Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.7642

Artigo de Pesquisa/Research Paper

Large Language Models na Identificação de Ideação Suicida em Textos não Clínicos em Inglês

Large Language Models in the Identification of Suicidal Ideation in Non-Clinical English Texts

Lais Carvalho Coutinho [Instituto Federal de Educação, Ciência e Tecnologia do Ceará | *lais.carvalho07@aluno.ifce.edu.br*]
Antonia Estefane Ribeiro Veras [Instituto Federal de Educação, Ciência e Tecnologia do Ceará |*estefaneribeiroveras@gmail.com*]
Rosana Celine Pinheiro Damasceno [Instituto Federal de Educação, Ciência e Tecnologia do Ceará |*celine.rosana08@aluno.ifce.edu.br*]
Adonias Caetano de Oliveira [Instituto Federal de Educação, Ciência e Tecnologia do Ceará |*adonias.oliveira@ifce.edu.br*]

Instituto Federal de Educação, Ciência e Tecnologia do Ceará, CE-187, s/n, Estádio, Tianguá, CE, 62320-000, Brazil.

Resumo. O suicídio é um fenômeno complexo que pode iniciar com ideação suicida e evoluir para planos e tentativas. Com a popularização das redes sociais, usuários frequentemente expressam sentimentos negativos nesses ambientes, possibilitando a identificação de sinais de risco. Este estudo tem como objetivo avaliar o desempenho de modelos baseados em *Transformers* e *Large Language Models* (LLMs) na detecção de ideação suicida em textos. A metodologia consistiu na aplicação e comparação de modelos discriminativos (ALBERT, Mental-RoBERTa e BERT-Large) e generativos (DeepSeek e ChatGPT-4o) em dois conjuntos de dados: *"Suicide and Depression Detection"* e *"Suicide vs Depression Classification"*, utilizando métricas como acurácia, precisão, sensibilidade e Medida-F1. Os resultados indicaram alto desempenho dos modelos discriminativos no primeiro conjunto (até 0,99), enquanto os LLMs apresentaram melhor equilíbrio na detecção da classe *suicide*, com F1 de até 0,74. Esses achados evidenciam o potencial dessas abordagens para apoiar sistemas de saúde mental.

Abstract. Suicide is a complex phenomenon that may begin with suicidal ideation and progress to planning and attempts. With the widespread use of social media, users often express negative feelings in these environments, enabling the identification of risk signals. This study aims to evaluate the performance of models based on *Transformers* and *Large Language Models* (LLMs) in detecting suicidal ideation in textual data. The methodology consisted of applying and comparing discriminative models (ALBERT, Mental-RoBERTa and BERT-Large) and generative models (DeepSeek and ChatGPT-4o) across two datasets: *"Suicide and Depression Detection"* and *"Suicide vs Depression Classification"*, using metrics such as accuracy, precision, recall and F1-score. The results indicated high performance of discriminative models in the first dataset (up to 0.99), while LLMs showed a better balance in detecting the *suicide* class, achieving F1-scores of up to 0.74. These findings highlight the potential of these approaches to support mental health systems.

Palavras-chave:
Aprendizagem de Máquina, Inteligência Artificial, *Large Language Models*, Ideação Suicida, Processamento de Linguagem Natural.

Keywords: Machine Learning, Artificial Intelligence, Large Language Models, Suicidal Ideation, Natural Language Processing.

Recebido/Received: 19 February 2026 • **Aceito/Accepted:** 03 May 2026 • **Publicado/Published:** 03 June 2026

1 Introdução

A ideação suicida, considerada uma manifestação inicial de atos suicidas, envolve pensamentos vagos sobre a morte e o desejo de morrer [Neeleman *et al*., 2004]. Esse estágio pode evoluir para um plano suicida concreto e culminar em uma tentativa de suicídio [Bertolote, 2016]. A ideação pode ser passiva, caracterizada por um desejo abstrato de morrer, ou ativa, envolvendo um plano específico para sua concretização [Rizvi *et al*., 2025].

Historicamente, indivíduos expressavam tais sentimentos por meio de notas de suicídio, cujo conteúdo tem sido amplamente estudado [Aladağ *et al*., 2018]. No entanto, com a ascensão das redes sociais, como Instagram [Arendt *et al*., 2021] e X (antigo *Twitter*) [Sinyor *et al*., 2021], tornou-se comum a manifestação de tristeza, desilusões e pensamentos suicidas nesses espaços digitais.

Nos últimos anos, tecnologias digitais têm se expandido na saúde mental, especialmente na prevenção do suicídio. Aplicativos móveis têm contribuído para melhorar o acesso aos cuidados e oferecer suporte personalizado a indivíduos em risco [Braciszewski, 2021; Bhaumik *et al*., 2023; Diniz *et al*., 2022].

Com o aumento do uso de redes sociais e dispositivos móveis, a Fenotipagem Digital (FD) possibilita a identificação de traços digitais de pensamento suicida, analisando uma ampla gama de dados comportamentais [Souza *et al*., 2021]. A FD surge como uma abordagem inovadora, definida como a quantificação momentânea do fenótipo humano ao nível individual, mediante a utilização de dados provenientes de *smartphones* e outros dispositivos digitais pessoais [Torous *et al*., 2016].

Aliada à FD, a Inteligência Artificial (IA) é uma abordagem promissora para a predição de manifestações digitais de transtornos mentais. Essa abordagem permite a análise de

grandes volumes de dados textuais e a identificação precoce de sinais de ansiedade e depressão em redes sociais [Shatte *et al*., 2019].

Os sistemas baseados em IA geralmente utilizam modelos de Aprendizado de Máquina (ML, do inglês *Machine Learning*) [Gohel *et al*., 2021]. Nesse contexto, surge a questão da caixa fechada *versus* caixa de vidro na aplicação desses modelos [de Oliveira *et al*., 2025]. Enquanto a abordagem tradicional opera como uma caixa fechada, fornecendo respostas sem explicar os processos internos [Gohel *et al*., 2021], modelos de caixa de vidro, como os lineares [Weisberg, 2005] e os baseados em árvores de decisão [Safavian and Landgrebe, 1991], oferecem maior interpretabilidade. No entanto, esses modelos geralmente apresentam desempenho inferior aos modelos de caixa fechada [Linardatos *et al*., 2021].

