistral 7B, modelo usado ao longo deste trabalho. Além disto, são descritos alguns dos parâmetros de configuração do modelo.

2.5.1 Modelos baseados em decodificação

Após o sucesso dos modelos baseados em *encoding*, tais como o BERT, a pesquisa da área se voltou fortemente para a área de predição de próxima palavra. Conforme descrito anteriormente, este é majoritariamente o papel do *decoder* do BERT. Como o *decoder* também possui camadas de atenção, ele pode ser adaptado para esta tarefa sem a necessidade de um *encoder*. Desta forma suge, em 2018, o modelo *Generative Pre-Trained Transformer* (GPT), onde retira-se o subbloco de atenção cruzada proveniente do *encoder* e realiza-se um ajuste fino no decodificador para tarefas de NLP (Radford *et al.*, 2018).

Até 2019, tarefas como NLP, tradução, e sumarização eram resolvidas com modelos treinados através de aprendizado supervisionado, utilizando conjuntos de dados específicos para cada tarefa. Em 2019, foi proposto o GPT-2, que trouxe ganhos significativos em relação aos modelos da época. No artigo em que foi apresentado (Radford *et al.*, 2019), os pesquisadores relataram terem treinado uma arquitetura *decoder* consideravelmente maior que o GPT-1, com aproximadamente 1.5 bilhão de parâmetros. Além disto, utilizaram um conjunto muito maior de dados coletados de milhões de páginas da internet, chamado de WebText. Na criação do conjunto de dados, os autores focaram em qualidade e utilizaram apenas páginas do Reddit que foram filtradas por outros humanos e receberam pelo menos três pontos (*karma*). O GPT-2foi capaz de aprender as diversas tarefas sem supervisão em seu treinamento. Quando guiado por documentos e perguntas, o GPT-2 excedeu diversos modelos do estado da arte em diversas tarefas para as quais não recebeu nenhum treinamento específico (*zero-shot*).

Em 2020, foi apresentado o GPT-3 (Brown *et al.*, 2020). O GPT-3 é um modelo de linguagem autorregressivo com 175 bilhões de parâmetros, treinado em com conjuto de dados filtrado de 570GB, obtido de múltiplos domínios para fornecer ao modelo uma

maior diversidade de assuntos. Com nenhum ou alguns exemplos via *prompt*, o modelo conseguiu realizar diversas tarefas de forma coerente em tarefas como tradução, resposta a perguntas, exercícios lógicos e aritméticos. O modelo também foi capaz de gerar amostras de artigos jornalísticos em que humanos não foram capazes de distiguir se a autoria era de um humano.

Apesar do grande avanço obtido com o GPT-3 e de sua base de dados treinada com dados de qualidade, grandes modelos de linguagem geravam saídas falsas, tóxicas ou inúteis para o usuário. Para resolver esse problema, pesquisadores da OpenAI propuseram o InstructGPT (Ouyang *et al.*, 2022). Na proposta, os pesquisadores mostraram que é possível utilizar *feedback* humano para fazer um ajuste fino utilizando aprendizado supervisionado nos modelos. Essa abordagem foi chamada de *Reinforcement Learning from Human Feedback*. O ajuste fino com *feedback* humano levou também a uma melhora na veracidade dos resultados e redução das saídas tóxicas. O resultado do InstructGPT, com 1.3 bilhões de parâmetros, foi considerado melhor que o GPT-3, com 175 bilhões, gerando também ganhos de eficiência.

Após isto, muitos pesquisadores buscaram otimizar a eficiência arquitetural e a escalabilidade dos modelos, permitindo desempenho de ponta em arquiteturas mais compactas. Neste cenário, pesquisadores da Mistral AI propuseram o Mistral 7B, um modelo aberto com apenas 7 bilhões de parâmetros. O modelo é capaz de realizar tarefas como raciocínio, matemática e geração de código comresultados competitivos com modelos substancialmente maiores (Jiang *et al.*, 2023).

2.5.2 Configuração dos modelos

Alguns dos parâmetros de configuração de LLMs são a temperatura e a amostragem do núcleo. Quando a LLM está selecionando o próximo token a ser gerado, ele utiliza uma distribuição de probabilidade em seu vocabulário. Em algumas aplicações, é importante que o modelo gere saídas diferentes para as mesmas entradas, mas ainda considerando as distribuições de probabilidade de seu vocabulário frente às instruções recebidas. Um dos parâmetros que ajusta esta distribuição é a temperatura.

A temperatura é um parâmetros de ajuste (achatamento ou alongamento) dessa distribuição para que aumentar ou diminuir a criatividade do modelo ao selecionar o próximo token. Usualmente, os valores da temperatura variam entre 0.0 e 2.0. Quando menor a temperatura, menos criativo e mais determinístico o gerador se torna. Por outro lado, quando mais próximo do limite superior, mais diversas são as respostas, mas também são menos coerentes, visto que o gerador pode selecionar palavras pouco prováveis no contexto semântico. Quando o valor da temperatura é zero, o modelo seleciona a saída mais provável (Zhang; Bao; Huang, 2024).

Contudo, o impacto da temperatura em diferentes problemas é diverso. Para proble-

mas envolvendo questões de múltipla escolha, foi demonstrado que o valor da temperatura ajustada entre 0 e 1 não leva a mudanças estatísticas significantes no desempenho das LLMs (Renze, 2024).

Conforme mencionado anteriormente, os modelos generativos passaram a se basear em instruções recebidas do usuário. Estas instruções são passadas via prompt. De forma geral, o prompt é dividido em mensagem de sistema e mensagem de usuário. A primeira é um conjunto de instruções e/ou restrições que o modelo deve seguir ao longo da conversa. A mensagem do usuário está relacionada com a sua tarefa a ser realizada em um dado momento. É comum que a mensagem de sistema seja inserida antes da mensagem do usuário (Qin *et al.*, 2025).

3 CONJUNTO DE DADOS

O conjunto de dados de ideação suicida foi coletado do Reddit, uma plataforma de mídia social que possui diversas comunidades. Cada comunidade possui um tópico central e os usuários da plataforma podem participar de diversas comunidades, compartilhando e discutindo assuntos relacionados ao tópico (Kumar *et al.*, 2015).

O conjunto de dados foi coletado utilizando duas comunidades: "SuicideWatch"e "Teenagers". Na primeira comunidade ocorrem conversas de pessoas com ideação suicida. Na segunda comunidade os usuários discutem temas relacionados à adolescência, abrangindo tópicos mais comuns entre adolescentes, não necessariamente ideação suicida.

Os dados foram coletados do Reddit usando a API "Pushshift". O título e o corpo das postagens foram unificados para formar a coluna de texto. A coluna "class"representa a classificação das postagens em duas classes: ideação suicida (*suicide*) e sem ideação suicida (*non-suicide*). O conjunto de dados é balanceado e possui 232.074 registros balanceados, onde cada classe contém 116.037 registros. Estes dados foram disponibilizados na plataforma Kaggle (Nikhileswar *et al.*, 2021). Vale destacar que os criadores do dataset declararam não ter realizado o label de forma manual, sendo o label vinculado á comunidade da qual a postagem foi extraída.

3.1 Tamanho das postagens

Ao analisar as postagens do conjunto de dados, percebeu-se que a quantidade média de tokens por postagem é de aproximadamente 131.55, com desvio padrão de 222.08. A menor quantidade de tokens encontrada foi de 1, enquanto a maior foi de 15.632.