A Inteligência Artificial Explicável (XAI, do inglês *Explainable Artificial Intelligence*) busca tornar os modelos de IA mais transparentes sem comprometer o desempenho [Adadi and Berrada, 2018]. A transparência é essencial em áreas críticas como saúde, justiça e finanças, permitindo a avaliação do desempenho e a justificativa das decisões dos modelos [Rudin, 2019; Gohel *et al*., 2021]. A interpretabilidade, por sua vez, refere-se à capacidade de explicar os resultados de forma compreensível [Linardatos *et al*., 2021].

A falta de explicabilidade em redes neurais profundas pode gerar desconfiança. Por isso, torna-se fundamental aprimorar a transparência para profissionais da saúde mental, explicando como os modelos identificam padrões de ideação suicida [Adadi and Berrada, 2020; de Oliveira *et al*., 2022; de Oliveira *et al*., 2025]. A clareza na fundamentação das respostas é essencial para fortalecer a confiança em aplicações médicas, garantindo decisões informadas e confiáveis [Gohel *et al*., 2021; de Oliveira *et al*., 2025]. Apesar dos avanços, ainda há desafios relacionados à generalização dos modelos para contextos clínicos, à explicabilidade robusta das decisões e à validação interdisciplinar com profissionais de saúde [de Oliveira *et al*., 2025].

Diante disso, este trabalho trata de um estudo na área de IA, no qual foi avaliado o desempenho dos Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) na identificação de ideação suicida em textos não clínicos, ou seja, textos produzidos fora de contextos formais de saúde, no idioma inglês [Calvo *et al*., 2017]. No escopo desta pesquisa, os textos não clínicos analisados foram extraídos de redes sociais (Reddit).

De forma mais específica, foram analisadas ferramentas baseadas em LLMs generativos, a saber, *Microsoft Copilot*, *Google Gemini*, *Deepseek* e ChatGPT-4o. Também foram considerados modelos baseados em *Bidirectional Encoder Representations from Transformers* (BERT), a saber, BERT-*Base*, BERT-*Large* [Devlin *et al*., 2019], *multilingual* BERT (mBERT) [Pires *et al*., 2019], *A Lite* BERT (AlBERT) [Lan *et al*., 2020], *Robustly Optimized BERT Approach* (RoBERTa) [Liu *et al*., 2019] e Mental-RoBERTa [Ji *et al*., 2022]. Adicionalmente, investigou-se a importância de *features* (palavras) de três sentenças no modelo Mental-RoBERTa, e uma sentença no modelo BERT-large, utilizando o método *Local Interpretable Model-Agnostic Explanations* (LIME) [Ribeiro *et al*., 2016].

O restante deste artigo está organizado da seguinte forma. Os trabalhos relacionados são discutidos na Seção 2. A Seção 3 descreve a metodologia utilizada no estudo, enquanto a Seção 4 apresenta e discute os resultados. Por fim, a Seção 5 conclui o artigo, apresentando também planos para trabalhos futuros.

2 Trabalhos Relacionados

Em um estudo conduzido porTang *et al*. [2024] foram exploradas informações médicas para avaliar a eficácia da XAI na previsão do risco de suicídio. Utilizando o método *SHapley Additive exPlanations* (SHAP) com *Random Forest*, os autores identificaram os principais fatores que influenciam o risco suicida, a saber, raiva, depressão e isolamento social. Por outro lado, indivíduos com renda elevada, profissionais conceituados e com ensino superior apresentaram menor risco.

Nielsen *et al*. [2023] investigaram preditores de tentativas de suicídio e suicídio. Os resultados indicaram que o risco de tentativa de suicídio aumentava tentativas anteriores. O risco de suicídio, por sua vez, aumentava inicialmente, mas diminuía após múltiplas tentativas. Fatores sociodemográficos e transtornos do humor apresentaram impactos distintos nos desfechos. De forma semelhante, Nordin *et al*. [2023] avaliaram preditores de tentativas de suicídio usando SHAP em modelos *Random Forest* e *Gradient Boosting*. Os principais preditores identificados foram histórico de tentativas, ideação suicida e etnia.

Gholi Zadeh Kharrat *et al*. [2024] investigaram a previsão de risco de suicídio na população de Quebec entre 2002 e 2019. Para isso, utilizaram modelos de, como *Logistic Regression*, *Random Forest*, *XGBoost* e *Multi Layer Perceptron*. A análise SHAP identificou variáveis-chave, como idade, consultas especializadas e psicoterapia psiquiátrica. Esses resultados evidenciam desafios na identificação do risco de suicídio e no acesso aos serviços de saúde mental [Gholi Zadeh Kharrat *et al*., 2024].

Lekkas *et al*. [2021] aplicaram modelos de ML para prever ideação suicida aguda no Instagram. A análise SHAP revelou que o engajamento e o número de seguidores são preditores relevantes. Esses achados sugerem padrões psicológicos associados ao tipo de conteúdo publicado.

No estudo de Chadaga *et al*. [2024], técnicas de ML foram usadas para identificar o Transtorno de Ansiedade Social (TAS). As análises mediante SHAP, LIME, ELI5 (“*Explain Like I’m Five*”) [Fan *et al*., 2019] e QLattice destacaram como principais atributos o questionário da Escala de Ansiedade Social de Liebowitz e o medo de falar em público.

Oliveira *et al*. [2022], mediante uso de ELI5, constataram que *Extra Trees* e *Random Forest* foram influenciados de forma semelhante por características compostas por um ou dois termos. Já o *Support Vector Machine* foi mais influenciado por características compostas por dois termos. De modo geral, termos como “suicídio”, “desejo de tirar a própria vida” e “tristeza” tiveram maior importância na indicação positiva para ideação suicida.

No contexto de LLMs, o estudo de Malhotra and Jindal [2024] analisou a interpretabilidade de modelos ajustados para detectar comportamentos depressivos e suicidas, utilizando as técnicas SHAP e LIME. Foram avaliados modelos como BERT, DistilBERT, RoBERTa, Mental-BERT, Psych-

BERT e PHSBERT. As técnicas permitiram compreender as causas dos erros de classificação e forneceram *insights* sobre a qualidade dos dados de treinamento. Observou-se que falsos positivos ocorreram apenas nos modelos treinados em um dos quatro conjuntos de dados, permitindo, por meio dos valores SHAP, uma análise detalhada das palavras ou características que influenciaram essas classificações errôneas.