O histograma da quantidade de tokens de cada postagem da classe de ideação suicida é mostrado na Figura 3. Postagens com tamanho superior a 500 tokens foram agrupadas na última barra do histograma. A menor quantidade de tokens em uma postagem foi de 1 e a maior de 9.684. A média de tokens para essa classe é de 202.16, com desvio padrão de 254.58. Nesta classe, 9.378 postagens tiveram 500+ tokens.

O histograma para classe sem ideação suicida é mostrado na Figura 4. Novamente, postagens com tamanho superior a 500 tokens foram agrupadas. A menor quantidade de tokens em uma postagem foi de 2 e a maior de 15.632. A média de tokens para essa classe é de 60.93, com desvio padrão de 154.45. Nesta classe, 947 postagens tiveram 500+ tokens.

A Figura 5 mostra a distribução da quantidade de tokens em cada classe através de um boxplot. Para a classe de ideação suicida, a mediana está localizada em 127 tokens, enquanto para a classe sem ideação a mediana é 31. Percebe-se uma considerável diferença entre as medidas centrais das duas classes, apontando na direção de que postagens com

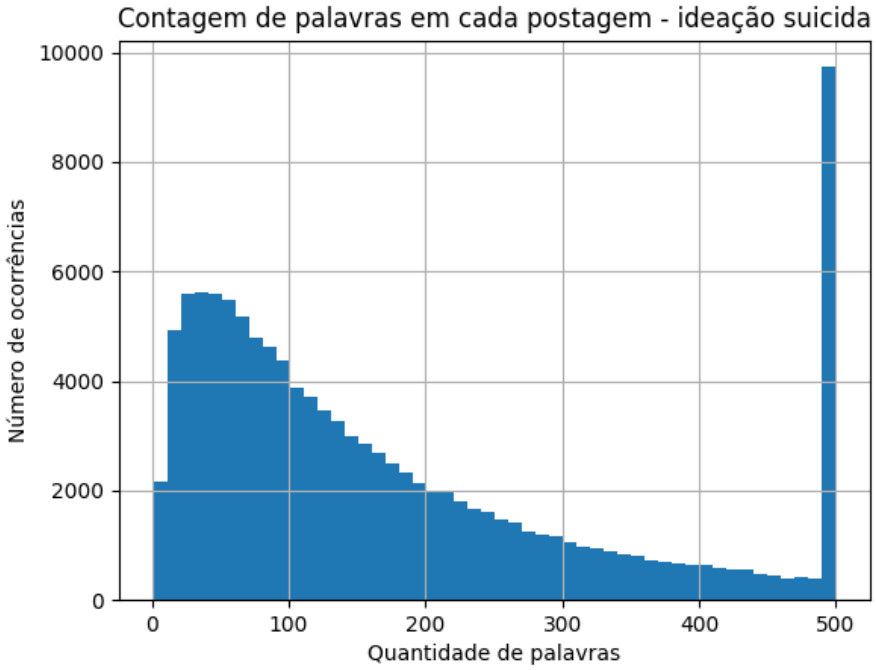

Figura 3 – Histograma da quantidade de tokens nas postagens da classe com ideação suicida. Ocorrências superiores a 500 tokens foram agrupados para melhor visualização.

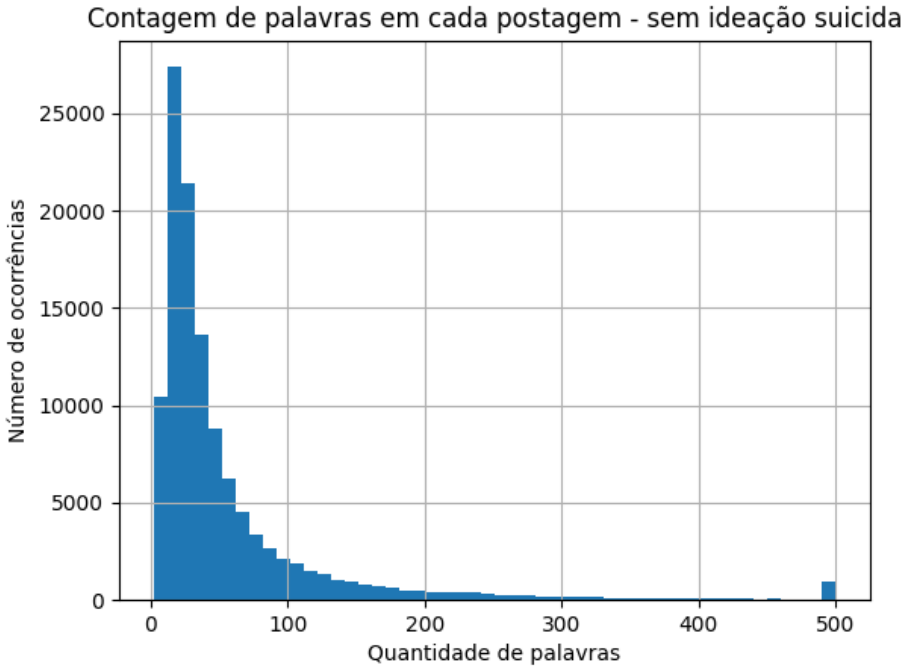

Figura 4 – Histograma da quantidade de tokens nas postagens da classe sem ideação suicida. Ocorrências superiores a 500 tokens foram agrupados para melhor visualização.

ideação suicida tendem a ser maiores.

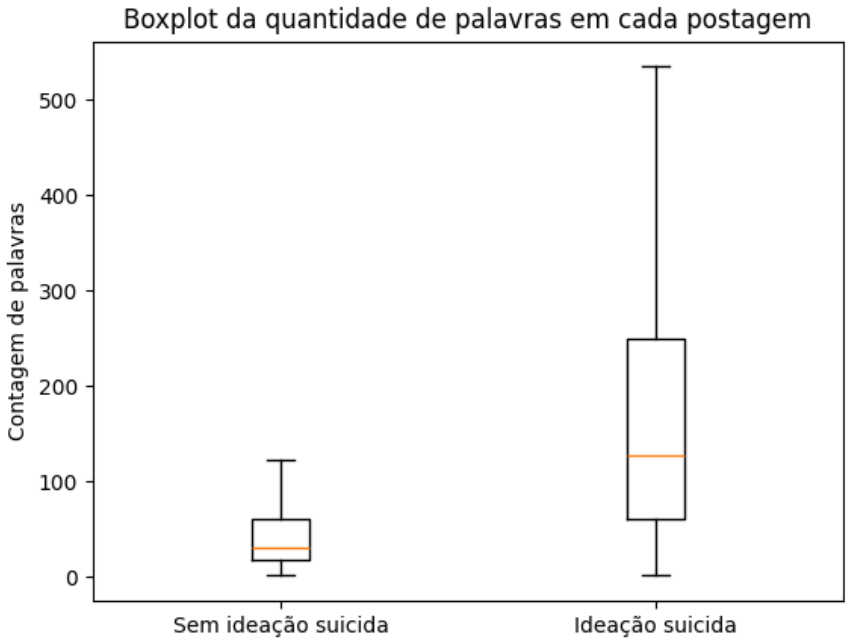

Figura 5 – Boxplot mostrando a distribuição da contagem de tokens das duas classes. Os *outliers* não foram postados para facilitar a visualização.

3.2 Frequência de palavras

Outra característica analisada foram as palavras predominantes de cada classe. Isso pode ser visualizado nos gráficos de barras exibidos nas Figuras 6 e 7. Para esta visualização, foram removidas *stopwords*, números e símbolos, e foi realizado o *lowercasing* das palavras. A frequência absoluta das dez palavras mais comuns na classe suicida e não suicida é mostrada nas Tabelas 3.2 e 3.2, respectivamente.

A comparação das Tabelas mostra diferenças relevantes entre as classes com e sem ideação suicida. Na classe suicida, observam-se termos fortemente associados à expressão de estados emocionais e reflexivos, como "want", "feel", "life"e "know", apontando na direção de uma maior presença de manifestações ligadas a desejos, sentimentos e reflexões existenciais.

Diferentemente, a classe sem ideação suicida apresenta predominância de termos cotidianos e conversacionais, como, por exemplo, "day", "people", "friend"e "really", refletindo uma linguagem em contextos social e cotidiano. Contraintuitivamente, percebe-se a presença da palavra "fuck"na classe não suicida, indicando maior ocorrência de expressões coloquiais sem relação direta com ideação suicida.

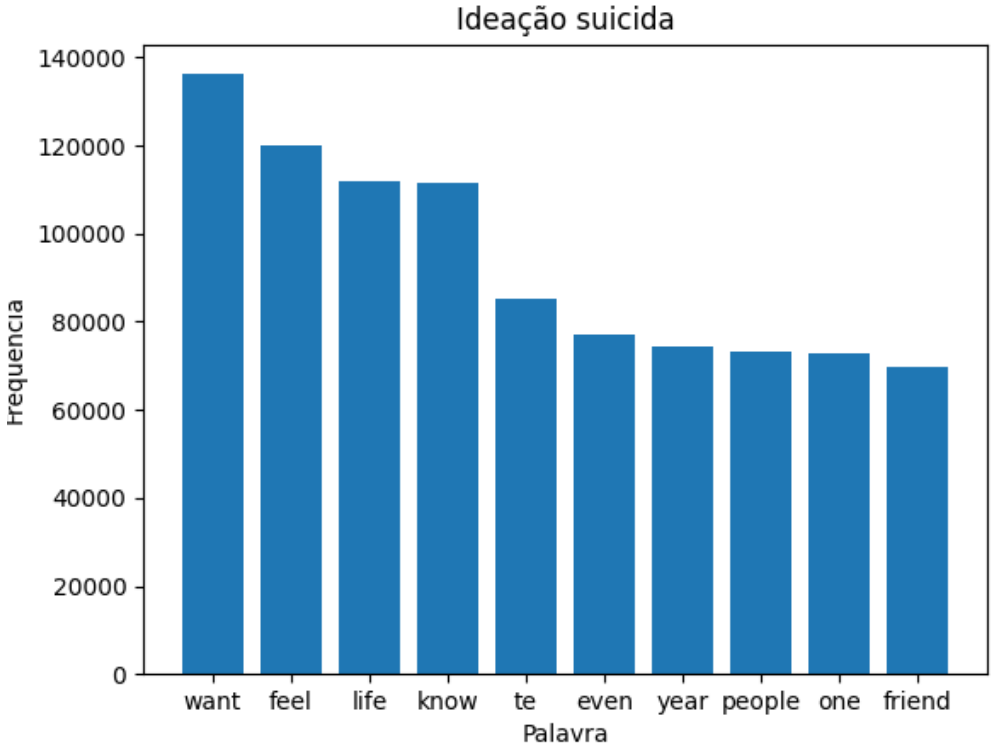

Figura 6 – Gráfico de barras com as palavras mais frequentes na classe com ideação suicida.

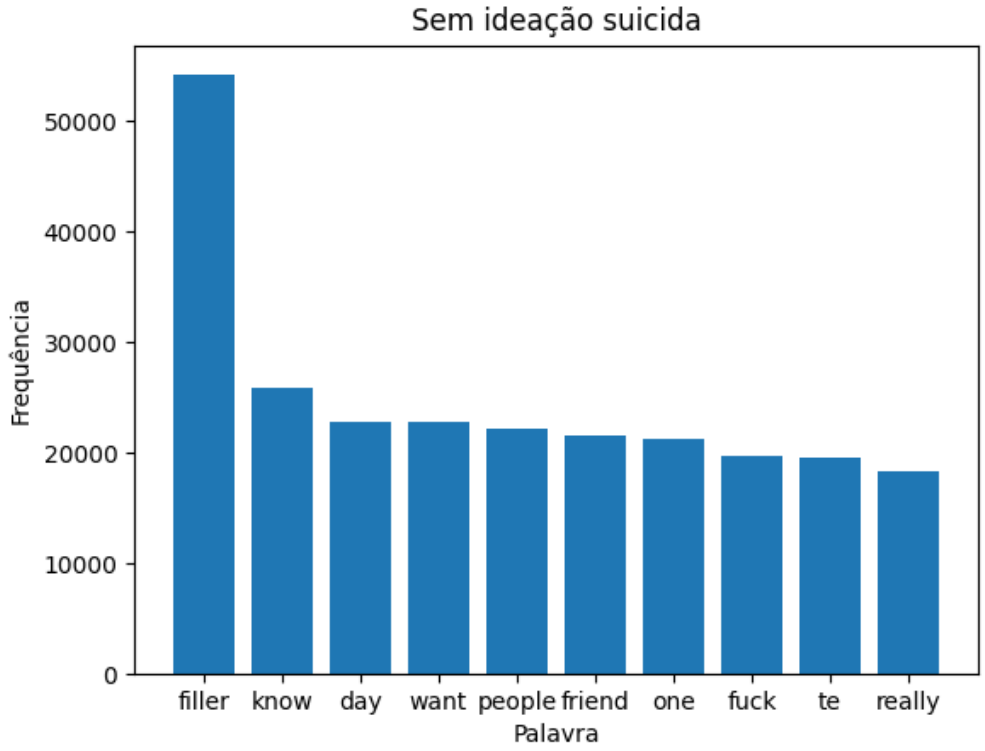

Figura 7 – Gráfico de barras com as palavras mais frequentes na classe sem ideação suicida.

Tabela 1 – Top 10 palavras mais frequentes na classe com ideação suicida

Palavra	**Frequência Absoluta**
want	136.179
feel	119.981
life	111.719
know	111.575
te	85.065
even	76.999
year	74.179
people	73.111
one	72.845
friend	69.898

Tabela 2 – Top 10 palavras mais frequentes na classe sem ideação suicida

Palavra	**Frequência Absoluta**
filler	54.198
know	25.918
day	22.833
want	22.766
people	22.240
friend	21.559
one	21.353
fuck	19.704
te	19.550
really	18.410

4 EXPERIMENTOS E RESULTADOS

Este capítulo descreve as duas abordagens testadas neste trabalho. A primeira abordagem utiliza o SBERT para geração dos *embeddings* e um regressor logístico para classificação. A segunda abordagem utiliza uma LLM para realizar a classificação das postagens.

4.1 Classificação de Ideação Suicida Utilizando SBERT

Conforme apresentado no Capítulo 2, o SBERT gera os *embeddings* do texto. Isto não é suficiente para determinar se a postagem contém ideação suicida ou não. Logo, é necessário acoplar ao SBERT um classificador. Este capítulo escreve os experimentos e resultados obtidos neste sentido.