Oliveira *et al*. [2024] analisaram o desempenho de três variações de modelos BERT e LLMs (Google Bard, Microsoft Bing/GPT-4 e OpenAI ChatGPT-3.5) na identificação de ideação suicida em textos não clínicos escritos em Português Brasileiro (PT-BR). Bing/GPT-4 obteve o melhor desempenho (98% de acurácia), seguido pelos modelos BERT ajustados: BERTimbau-Large (96%), BERTimbau-Base (94%) e mBERT (87%). ChatGPT-3.5 alcançou 81%, enquanto Bard teve o pior desempenho (62%). A alta capacidade de *recall* dos modelos sugeriu baixa taxa de erros de classificação de pacientes em risco, fator essencial para prevenir falhas de intervenção.

Este estudo apresenta semelhanças e diferenças em relação aos trabalhos mencionados. Assim como os estudos de assim como os estudos de Tang *et al*. [2024], Nielsen *et al*. [2023], Nordin *et al*. [2023], Gholi Zadeh Kharrat *et al*. [2024] e Oliveira *et al*. [2024] , foca na identificação de ideação suicida utilizando modelos de ML e técnicas explicáveis. Ademais, foi adotada a técnica LIME para a interpretabilidade dos melhores modelos. Essa abordagem também foi utilizada no estudo de Chadaga *et al*. [2024].

Diferentemente dos estudos mencionados que utilizam informações médicas ou dados estruturados, como Tang *et al*. [2024] e Nielsen *et al*. [2023], este trabalho limita-se à análise de como palavras de algumas sentenças influenciam os modelos BERT. Enquanto os estudos de Oliveira *et al*. [2024] analisaram o desempenho de LLMs na identificação de ideação suicida em textos de PT-BR, este estudo analisou o desempenho dos LLMs *Copilot*, *Gemini*, *Deepseek* e ChatGPT-4o, BERT-*Base*, BERT-*Large*, mBERT, AlBERT, RoBERTa e Mental-RoBERTa em dois conjuntos de dados de idioma inglês. Diferentemente de Chadaga *et al*. [2024], que usou o SHAP, e de Oliveira *et al*. [2022], que usou ELI5, este estudo emprega somente o LIME como ferramenta explicável. Essa abordagem é usada para avaliar a influência de palavras-chave nas predições dos melhores modelos LLMs.

Por fim, identificamos limitações, como dependência de palavras-chave e dificuldades no processamento de textos longos e ambíguos, aspectos pouco discutidos nos trabalhos anteriores.

3 Materiais e Métodos

Esta seção descreve os materiais[1] e procedimentos metodológicos, em que as seguintes etapas foram adotadas: (*i*) Seleção de conjuntos de dados, (*ii*) Pré-processamento de dados; (*iii*) Criação dos modelos; (*iv*) e Avaliação e Explicação de modelos.

[1]Disponível em `https://github.com/adonias-caetano/adonias-caetano-ifce-pibic-llms-suicide.git`

3.1 Seleção e preparação de conjunto de dados

Dois conjuntos de dados textuais foram utilizados neste estudo. O primeiro conjunto de dados foi o *"Suicide and Depression Detection"* (SDD) [Nikhileswar *et al*., 2021], disponível no repositório Kaggle [Komati, 2021]. Ele é composto por postagens dos *subreddits* "*SuicideWatch*" e "*depression*" da rede social *Reddit*. O conjunto possui duas colunas principais: "*text*", que contém as postagens, e "*class*", que indica se o conteúdo é classificado como "*suicide*" ou'*non-suicide*" [Nikhileswar *et al*., 2021].

O processo de rotulação desse conjunto não foi validado por profissionais de saúde mental. Essa limitação pode introduzir ruídos nas classes, pois a ideação suicida nem sempre é explicitamente declarada nos textos. Além disso, postagens em fóruns *online* podem conter ambiguidades, ironias ou expressões indiretas, o que dificulta a distinção entre conteúdos depressivos e suicidas.

O segundo conjunto de dados foi o *"Suicide vs Depression Classification"* (SDCNL) [Haque *et al*., 2021b,a]. Ele reúne postagens de usuários extraídas de *subreddits* "*r/SuicideWatch*" e "*r/depression*" da rede social *Reddit*, resultando em 1.895 amostras. Nesse conjunto, textos do "*r/SuicideWatch*" foram rotulados como suicidas, enquanto os do "*r/Depression*" foram classificados como depressivos.

Para validar a correção de rótulos, foi utilizado o conjunto de dados *Reddit Suicide C-SSRS*, composto por 500 postagens do *subreddit* "*r/Depression*", classificadas por psicólogos de acordo com a *Columbia Suicide Severity Rating Scale*. Adicionalmente, a validação foi reforçada com o conjunto de dados *IMDB Large Movie Dataset* [Pal *et al*., 2020], um *benchmark* de Processamento de Linguagem Natural (PLN) contendo 50.000 avaliações polarizadas de filmes. O SDCNL resulta na combinação desses conjuntos, totalizando 8.853 sentenças. Ele foi adaptado para uma tarefa de classificação binária de textos clinicamente saudáveis e suicidas. Para isso, foram incluídas postagens dos *subreddits* "*r/SuicideWatch*" e "*r/CasualConversation*". Este, por sua vez, um *subreddit* de conversas gerais usado como referência para textos de classe clinicamente saudável.

As amostras resultantes do SDCNL foram organizadas em duas colunas: *"text"*, contendo o conteúdo textual da postagem, e *"target"*, com o rótulo atribuído à amostra. As postagens foram rotuladas por especialistas de forma binária, sendo "0" atribuído à ausência de ideação suicida (classe negativa) e "1" à presença de ideação suicida (classe positiva) [Haque *et al*., 2021a]. Neste trabalho, esses rótulos foram ajustados como "*suicide*" e "*non-suicide*", a fim de manter consistência com o SDD.

Em ambos os conjuntos de dados, foram aplicadas técnicas de PLN para reduzir ruídos e padronizar os textos. Essas técnicas incluem remoção de elementos irrelevantes (endereços eletrônicos, emojis e caracteres especiais), normalização para minúsculas, tokenização, remoção de *stopwords*, os quais não contribuem para a análise, e *stemming*, para reduzir as palavras à sua forma básica [Birjali *et al*., 2021].

Todos os experimentos computacionais de preparação de dados, ajuste fino (do inglês *fine-tuning*) de modelos BERT e avaliação dos *chatbots* baseados em LLMs foram realiza-

dos no ambiente de desenvolvimento *Google Colaboratory* utilizando a linguagem de programação Python.