4.1.1 Experimentos

O pré-processamento dos dados envolveu a substituição das URLs contidas nas postagens por tokens $< URL >$. Além disto, utilizou-se apenas os primeiros 128 tokens de cada postagem, visto que postagens muito longas podem conter muitas senteças com sentidos diferentes, o que pode levar a geração de *embeddings* de pior qualidade para a detecção de ideação suicida. Por outro lado, a remoção de tokens leva a perda de informação que poderia ser útil para a identificação de ideação suicida.

Utilizou-se a biblioteca sentence-transformer (Reimers; Gurevych, 2019) para geração dos *embeddings*. Desta biblioteca utilizou-se o sentence embedding usando o modelo *all-MiniLM-L6-v2*. Os dados foram divididos 80% para treino e 20% para validação (46.415 amostras), mantando-se a proporção das classes em ambos conjuntos. Como classificador, utilizou-se um regressor logístico. A identificação da ideação suicida na maior quantidade de casos possíveis é mais importante que os erros de classificação de ideação não suicida. Com isto, considera-se que a revocação é a métrica mais adequada para avaliar a qualidade dos classificadores para o problema em questão. As demais métricas (acurácia e precisão) possuem importância secundária.

4.1.2 Resultados - SBERT

A Figura 8 mostra a matriz de confusão obtida com o modelo treinado. Das amostas de validação, 21.782 e 21.717 foram classificadas corretamente como não suicida e suicida, respectivamente. Por outro lado, 1.490 e 1.426 foram classificadas incorretamente como não suicida e suicida, respectivamente. Com isto, o classificador mostrou uma acurácia de $93,72\%$, uma precisão de $93,84\%$ e revocação de $93,58\%$.

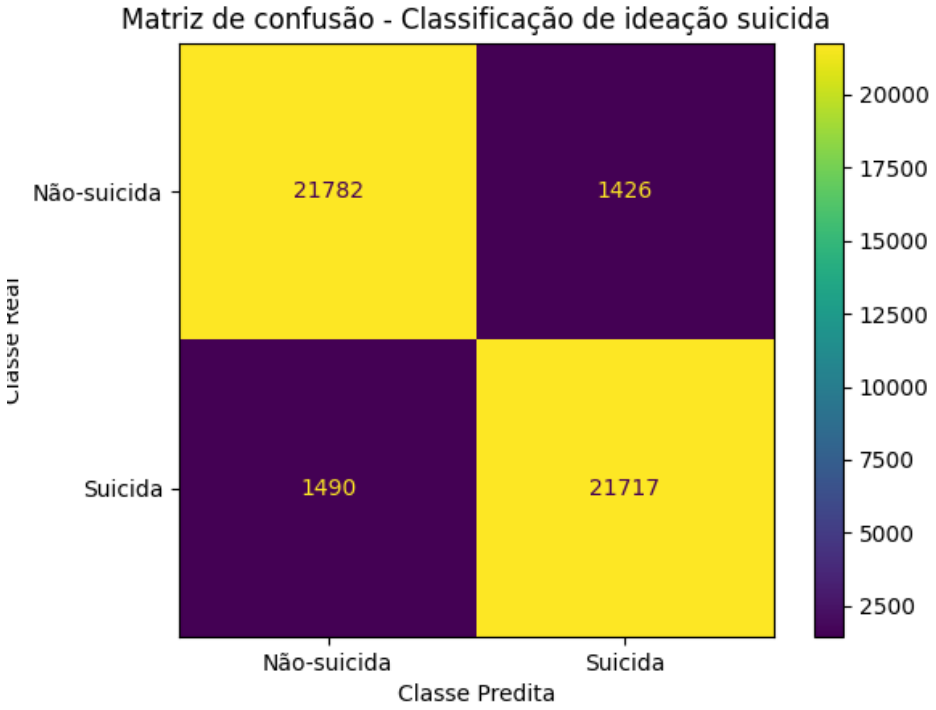

Figura 8 – Matriz de confusão obtida a partir da codificação utilizando SBERT e regressão logística.

Através da análise da matriz de confusão, percebe-se que o modelo demonstra um balanceamento adequado entre as categorias. A quantidade de acertos, tanto de postagens identificados corretamente como suicidas, quanto daqueles classificados corretamente como não suicidas, é consideravelmente alta em comparação com o conjunto de dados. Isso sugere que o classificador compreendeu de forma eficaz os padrões de significado presentes nos *embeddings* SBERT.

A acurácia de $93,72\%$ indica que o modelo apresenta um bom desempenho na identificação correta dos casos. No entanto, esta métrica, por si só, não é suficiente para avaliar modelos envolvendo situações delicadas como a identificação de ideação suicida.

A revocação de $93,58\%$ demonstra que a vasta maioria dos tuítes com conteúdo suicida foram identificados, reduzindo a chance de falsos negativos. Isto é essencial para soluções neste sentido, já que não identificar um caso de suicídio pode ter efeitos piores do que um falso positivo.

O equilíbrio entre precisão e revocação sugere que o classificador não está excessivamente enviesado para uma das classes. Apesar de ainda ocorrerem falhas de identificação (com 1.490 falsos negativos passar e 1.426 falsos negativos, percebe-se o potencial desta abordagem na seleção automática inicial ou como suporte em plataformas de acompanhamento.

4.2 Classificação de Ideação Suicida Utilizando LLM

A LLM utilizada para realizar a tarefa é o Mistral 7B, mencionada nos capítulos anteriores. A metodologia para execução é direta e envolve o pré-processamento dos dados, os parâmetros de configuração do modelo e os resultados.

4.2.1 Configuração da LLM

Primeiramente os dados são carregados e, em seguida, realiza-se um breve tratamento. Neste processo, as URLs de sites foram removidas do texto. Além disto, o texto foi limitado em 128 tokens para fins tanto de eficiência quanto para padronização das entradas.

Para configuração da LLM, considerou-se o valor da temperatura como 0.0, visando uma geração estável de saídas. Foram utilizados dois prompts e a postagem foi inserida no campo «tweet>":

Tabela 3 – Prompts testados para classificação de ideação suicida.

Prompt	**Mensagem de sistema**	**Mensagem de usuário**
Prompt A	You are a psychology expert who answers questions with 'Yes' or 'No' only, without adding any punctuation or any other comment.	Do you consider that the tweet <tweet> shows suicidal ideation of its author?
Prompt B	You are a psychology expert. Classify ONLY with one of these two exact words: Yes or No. "Yes"= the tweet's author seems to express desire, intention, plan, or preparation for self-harm or suicide. "No"= any other case, including: metaphor/hyperbole ('I'm dying of laughter'), quotes, song lyrics, news reports, general discussions, jokes/sarcasm, support/encouragement to others, prevention, generic requests for help without intention, references to third parties (not the author), or ambiguous phrases without sufficient context. Do not explain. Do not use punctuation. Respond only with Yes or No.	Does the tweet below show potential suicidal ideation by the AUTHOR? [TWEET]: <tweet>

A primeira instrução definiu um comando de sistema sucinto, limitando o modelo a responder apenas com "Yes"ou "No", sem incluir qualquer pontuação extra ou observações. Esta abordagem tem o benefício de ser direta e fácil de usar em problemas de classificação

binária, diminuindo a variação nas respostas e promovendo a repetição dos resultados. Porém, como não detalha as condições que mostram pensamentos suicidas, o modelo pode ter mais dúvidas em situações incertas, como expressões figuradas ou referências indiretas a outras pessoas, o que pode afetar o desempenho.