Diante das limitações dos dados, o pré-processamento se torna uma etapa essencial. Ele contribui para melhorar a qualidade das representações textuais e, consequentemente, o desempenho dos modelos.

3.2 Avaliação e Explicação dos modelos

Para o conjunto de dados SDD, foi realizado o ajuste fino de quatro modelos baseados em BERT: BERT-*Base*, RoBERTa, Mental-RoBERTa e AlBERT. Para a avaliação desses modelos, o conjunto SDD foi dividido em 80% para treinamento e 20% para testes. Do conjunto de treinamento, 20% foram reservados para validação.

No conjunto SDCNL, foram aplicados os modelos BERT-*Large*, mBERT, RoBERTa, *Mental*-RoBERTa e AlBERT. Devido à dimensão do SDCNL, os modelos foram treinados com 1.795 sentenças (930 frases suicidas e 865 frases não suicidas) e testados com 100 sentenças, pré-selecionadas aleatoriamente (50 sentenças de cada classe). Do subconjunto de treinamento, 20% foram reservados para validação em cada época de treinamento.

O ajuste fino foi realizado ao final de cada época utilizando o conjunto de validação. Foi adotado tamanho de lote (*batch size*) de 16, com ajustes nos pesos para preservar a generalização. Todos os modelos foram configurados com taxa de aprendizagem de 2e-6, visando estabilidade na convergência. As entradas foram tokenizadas com limite de 128 tokens, garantindo consistência na representação vetorial. Dos modelos aplicados no conjunto SDD, somente o AlBERT teve um ajuste fino em 13 épocas, enquanto os demais modelos foram treinados ao longo de 4 épocas. Já os modelos aplicados no SDCNL, somente o mBERT foi treinado em 10 épocas e os demais modelos em 8 épocas.

Diante do desempenho da maioria dos modelos BERT no conjunto de dados SDCNL, este trabalho explorou os *chatbots* baseados em LLMs generativos, a saber, ChatGPT-4o, *Gemini* 1.5 *Flash*, *Copilot* (*Azure Cognitive Services*), e o *DeepSeek*.

Foram selecionadas 100 sentenças (50 de cada classe) do conjunto de teste SDCNL, a fim de serem utilizadas no teste das ferramentas baseadas em LLM generativos. Essa abordagem foi inspirada no trabalho de Oliveira *et al*. [2024] por buscar garantir o equilíbrio entre as classes e reduzir viés de amostragem. Adicionalmente, essa amostragem reduzida justifica-se pela limitação de inserção dos dados no modelo DeepSeek, garantindo assim uma comparação justa e consistente entre todos os modelos avaliados para esse conjunto de dados.

A aplicação dos LLMs generativos exigiu o uso da engenharia de *prompts* para direcionar a saída dos modelos [Heston and Khun, 2023]. Foi adotada uma abordagem *zero-shot*, sem exemplos de sentenças positivas ou negativas, conforme a Tabela 1. Esta escolha metodológica visou avaliar a capacidade de raciocínio intrínseco e o conhecimento latente dos modelos sobre sinais de ideação suicida.

Todos os modelos generativos, com exceção do *Deepseek*, foram utilizados por meio de via *Application Programming Interface* (API) no ambiente *Google Colaboratory*, utilizando requisições *Hypertext Transfer Protocol* (HTTP) no formato *prompt-response* e *scripts Python* para automatizar as consultas. O *Gemini* e o O ChatGPT-4o foram acessados via API *Key* com uso do *prompt* definido, enquanto o *Copilot* foi utilizado no mesmo ambiente sem necessidade de personalização adicional de prompt, devido a disponibilização do modelo pela plataforma da *Microsoft Azure*.

O *Deepseek* foi o único modelo acessado manualmente, devido à indisponibilidade de API pública estável no momento dos experimentos. Nesse caso, as 100 sentenças do conjunto de teste foram inseridas manualmente na interface do modelo, uma a uma, seguindo o mesmo prompt definido. Esse procedimento garantiu consistência experimental e comparabilidade com os demais modelos. As respostas geradas foram registradas manualmente e organizadas para análise quantitativa.

Tabela 1. Comando de *Prompt* aplicado

"Você é um modelo de IA treinado para classificar frases quanto à ideação suicida. Regras: Se a frase indicar ideação suicida, retorne exatamente '1'. Se a frase não indicar ideação suicida, retorne exatamente '0'. Não adicione explicações ou qualquer outro texto. Somente retorne '0' ou '1'. Frase: 'frase'."

O desempenho dos modelos foi avaliado por meio das métricas de acurácia, precisão, sensibilidade e medida-F1 [Sokolova and Lapalme, 2009]. O modelo Mental-RoBERTa foi selecionado para explicação com LIME por apresentar o melhor desempenho em termos de medida-F1. Essa técnica XAI visa tornar modelos complexos mais compreensíveis ao destacar características relevantes nas predições. Aproximando o comportamento do modelo original por meio de perturbações e utilizando um modelo linear simplificado ajustado à amostra específica [Molnar, 2020]. Para realizar a interpretação do modelo foram selecionadas duas sentenças do conjunto SDD (uma de cada classe) classificadas corretamente.

4 Resultados e Discussão

A Tabela 2 apresenta o desempenho de quatro modelos BERT no conjunto SDD. Os modelos AlBERT e Mental-RoBERTa obtiveram um desempenho de 99% em todas as métricas. Esse resultado indica desempenho superior em relação aos demais. O BERT-*Base* apresentou o segundo melhor desempenho, com valores entre 97% e 98%. Esse modelo apresentou maior precisão para sentenças sem ideação suicida. Também obteve melhor sensibilidade para sentenças com ideação suicida em comparação ao RoBERTa.

Todos os modelos apresentaram elevada sensibilidade. Isso indica capacidade de identificar corretamente a maioria dos textos com ideação suicida. Esse aspecto é essencial em cenários clínicos e de prevenção. Além disso, a precisão igualmente alta indica que os modelos minimizam falsos positivos, contribuindo para uma triagem mais confiável.

Na Tabela 3, o BERT-*Large* e AlBERT se destacam com acurácia de 64%, superior aos demais modelos BERT. O BERT-*Large* apresenta a maior precisão (78%) para sentenças sem ideação suicida, enquanto AlBERT mantém equilíbrio entre precisão e sensibilidade, apresentando uma medida-F1 de 63% para sentenças suicidas, tornando-se mais confiável para essa detecção.