A segunda instrução foi criada de forma mais detalhada e descritiva, especificando claramente os critérios para uma classificação tanto positiva quanto negativa, com exemplos de situações que não devem ser vistas como pensamentos suicidas, como, por exemplo, exageros, notícias ou piadas. Esse formato possui a vantagem de dar ao modelo clareza ao definir os limites da classificação, o que tende a diminuir os erros e melhorar a consistência na interpretação.

Para processamento das respostas da LLM, serão considerados: "Yes"e "Yes."como classificação positiva para ideação suicida; respostas divergentes destas foram consideradas como negativas para ideação suicida.

4.2.2 Resultados - LLM

A classificação via LLM foi realizada utilizando o mesmo conjunto de dados de validação utilizado para treino do classificador logístico. Os resultados obtidos são mostrados na Tabela 4.2.2 e as matrizes de confusão são mostradas nas Figuras 9 e 10.

Tabela 4 – Resultados de classificação por prompt

Prompt	**Acurácia**	**Precisão**	**Revocação**
Prompt A	0.8980	0.8433	0.9776
Prompt B	0.8767	0.8284	0.9504

Para o Prompt A, 18.992 e 22.688 postagens foram classificadas corretamente nas classes "não suicida"e "suicida", respectivamente, enquanto 519 e 4.216 foram classificadas erroneamente nas classes "não suicida"e "suicida", respectivamente. A acurácia do modelo foi de $89,80\%$, enquanto a precisão foi de $84,33\%$. A revocação mostra que $97,76\%$ dos tweets suicidas foram corretamente identificados pela LLM.

O desempenho da LLM com o Prompt A demonstra que instruções concisas foram mais eficazes para a tarefa de classificação. A revocação de $97,76\%$ evidencia que o modelo foi capaz de identificar quase a totalidade das postagens com ideação suicida, minimizando significativamente a ocorrência de falsos negativos. Esse resultado é extremamente relevante para aplicações voltadas para a prevenção, onde a sensibilidade do modelo é prioritária.

Por outro lado, a precisão de $84,33\%$ sugere maior ocorrência de falsos positivos em comparação à abordagem com SBERT. Ainda assim, este comportamento pode ser considerado aceitável em cenários críticos, conforme mencionado anteriormente.

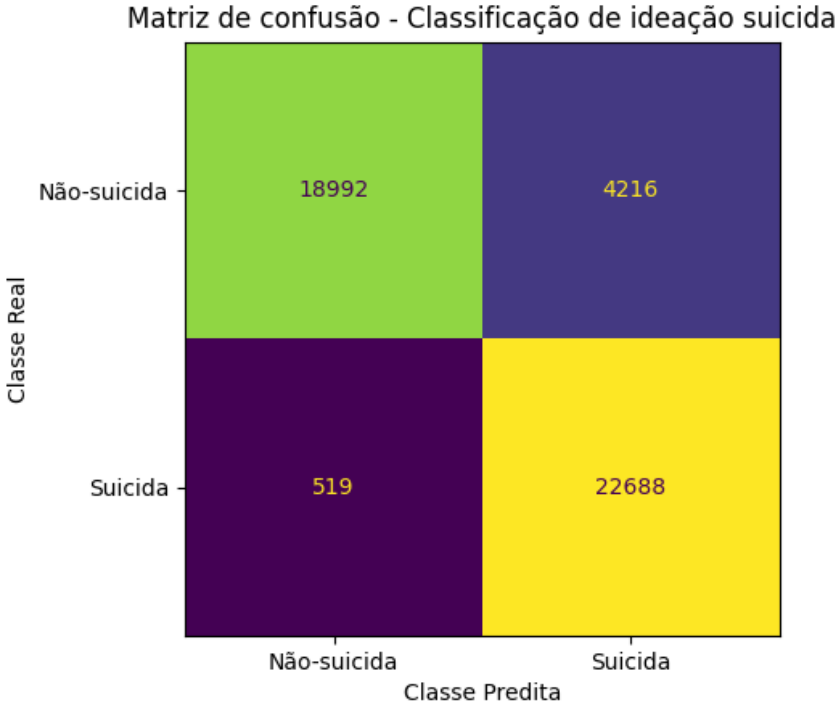

Figura 9 – Matriz de confusão para o prompt A

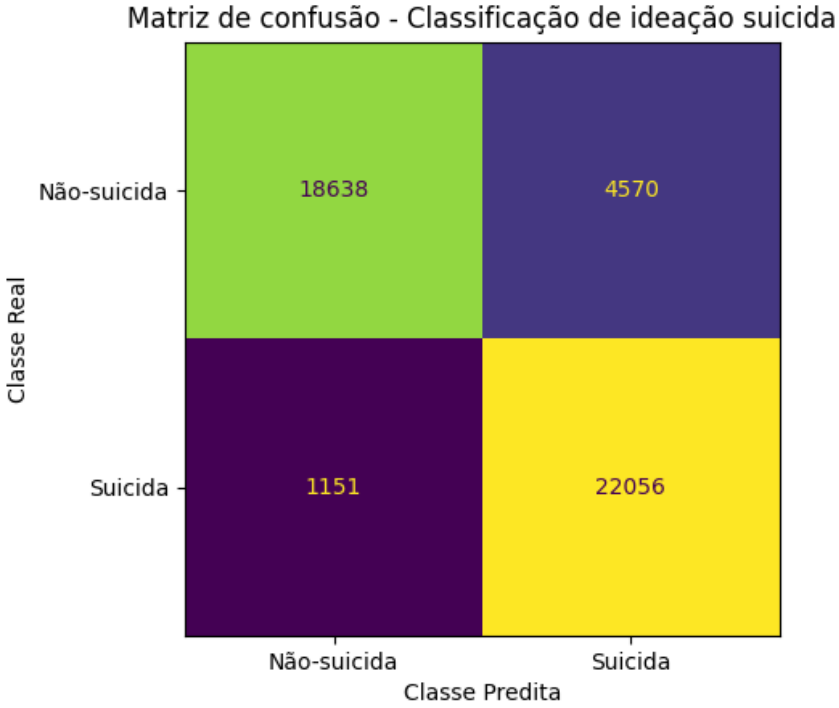

Figura 10 – Matriz de confusão para o prompt B

Para o Prompt B, 18.638 foram classificadas corretamente como "não suicidas"e 22.056 corretamente como "suicida". Por outro lado, o modelo classificou erroneamente 1.151 e 4.570 nas classes "não suicida"e "suicida", respectivamente. Isso leva a uma acurácia de $87,67\%$; uma precisão de $82,84\%$ e uma revocação de $95,04\%$.

Apesar de, em tese, comandos mais detalhados e exemplificados terem o potencial de guiar o modelo com mais precisão, no teste atual, o Mistral exibiu resultados melhores com o Prompt A, que possui solicitações concisas e objetivas. Uma das explicações para isto está no fato de o aprendizado por *few-shot* ser bastante instável frente à pequenas mudanças no prompt (Zhao *et al.*, 2021). Outra possibilidade é que o modelo esteja mais habituado a comandos simples (zero-shot) do que a situações com exemplos repletos de detalhes.

4.3 Exemplos

A Tabela 5 mosta alguns exemplos de postagens classificadas tanto pela LLM quando pelo SBERT. Nos exemplos 1-4, é possível visualizar várias postagens corriqueiras sem o mínimo sinal de ideação suicida. Tanto a LLM em ambos prompts quanto o método baseado no SBERT classificaram estas postagens de forma correta.