Tabela 2. Desempenho dos LLMs na identificação de ideação suicida no SDD

Modelos	Classes	Acurácia	Precisão	Sensibilidade	Medida-F1
BERT-*Base*	*Non-suicide*	0,98	0,98	0,97	0,97
	Suicide		0,97	0,98	0,98
AlBERT	*Non-suicide*	0,99	0,99	0,99	0,99
	Suicide		0,99	0,99	0,99
RoBERTa	*Non-suicide*	0,97	0,97	0,97	0,97
	Suicide		0,97	0,97	0,97
Mental-RoBERTa	*Non-suicide*	0,99	0,99	0,99	0,99
	Suicide		0,99	0,99	0,99

Tabela 3. Desempenho dos LLMs na identificação de ideação suicida no SDCNL

Modelos	Classes	Acurácia	Precisão	Sensibilidade	Medida-F1
BERT-*Large*	*Non-suicide*	**0,64**	**0,78**	0,61	**0,68**
	Suicide		0,50	0,69	0,58
mBERT	*Non-suicide*	0,56	0,64	0,55	0,59
	Suicide		0,48	0,57	0,52
AlBERT	*Non-suicide*	**0,64**	0,66	0,63	0,65
	Suicide		0,62	0,65	0,63
RoBERTa	*Non-suicide*	0,63	0,62	0,63	0,63
	Suicide		0,64	0,63	0,63
Mental-RoBERTa	*Non-suicide*	**0,56**	**0,84**	0,54	0,66
	Suicide		**0,28**	0,64	0,39
Gemini 1,5 *flash*	*Non-suicide*	0,59	0,30	0,71	0,42
	Suicide		**0,88**	0,56	**0,68**
Copilot	*Non-suicide*	0,54	0,44	0,55	0,49
	Suicide		0,64	0,53	0,58
DeepSeek	*Non-suicide*	**0,70**	0,54	0,79	**0,64**
	Suicide		0,86	0,65	**0,74**
ChatGPT-4o	*Non-suicide*	0,68	0,48	**0,80**	0,60
	Suicide		**0,88**	0,63	0,73

O RoBERTa, por sua vez, exibe métricas estáveis em ambas as classes, sugerindo consistência, embora sem superar os melhores resultados. O mBERT obteve desempenho inferior, com acurácia de 56% e medida-F1 de 52% na classe suicida, indicando menor capacidade de generalização no contexto analisado. O Mental-RoBERTa apresentou um comportamento contrastante. O modelo alcançou alta precisão (84%) para a classe não suicida. No entanto, seu desempenho caiu significativamente para a classe suicida, com medida-F1 de 39%. Esse resultado evidencia dificuldades em lidar com cenários desbalanceados e limita sua aplicabilidade em tarefas críticas de detecção de ideação suicida.

O *DeepSeek* apresentou o melhor desempenho entre os *chatbots* de LLMs generativos, com 70% de acurácia, seguido pelo ChatGPT-4o (68%). Ambos demonstram alta sensibilidade na identificação de sentenças suicidas, indicando capacidade de detectar padrões linguísticos associados à ideação suicida. Por outro lado, o *Gemini* apresenta um desempenho misto: sua sensibilidade para sentenças não suicidas foi alta (71%), porém sua precisão é baixa (30%), indicando alto número de falsos positivos. Adicionalmente, o *Copilot* apresentou o menor desempenho entre os LLMs generativos avaliados. Nesse sentido, o elevado número de falsos positivos observado no *Copilot* pode gerar alarmes excessivos em aplicações práticas, sobrecarregando profissionais de saúde.

O menor desempenho dos LLMs no conjunto de dados SDCNL quando comparado com SDD levanta algumas suposições. Apesar da validação parcial por especialistas, a construção do SDCNL envolve a combinação de múltiplas fontes. Isso pode introduzir heterogeneidade nos padrões linguísticos. Além disso, a associação direta entre *subreddits* e rótulos pode gerar vieses, já que nem todas as postagens refletem necessariamente a condição atribuída. A identificação das classes também é desafiadora devido à sobreposição entre sinais de depressão e ideação suicida. Por outro lado, a baixa sensibilidade do Mental-RoBERTa no SDCNL indica risco de falsos negativos. Em cenários clínicos, esse erro é crítico, pois pode resultar na não detecção de indivíduos em risco.

O LIME foi utilizado para analisar as predições do modelo e a relevância de cada *token*. A Figura 1 apresenta a seguinte sentença: "*Finally 2020 is almost over... So I can never hear* "2020 *has been a bad year*"*ever again. I swear to fucking God its so annoying*.". O Mental-RoBERTa classificou a sentença corretamente como 99% não suicida. O LIME destacou palavras que impactaram a decisão do modelo, como "2020", "*So*", "*hear*", "*year*", "*again*", "*its*", e "*annoying*". Essas palavras tiveram um peso na classificação, indicando que a estrutura e o contexto influenciaram a decisão.

De forma análoga, a Figura 2 exibe a análise explicativa da classificação da sentença: "*My family doesn't know that I want to end my life*." Nesse caso, o Mental-RoBERTa classificou corretamente o texto como suicida, com 98% de probabilidade. Os termos "*end*" (0,39) e "*life*" (0,27) apresentaram maior contribuição para a classificação, sugerindo