Os exemplos 5-7 mostram casos em que as abordagens classificaram erroneamente as postagens como suicidas. No caso do exemplo 8, o prompt A levou a um acerto, enaquanto o B a erro. Os exemplos 5-7 mostram postagens de pessoas com uma forte sensação negativa de si. Estes sentimentos não estão necessariamente envolvidos com ideação suicida, mas ser indicativos de alerta no futuro. No caso do exemplo 8, apenas a LLM com prompt B errou.

Os exemplos 9, 10 e 12 mostram casos classificados erroneamente como não suicidas. Estes exemplos se destacam visto que por si só não dão indicativo claro de ideação suicida. Isso pode ser explicado pela forma como os dados foram coletados. No contexto de uma comunidade voltada para assuntos sobre o suicídio, essas mensagens podem indicar potenciais ideações suicidas. Fora deste contexto, o entendimento pode ser outro. No caso do exemplo 11, o a LLM (prompt A) e o SBERT classificaram corretamente a postagem como suicida.

Por fim, os exemplos 13-16 refletem casos onde todas as abordagens classificaram as postagens corretamente. Em todos os casos, fica muito nítida essa ideação, em parte pela utilização de termos como "kill"e "suicidal".

4.4 Comparativo com modelos de referência

Esta seção busca comparar o desempenho dos modelos apresentados neste projeto com modelos propostos na literatura. Serão consideradas duas referências: (1) solução

Tabela 5 – Exemplos postagens retiradas dos dados de validação e suas respectivas classificações (S - suicida, NS - não suicida) a partir do prompt A (PA), prompt B (PB) e SBERT + classificador (SBERT).

Número	Postagem	PA	PB	SBERT	Classe
1	"Y'all, I'm feeling Groovy"	NS	NS	NS	NS
2	"reddit updated againnnn its weirddddddd"	NS	NS	NS	NS
3	"hows your day? anything exciting happen?"	NS	NS	NS	NS
4	"one hour starts now btw oop"	NS	NS	NS	NS
5	"I'm a piece of shit. i don't feel so good :("	S	S	S	NS
6	"I hate myself I feel bad"	S	S	S	NS
7	"I hate my life Currently."	S	S	S	NS
8	"I CAN'T SLEEP It's 3:50 please HELP."	NS	S	NS	NS
9	"HeyMessage me to call me"	NS	NS	NS	S
10	"Welp.Seriously, fuck it."	NS	NS	NS	S
11	"BridgeFall to my death."	S	NS	S	S
12	"HeeeyyyyCan I please talk to someone ;-;"	NS	NS	NS	S
13	"Im about to do itI see no other way"	S	S	S	S
14	"Might kill myself todayYes."	S	S	S	S
15	"Sometimes i feel like ending it allI need help"	S	S	S	S
16	"suicidal thoughts for 7 yearsforget it"	S	S	S	S

proposta por (Nikhileswar *et al.*, 2021) e (2) solução proposta por (Samuel *et al.*, 2024). Na proposta (1), duas técnicas apresentaram resultados significativos: *Support Vector Machine* (SVM) e *Fully Connected Neural Network* (FCNN). Na segunda, combina-se BERT + BiLSTM + CNN.

Para fins de melhor comparação, os resultados são reapresentados na Tabela 4.4. Percebe-se que o SBERT apresentou resultados levemente inferiores aos três modelos de referência. No caso da LLM, esta apresentou acurácia e precisão consideravelmente inferiores aos três modelos de referência. Contudo, a revocação foi superior àquela obtida pela SVM e FCNN, mas levemente inferior ao BERT.

Tabela 6 – Comparativo entre as técnicas propostas e modelos de referência.

Modelo	**Acurácia**	**Precisão**	**Revocação**
SBERT	0,9372	0,9384	0,9358
LLM	0,8980	0,8433	0,9776
SVM	0,938	0,94	0,94
FCNN	0,9416	0,945	0,945
BERT	0,9699	0,960	0,980

5 CONSIDERAÇÕES FINAIS

Neste trabalho foram apresentadas duas abordagens (LLM e SBERT + classificador) para realizar a classificação de ideação suicida em postagens. Os experimentos realizados demonstraram resultados satisfatórios e consistentes para a tarefa de detecção de ideação suicida.

O modelo baseado em SBERT apresentou desempenho bastante equilibrado entre precisão e revocação, alcançando valores superiores a 93% em acurácia, precisão e revocação. Isto indica um comportamento equilibrado e robusto, capaz de identificar corretamente a maioria dos casos de ideação sem comprometer excessivamente a taxa de falsos positivos, tornando esta técnicas adequada para aplicações em cenários que demandem maior nível de confiabilidade e estabilidade.

Diferentemente, a classificação utilizando a LLM Mistral 7B apresentou acurácia e precisão levemente inferiores. Entretanto, a principal vantagem observada foi a revocação extremamente elevada, em especial para o Prompt A, onde o modelo identificou quase todos os casos positivos de ideação suicida. Esta abordagem é relevante em aplicações voltadas para a prevenção, já que minimizar a ocorrência de falsos negativos leva à redução de riscos de fatalidades. Apesar disto, esta abordagem apresentou um número maior de falsos positivos, indicando a necessidade de estratégias complementares de filtragem para minimizar a geração de falsos positivos.

Um ponto interessante observado foi o resultado contraintuitivo obtido a partir de cada prompt. O Mistral 7B demonstrou um desempenho superior com orientações claras e objetivas, ao invés de solicitações complexas, o que evidencia a relevância da elaboração do comando no desempenho de tarefas envolvendo LLMs.

Por fim, enquanto o SBERT se mostra uma solução sólida e balanceada, as LLMs apresentam potencial para cenários em que a prioridade é a máxima cobertura dos casos de risco, ainda que consideravelemente sujeitos a falsos alarmes. Com isto, o trabalho mostra não apenas a viabilidade técnica de ambas as abordagens, mas também ressalta que a escolha entre elas deve considerar o contexto prático de aplicação, equilibrando precisão, revocação e custo de eventuais erros.

Para trabalhos futuros, pretende-se utilizar LLMs maiores para a tarefa de classificação, bem como outras técnicas de *embeddings*, classificadores e redes complexas.

REFERÊNCIAS

ABUBAKAR, H. D.; UMAR, M. Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. **SLU Journal of Science and Technology**, Sule Lamido University Kafin Hausa, v. 4, n. 1 and 2, p. 27–33, ago. 2022. Disponível em: http://dx.doi.org/10.56471/slujst.v4i.266.

ALDHYANI, T. H. H. *et al.* Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. **International Journal of Environmental Research and Public Health**, MDPI AG, v. 19, n. 19, p. 12635, out. 2022. ISSN 1660-4601. Disponível em: http://dx.doi.org/10.3390/ijerph191912635.

ALLAM, H. *et al.* Ai-driven mental health surveillance: Identifying suicidal ideation through machine learning techniques. **Big Data and Cognitive Computing**, MDPI AG, v. 9, n. 1, p. 16, jan. 2025. ISSN 2504-2289. Disponível em: http://dx.doi.org/10.3390/bdcc9010016.

ALLARAKHA, S.; UTTEKAR, P. **What Is a Suicidal Ideation Scale?** 2021. Accessed on March 9, 2025. Disponível em: https://www.medicinenet.com/what_is_a_suicidal_ideation_scale/article.htm.