Figura 1. Explicação do Mental-RoBERTa com LIME para sentença não suicida

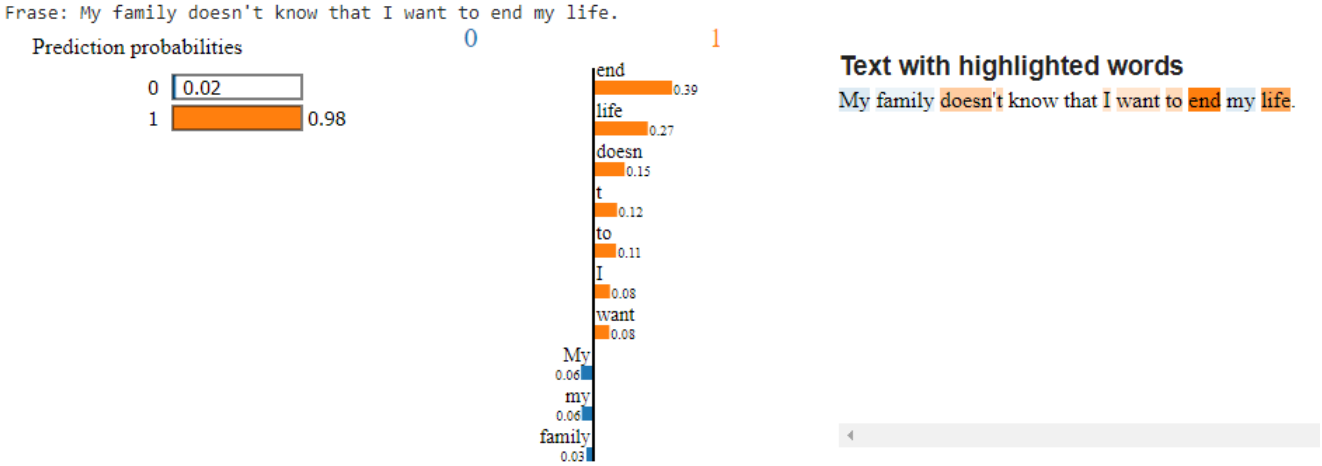

Figura 2. Explicação do Mental-RoBERTa com LIME para sentença suicida

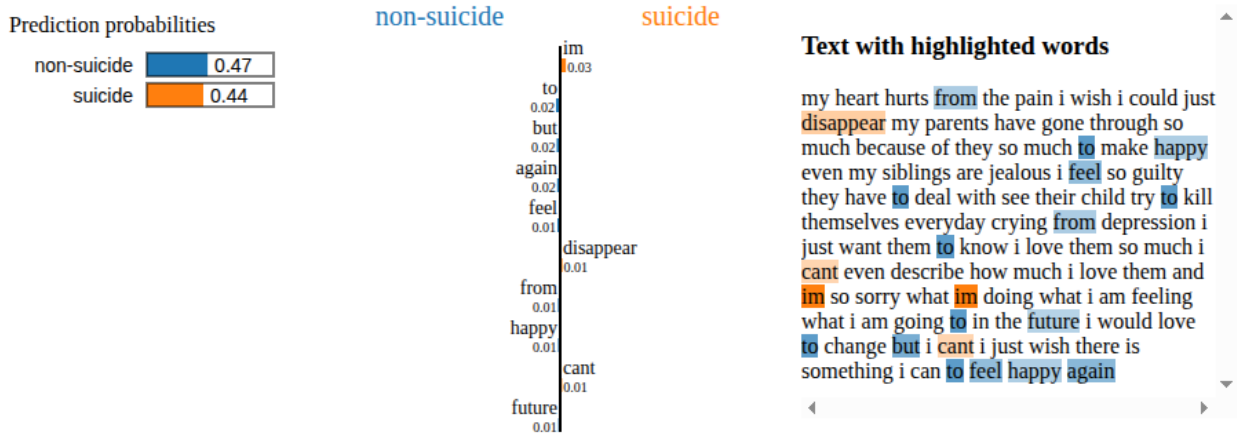

Figura 3. Explicação do BERT-Large com LIME para o conjunto SDCNL

que o modelo identificou esses termos como indicativos de ideação suicida. O termo "*doesn't*" também aparece destacado, indicando que negações influenciam a interpretação do modelo.

Esses resultados demonstram a eficácia do Mental-RoBERTa no SDD, especialmente em sentenças curtas No entanto, também revelam dependência de palavras-chave associadas à ideação suicida. Essa característica pode limitar o desempenho do modelo em contextos mais complexos, nos quais coexistem expressões de sofrimento e de afeto. Essa limitação torna-se evidente no conjunto de dados SDCNL, em que os textos analisados são mais extensos e apresentam maior ambiguidade semântica, desafiando os modelos a manter consistência em suas predições.

No conjunto SDCNL, o LIME foi aplicado para analisar as predições. A Figura 3 apresenta o gráfico de importância das *features* do BERT-*Large* para a sentença: "*my heart hurts from the pain i wish i could just disappear my parents have gone through so much because of they so much to make happy even my siblings are jealous i feel so guilty they have to deal with see their child try to kill themselves everyday crying from depression i just want them to know i love them so much i cant even describe how much i love them and im so sorry what im doing what i am feeling what i am going to in the future i would love to change but i cant i just wish there is something i can to feel happy again*".

O modelo evidenciou incerteza na classificação ao atribuir probabilidades próximas para ambas as classes: 0,47 para não suicida e 0,44 para suicida. Termos como "*disappear*" e "*can't*" tiveram maior contribuição para a classe suicida,

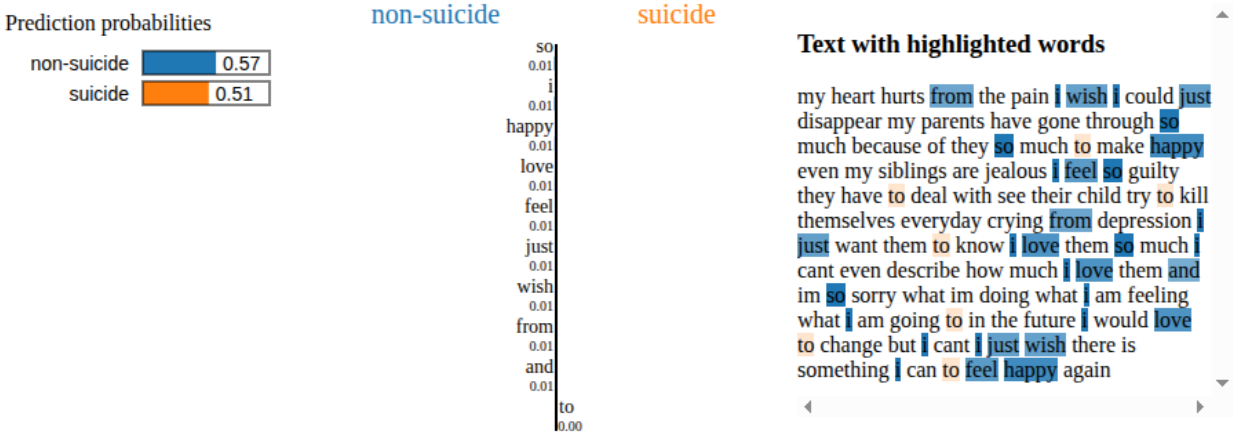

Figura 4. Explicação do Mental-RoBERTa com LIME para o conjunto SDCNL

refletindo sentimentos de desesperança. Por outro lado, palavras como "*happy*", "*feel*", "*future*" e "*to feel happy again*" favoreceram a classificação como não suicida, ilustrando a ambiguidade semântica presente no texto.