BALCı, E.; SARAç, E. Automated depression detection from tweets: a comparison of nlp techniques. *In*: **2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)**. IEEE, 2024. p. 1–5. Disponível em: http://dx.doi.org/10.1109/IDAP64064.2024.10711029.

BAYRAM, U.; BENHIBA, L. Determining a person's suicide risk by voting on the short-term history of tweets for the clpsych 2021 shared task. *In*: **Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access**. Association for Computational Linguistics, 2021. p. 81–86. Disponível em: http://dx.doi.org/10.18653/v1/2021.clpsych-1.8.

BENJACHAIRAT, P. *et al.* Classification of suicidal ideation severity from twitter messages using machine learning. **International Journal of Information Management Data Insights**, Elsevier BV, v. 4, n. 2, p. 100280, nov. 2024. ISSN 2667-0968. Disponível em: http://dx.doi.org/10.1016/j.jjimei.2024.100280.

BEURS, D. D. Network analysis: A novel approach to understand suicidal behaviour. **International Journal of Environmental Research and Public Health**, MDPI AG, v. 14, n. 3, p. 219, fev. 2017. ISSN 1660-4601. Disponível em: http://dx.doi.org/10.3390/ijerph14030219.

BREIMAN, L. *et al.* **Classification And Regression Trees**. Routledge, 2017. ISBN 9781315139470. Disponível em: http://dx.doi.org/10.1201/9781315139470.

BROWN, T. B. *et al.* Language models are few-shot learners. *In*: **Advances in Neural Information Processing Systems (NeurIPS)**. [*S.l.: s.n.*], 2020. Disponível em: https://arxiv.org/abs/2005.14165.

CHATTERJEE, M. *et al.* Suicide ideation detection from online social media: A multi-modal feature based technique. **International Journal of Information Management Data Insights**, Elsevier BV, v. 2, n. 2, p. 100103, nov. 2022. ISSN 2667-0968. Disponível em: http://dx.doi.org/10.1016/j.jjimei.2022.100103.

CHO, K. *et al.* Learning phrase representations using rnn encoder–decoder for statistical machine translation. *In*: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [*S.l.: s.n.*], 2014. p. 1724–1734.

CHOUDHURY, M. D. *et al.* Discovering shifts to suicidal ideation from mental health content in social media. *In*: **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems**. ACM, 2016. (CHI'16), p. 2098–2110. Disponível em: http://dx.doi.org/10.1145/2858036.2858207.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. Cambridge University Press, 2000. ISBN 9780511801389. Disponível em: http://dx.doi.org/10.1017/CBO9780511801389.

DataCamp. **How Transformers Work: A Detailed Exploration of Transformer Architecture**. 2024. https://www.datacamp.com/tutorial/how-transformers-work. Acessado em 6 de setembro de 2025.

DEJONG, T. M.; OVERHOLSER, J. C.; STOCKMEIER, C. A. Apples to oranges?: A direct comparison between suicide attempters and suicide completers. **Journal of Affective Disorders**, Elsevier BV, v. 124, n. 1–2, p. 90–97, jul. 2010. ISSN 0165-0327. Disponível em: http://dx.doi.org/10.1016/j.jad.2009.10.020.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. arXiv, 2018. Disponível em: https://arxiv.org/abs/1810.04805.

DUONG, H.-T.; NGUYEN-THI, T.-A. A review: preprocessing techniques and data augmentation for sentiment analysis. **Computational Social Networks**, Springer Science and Business Media LLC, v. 8, n. 1, jan. 2021. ISSN 2197-4314. Disponível em: http://dx.doi.org/10.1186/s40649-020-00080-x.

EKMAN, M. **Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow**. Pearson Education, Limited, 2021. ISBN 9780137470259. Disponível em: https://books.google.com.br/books?id=4qaUzgEACAAJ.

ELYOSEPH, Z.; SHOVAL, D. H.; LEVKOVICH, I. Beyond personhood: Ethical paradigms in the generative artificial intelligence era. **The American Journal of Bioethics**, Informa UK Limited, v. 24, n. 1, p. 57–59, jan. 2024. ISSN 1536-0075. Disponível em: http://dx.doi.org/10.1080/15265161.2023.2278546.

FRANKLIN, J. C. *et al.* Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. **Psychological Bulletin**, American Psychological Association (APA), v. 143, n. 2, p. 187–232, 2017. ISSN 0033-2909. Disponível em: http://dx.doi.org/10.1037/bul0000084.

ISLAM, M. R. *et al.* Machine learning with explainability for suicide ideation detection from social media data. *In*: **2023 10th International Conference on Behavioural and Social Computing (BESC)**. IEEE, 2023. p. 1–6. Disponível em: http://dx.doi.org/10.1109/BESC59560.2023.10386773.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. Springer US, 2021. ISSN 2197-4136. ISBN 9781071614181. Disponível em: http://dx.doi.org/10.1007/978-1-0716-1418-1.

JI, S. *et al.* Suicidal ideation detection: A review of machine learning methods and applications. **IEEE Transactions on Computational Social Systems**, Institute of Electrical and Electronics Engineers (IEEE), v. 8, n. 1, p. 214–226, fev. 2021. ISSN 2373-7476. Disponível em: http://dx.doi.org/10.1109/TCSS.2020.3021467.

JI, S. *et al.* Supervised learning for suicidal ideation detection in online user content. **Complexity**, Wiley, v. 2018, n. 1, jan. 2018. ISSN 1099-0526. Disponível em: http://dx.doi.org/10.1155/2018/6157249.

JIANG, A. Q. *et al.* **Mistral 7B**. 2023. Disponível em: https://arxiv.org/abs/2310.06825.

KOMATI, N. **Suicide Watch Dataset**. 2020. https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. Acessado em: 4 de setembro de 2025.

KRUTHIKA, B. *et al.* Exploring multi-vectorization approaches for suicide ideation detection in tweets. *In*: **2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. IEEE, 2024. p. 1–7. Disponível em: http://dx.doi.org/10.1109/ICCCNT61001.2024.10724204.

KUMAR, D. K. **Limitations of Word2vec**. 2025. https://medium.com/@dhananjaikrishnakumar/limitations-of-word2vec-33db20ac63cc. [Online; accessado em 01-Junho-2025].

KUMAR, M. *et al.* Detecting changes in suicide content manifested in social media following celebrity suicides. *In*: **Proceedings of the 26th ACM Conference on Hypertext; Social Media - HT '15**. ACM Press, 2015. (HT '15), p. 85–94. Disponível em: http://dx.doi.org/10.1145/2700171.2791026.

MCHUGH, C. M. *et al.* Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. **BJPsych Open**, Royal College of Psychiatrists, v. 5, n. 2, jan. 2019. ISSN 2056-4724. Disponível em: http://dx.doi.org/10.1192/bjo.2018.88.

MIKOLOV, T. *et al.* **Distributed Representations of Words and Phrases and their Compositionality**. arXiv, 2013. Disponível em: https://arxiv.org/abs/1310.4546.

NASEEM, U. *et al.* Hybrid text representation for explainable suicide risk identification on social media. **IEEE Transactions on Computational Social Systems**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, n. 4, p. 4663–4672, ago. 2024. ISSN 2373-7476. Disponível em: http://dx.doi.org/10.1109/TCSS.2022.3184984.

NIKHILESWAR, K. *et al.* Suicide ideation detection in social media forums. *In*: **2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)**. [*S.l.: s.n.*], 2021. p. 1741–1747.

OMAR, M. *et al.* Applications of large language models in psychiatry: a systematic review. **Frontiers in Psychiatry**, Frontiers Media SA, v. 15, jun. 2024. ISSN 1664-0640. Disponível em: http://dx.doi.org/10.3389/fpsyt.2024.1422807.

ORGANIZATION, W. H. **Suicide**. 2024. Accessed on March 9, 2025. Disponível em: https://www.who.int/news-room/fact-sheets/detail/suicide.

OUYANG, L. *et al.* **Training language models to follow instructions with human feedback**. 2022. Disponível em: https://arxiv.org/abs/2203.02155.

O'DEA, B. *et al.* Detecting suicidality on twitter. **Internet Interventions**, Elsevier BV, v. 2, n. 2, p. 183–188, maio 2015. ISSN 2214-7829. Disponível em: http://dx.doi.org/10.1016/j.invent.2015.03.005.

POLANCO-ROMAN, L. *et al.* Association of childhood adversities with suicide ideation and attempts in puerto rican young adults. **JAMA Psychiatry**, American Medical Association (AMA), v. 78, n. 8, p. 896, ago. 2021. ISSN 2168-622X. Disponível em: http://dx.doi.org/10.1001/jamapsychiatry.2021.0480.

QIN, Y. *et al.* Sysbench: Can large language models follow system messages? *In*: **Proceedings of the International Conference on Learning Representations (ICLR)**. [*S.l.: s.n.*], 2025. Disponível em: https://arxiv.org/abs/2408.10943.

RADFORD, A. *et al.* **Improving Language Understanding by Generative Pre-Training**. [*S.l.*], 2018. Disponível em: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

RADFORD, A. *et al.* **Language Models are Unsupervised Multitask Learners**. [*S.l.*], 2019. Disponível em: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

RASCHKA, S. **Naive Bayes and Text Classification I - Introduction and Theory**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1410.5329.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *In*: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: https://arxiv.org/abs/1908.10084.

RENZE, M. The effect of sampling temperature on problem solving in large language models. *In*: **Findings of the Association for Computational Linguistics: EMNLP 2024**. Association for Computational Linguistics, 2024. p. 7346–7356. Disponível em: http://dx.doi.org/10.18653/v1/2024.findings-emnlp.432.

ROBINSON, J. *et al.* Social media and suicide prevention: a systematic review. **Early Intervention in Psychiatry**, Wiley, v. 10, n. 2, p. 103–121, fev. 2015. ISSN 1751-7893. Disponível em: http://dx.doi.org/10.1111/eip.12229.

SAIF, H. *et al.* On stopwords, filtering and data sparsity for sentiment analysis of Twitter. *In*: CALZOLARI, N. *et al.* (ed.). **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 810–817. Disponível em: https://aclanthology.org/L14-1265/.

SAMUEL, R. J. R. *et al.* Nlp-ml hybrid: Identifying signs of suicidal thoughts in social media content. *In*: **2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)**. [*S.l.: s.n.*], 2024. p. 1373–1379.

SCHUSTER, M.; NAKAJIMA, K. Japanese and korean voice search. *In*: **2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2012. Disponível em: http://dx.doi.org/10.1109/ICASSP.2012.6289079.

SETIAWAN, M. D. *et al.* Transfer learning using indobert with long short-term memory classifier for detection of suicide ideation themed indonesian twitter posts. *In*: **2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)**. IEEE, 2024. p. 276–281. Disponível em: http://dx.doi.org/10.1109/ICSINTESA62455.2024.10747816.

SINGH, M. *et al.* Suicide detection using twitter sentiment analysis. *In*: **2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)**. IEEE, 2024. p. 808–813. Disponível em: http://dx.doi.org/10.23919/INDIACom61295.2024.10499165.

SOFIA *et al.* Machine learning based model for detecting depression during covid-19 crisis. **Scientific African**, Elsevier BV, v. 20, p. e01716, jul. 2023. ISSN 2468-2276. Disponível em: http://dx.doi.org/10.1016/j.sciaf.2023.e01716.

SRINIVASARAO, T. *et al.* Suicidal tendency detection from twitter posts using a long short memory vectors. *In*: **2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)**. IEEE, 2024. p. 422–427. Disponível em: http://dx.doi.org/10.1109/ESIC60604.2024.10481661.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. *In*: **Advances in Neural Information Processing Systems (NeurIPS)**. [*S.l.: s.n.*], 2014. Disponível em: https://arxiv.org/abs/1409.3215.

TADESSE, M. M. *et al.* Detection of suicide ideation in social media forums using deep learning. **Algorithms**, MDPI AG, v. 13, n. 1, p. 7, dez. 2019. ISSN 1999-4893. Disponível em: http://dx.doi.org/10.3390/a13010007.

TAN, K. L.; LEE, C. P.; LIM, K. M. A survey of sentiment analysis: Approaches, datasets, and future research. **Applied Sciences**, MDPI AG, v. 13, n. 7, p. 4550, abr. 2023. ISSN 2076-3417. Disponível em: http://dx.doi.org/10.3390/app13074550.

TAYIM, N. *et al.* A cross-sectional network analysis of intimate partner violence and suicidal ideation among arab women. **Clinical Psychology; Psychotherapy**, Wiley, v. 32, n. 1, jan. 2025. ISSN 1099-0879. Disponível em: http://dx.doi.org/10.1002/cpp.70037.

TEIXEIRA, A. S. *et al.* Revealing semantic and emotional structure of suicide notes with cognitive network science. **Scientific Reports**, Springer Science and Business Media LLC, v. 11, n. 1, set. 2021. ISSN 2045-2322. Disponível em: http://dx.doi.org/10.1038/s41598-021-98147-w.

VANDANA; MARRIWALA, N.; CHAUDHARY, D. A hybrid model for depression detection using deep learning. **Measurement: Sensors**, Elsevier BV, v. 25, p. 100587, fev. 2023. ISSN 2665-9174. Disponível em: http://dx.doi.org/10.1016/j.measen.2022.100587.

VASWANI, A. *et al.* Attention is all you need. *In*: **Advances in Neural Information Processing Systems (NeurIPS)**. [*S.l.: s.n.*], 2017. p. 5998–6008. Disponível em: https://arxiv.org/abs/1706.03762.

WALLACE, E. *et al.* **Do NLP Models Know Numbers? Probing Numeracy in Embeddings**. 2019. Disponível em: https://arxiv.org/abs/1909.07940.

WANG, N. *et al.* Learning models for suicide prediction from social media posts. *In*: **Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access**. Association for Computational Linguistics, 2021. Disponível em: http://dx.doi.org/10.18653/v1/2021.clpsych-1.9.

ZHANG, S.; BAO, Y.; HUANG, S. **EDT: Improving Large Language Models' Generation by Entropy-based Dynamic Temperature Sampling**. 2024. Disponível em: https://arxiv.org/abs/2403.14541.

ZHAO, T. Z. *et al.* **Calibrate Before Use: Improving Few-Shot Performance of Language Models**. 2021. Disponível em: https://arxiv.org/abs/2102.09690.

ZHAO, W. X. *et al.* A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023. Disponível em: https://arxiv.org/abs/2303.18223.

ZHONG, S. *et al.* A network analysis of depressive symptoms, psychosocial factors, and suicidal ideation in 8686 adolescents aged 12–20 years. **Psychiatry Research**, v. 329, p. 115517, 2023. ISSN 0165-1781. Disponível em: https://www.sciencedirect.com/science/article/pii/S0165178123004675.