De forma semelhante, a Figura 4 mostra o gráfico de importância dos atributos do Mental-RoBERTa para a mesma sentença. O modelo apresentou *scores* ainda mais próximos entre as classes (0,57 para não suicida e 0,51 para suicida), destacando expressões como "*just wish*", "*feel happy*", "*feel so*" e "*i love*" para a classe não suicida, termos associados a afeto e desejo de mudança.

Os resultados mostram que os modelos identificam termos associados à ideação suicida. No entanto, a presença simultânea de sofrimento e esperança dificulta a decisão final. Isso revela vulnerabilidade em textos longos e emocionalmente complexos.

Esses achados reforçam a presença de erros sistemáticos nos modelos, sobretudo em textos que combinam expressões de sofrimento e de esperança. Casos envolvendo negação, ironia ou linguagem figurada também se mostraram desafiadores, apontando para limitações na compreensão semântica mais profunda dos LLMs.

5 Conclusão

Este estudo avaliou o desempenho de LLMs na identificação de ideação suicida em textos em inglês, demonstrando resultados promissores. No conjunto SDD, AlBERT e Mental-RoBERTa alcançaram 99% em todas as métricas, destacando-se pela precisão e sensibilidade. No SDCNL, BERT-*Large* foi eficaz na minimização de falsos positivos, enquanto AlBERT manteve um equilíbrio entre suas métricas.

Entre os modelos generativos, *DeepSeek* e ChatGPT-4o apresentaram maior acurácia e sensibilidade na identificação de padrões suicidas. Os achados indicam que a implementação de LLMs pode aprimorar sistemas automatizados de suporte à saúde mental, possibilitando detecção precoce e intervenções mais assertivas. Além disso, a análise do LIME evidenciou alta sensibilidade em detectar padrões suicidas e a interpretabilidade. Por conseguinte, tanto o Mental-RoBERTa quanto o BERT-*Large* são fortemente influenciados pelo contexto e pela presença de termos relacionados à ideação suicida, indicando que ambos os modelos são confiáveis em textos curtos e objetivos. No entanto, apresentam limitações na classificação de textos longos ou semanticamente ambíguos. Esses fatores restringem a generalização para cenários reais de saúde mental.

Como trabalho futuro, técnicas de interpretabilidade como SHAP e Captum podem ser exploradas. Essas abordagens podem fornecer insights sobre a influência das entradas nas predições. Isso pode aumentar a transparência, a robustez e a aplicabilidade prática. Além disso, LLMs generativos, tais com LLaMa e Qwen, podem ser avaliados nesta tarefa de classificação.

Declarações complementares

Agradecimentos

Agradecemos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC/IFCE) e ao Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI/IFCE) pelo apoio financeiro concedido mediante as bolsas de pesquisa, permitindo o desenvolvimento deste estudo.

Financiamento

Esta pesquisa foi financiada pelo PIBIC/IFCE e PIBITI/IFCE, por meio da concessão de bolsas de pesquisa que viabilizaram o desenvolvimento deste estudo.

Contribuições dos autores

Lais CC: Coleta de dados; organização do material empírico; experimentos computacionais; análise dos resultados; redação, edição e revisão do manuscrito. Antonia Estefane RV: Coleta de dados; organização do material empírico; experimentos computacionais; análise dos resultados; redação, edição e revisão do manuscrito. Rosana CPD: Organização dos dados; redação, edição e revisão do manuscrito. Adonias CO: Concepção e coordenação do estudo; definição da metodologia; supervisão dos experimentos computacionais; análise e interpretação dos resultados; redação, edição, revisão e submissão do manuscrito.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante o estudo atual estão disponíveis em `https://github.com/adonias-caetano/adonias-caetano-ifce-pibic-llms-suicide.git`.

Outras informações relevantes

Durante a preparação deste trabalho, os autores usaram o Microsoft Copilot e ChatGPT para aprimorar a redação e estrutura do texto, além de corrigir erros ortográficos e gramaticais. Após utilizar esta ferramenta/serviço, os autores revisaram e editaram o conteúdo conforme necessário e assumem total responsabilidade pelo conteúdo do artigo publicado.

Referências

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

Adadi, A. and Berrada, M. (2020). Explainable ai for healthcare: From black box to interpretable models. In Bhateja, V., Satapathy, S. C., and Satori, H., editors, *Embedded Systems and Artificial Intelligence*, pages 327–337, Singapore. Springer Singapore. DOI: 10.1007/978-981-15-0947-6_31.

Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O., and Bingol, H. O. (2018). Detecting suicidal ideation on forums: Proof-of-concept study. *J Med Internet Res*, 20(6):e215. DOI: 10.2196/jmir.9840.

Arendt, F., Markiewitz, A., and Scherr, S. (2021). Investigating Suicide-Related Subliminal Messages on Instagram. *Crisis*, 42(4):263–269. DOI: 10.1027/0227-5910/a000717.

Bertolote, J. (2016). *O suicídio e sua prevenção*. Editora Unesp.

Bhaumik, R., Srivastava, V., Jalali, A., Ghosh, S., and Chandrasekharan, R. (2023). Mindwatch: A smart cloud-based ai solution for suicide ideation detection leveraging large language models. *medRxiv*. DOI: 10.1101/2023.09.25.23296062.

Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134. DOI: 10.1016/j.knosys.2021.107134.

Braciszewski, J. M. (2021). Digital technology for suicide prevention. *Advances in psychiatry and behavioral health*, 1(1):53–65. DOI: 10.1016/j.ypsc.2021.05.008.

Calvo, R., Milne, D., Hussain, S., and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, pages 1–37. DOI: 10.1017/S1351324916000383.

Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., Bhat, D., Sharma, A. K., and Swathi, K. (2024). SADXAI: Predicting social anxiety disorder using multiple interpretable artificial intelligence techniques. *SLAS Technology*, 29(2):100129. DOI: 10.1016/j.slast.2024.100129.

de Oliveira, A. C., Azevedo, J. P. C., Ruback, L., Moreira, R., Teixeira, S. S., and Teles, A. S. (2025). Effect of explainable artificial intelligence on trust of mental health professionals in an ai-based system for suicide prevention. *IEEE Access*, 13:60987–61005. DOI: 10.1109/ACCESS.2025.3556245.

de Oliveira, A. C., Diniz, E. J., Teixeira, S., and Teles, A. S. (2022). How can machine learning identify suicidal ideation from user's texts? towards the explanation of the boamente system. *Procedia Computer Science*, 206:141–150. International Society for Research on Internet Interventions 11th Scientific Meeting. DOI: 10.1016/j.procs.2022.09.093.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Disponível em: `https://arxiv.org/abs/1810.04805`.

Diniz, E. J., Fontenele, J. E., de Oliveira, A. C., Bastos, V. H., Teixeira, S., Rabêlo, R. L., Calçada, D. B., Dos Santos, R. M., de Oliveira, A. K., and Teles, A. S. (2022). Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. In *Healthcare*, volume 10, page 698. MDPI. DOI: 10.3390/healthcare10040698.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). ELI5: long form question answering. *CoRR*, abs/1907.09190. Disponível em: `http://arxiv.org/abs/1907.09190`.

Gholi Zadeh Kharrat, F., Gagne, C., Lesage, A., Gariépy, G., Pelletier, J.-F., Brousseau-Paradis, C., Rochette, L., Pelletier, E., Lévesque, P., Mohammed, M., *et al*. (2024). Explainable artificial intelligence models for predicting risk of suicide using health administrative data in quebec. *PLoS one*, 19(4):e0301117. DOI: 10.1371/journal.pone.0301117.

Gohel, P., Singh, P., and Mohanty, M. (2021). Explainable ai: current status and future directions. *arXiv preprint arXiv:2107.07045*. DOI: 10.48550/arXiv.2107.07045.

Haque, A., Reddi, V., and Giallanza, T. (2021a). Deep learning for suicide and depression identification with unsupervised label correction. In Farkaš, I., Masulli, P., Otte, S., and Wermter, S., editors, *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 436–447, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-86383-8_35.

Haque, A., Reddi, V., and Giallanza, T. (2021b). SDCNL (Suicide vs Depression Classification). Disponível em: `https://github.com/ayaanzhaque/SDCNL`. Acesso em 05 de junho de 2025.

Heston, T. F. and Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, 2(3):198–205. DOI: 10.3390/ime2030019.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association. Disponível em: `https://aclanthology.org/2022.lrec-1.778/`.

Komati, N. (2021). Suicide and depression detection. Repositório Kaggle. Disponível em: `https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch`. Acesso em 05 de junho de 2025.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. Disponível em: `https://arxiv.org/abs/1909.11942`.

Lekkas, D., Klein, R. J., and Jacobson, N. C. (2021). Pre-

dicting acute suicidal ideation on instagram using ensemble machine learning models. *Internet Interventions*, 25:100424. DOI: 10.1016/j.invent.2021.100424.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1). DOI: 10.3390/e23010018.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. Disponível em: `https://arxiv.org/abs/1907.11692`.

Malhotra, A. and Jindal, R. (2024). Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cognitive Systems Research*, 84:101186. DOI: 10.1016/j.cogsys.2023.101186.

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub.

Neeleman, J., de Graaf, R., and Vollebergh, W. (2004). The suicidal process; prospective comparison between early and later stages. *Journal of affective disorders*, 82(1):43–52. DOI: 10.1016/j.jad.2003.09.005.

Nielsen, S. D., Christensen, R. H. B., Madsen, T., Karstoft, K.-I., Clemmensen, L., and Benros, M. E. (2023). Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide danish registers. *Acta Psychiatrica Scandinavica*, 148(6):525–537. DOI: 10.1111/acps.13629.

Nikhileswar, K., Vishal, D., Sphoorthi, L., and Fathimabi, S. (2021). Suicide ideation detection in social media forums. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1741–1747. DOI: 10.1109/ICOSEC51865.2021.9591887.

Nordin, N., Zainol, Z., Mohd Noor, M. H., and Chan, L. F. (2023). An explainable predictive model for suicide attempt risk using an ensemble learning and shapley additive explanations (shap) approach. *Asian Journal of Psychiatry*, 79:103316. DOI: 10.1016/j.ajp.2022.103316.

Oliveira, A. C. d., Bessa, R. F., and Teles, A. S. (2024). Comparative analysis of bert-based and generative large language models for detecting suicidal ideation: a performance evaluation study. *Cadernos de Saúde Pública*, 40(10):e00028824. DOI: 10.1590/0102-311XEN028824.

Oliveira, A. C. d., Diniz, E. J., Teixeira, S., and Teles, A. S. (2022). How can machine learning identify suicidal ideation from user's texts? towards the explanation of the boamente system. *Procedia Computer Science*, 206:141–150. International Society for Research on Internet Interventions 11th Scientific Meeting. DOI: 10.1016/j.procs.2022.09.093.

Pal, A., Barigidad, A., and Mustafi, A. (2020). IMDb Movie Reviews Dataset. Disponível em: `https://dx.doi.org/10.21227/zm1y-b270`.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics. DOI: 10.18653/v1/P19-1493.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2939672.2939778.

Rizvi, A., Harmer, B., and Saadabadi, A. (2025). Suicidal Ideation. StatPearls Publishing, Treasure Island (FL). Disponível em: `https://www.ncbi.nlm.nih.gov/sites/books/NBK565877/`.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. DOI: 10.1038/s42256-019-0048-x.

Safavian, S. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674. DOI: 10.1109/21.97458.

Shatte, A. B. R., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9):1426–1448. DOI: 10.1017/S0033291719000151.

Sinyor, M., Williams, M., Zaheer, R., Loureiro, R., Pirkis, J., Heisel, M. J., Schaffer, A., Redelmeier, D. A., Cheung, A. H., and Niederkrotenthaler, T. (2021). The association between twitter content and suicide. *Australian & New Zealand Journal of Psychiatry*, 55(3):268–276. PMID: 33153274. DOI: 10.1177/0004867420969805.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

Souza, V., Nobre, J., and Becker, K. (2021). A deep learning ensemble to classify anxiety, depression, and their comorbidity from texts of social networks. *Journal of Information and Data Management*, 12. DOI: 10.5753/jidm.2021.1901.

Tang, H., Miri Rekavandi, A., Rooprai, D., Dwivedi, G., Sanfilippo, F. M., Boussaid, F., and Bennamoun, M. (2024). Analysis and evaluation of explainable artificial intelligence on suicide risk assessment. *Scientific Reports*, 14(1):6163. DOI: 10.1038/s41598-024-53426-0.

Torous, J., Kiang, M. V., Lorme, J., Onnela, J.-P., *et al.* (2016). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health*, 3(2):e5165. DOI: 10.2196/mental.5165.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley.