UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

João Pedro Cavalcanti Azevedo

Explorando Métodos de Aprendizado Semi-supervisionado e Inteligência Artificial Generativa na Identificação de Ideação Suicida em Textos Não-Clínicos

São Luís
2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO MARANHÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Explorando Métodos de Aprendizado Semi-supervisionado e Inteligência Artificial Generativa na Identificação de Ideação Suicida em Textos Não-Clínicos

João Pedro Cavalcanti Azevedo

São Luís, 2025

João Pedro Cavalcanti Azevedo

Explorando Métodos de Aprendizado Semi-supervisionado e Inteligência Artificial Generativa na Identificação de Ideação Suicida em Textos Não-Clínicos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ariel Soares Teles
UFMA/IFMA

São Luís
2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Cavalcanti Azevedo, João Pedro.
Explorando Métodos de Aprendizado Semi-supervisionado e Inteligência Artificial Generativa na Identificação de Ideação Suicida em Textos Não-Clínicos / João Pedro Cavalcanti Azevedo. - 2025.
72 f.

Orientador(a): Ariel Soares Teles.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luís, 2025.

1. Ideação Suicida. 2. Aprendizado Semi-supervisionado. 3. Modelos de Linguagem Generativos. 4. Saúde Mental. 5. Explicabilidade de Modelos. I. Soares Teles, Ariel. II. Título.

João Pedro Cavalcanti Azevedo

Explorando Métodos de Aprendizado Semi-supervisionado e Inteligência Artificial Generativa na Identificação de Ideação Suicida em Textos Não-Clínicos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Trabalho Aprovado. São Luís, 30 de Junho de 2025:

Orientador: Prof. Dr. Ariel Soares Teles
Orientador
Universidade Federal do Maranhão

Prof. Dr. Antônio de Abreu Batista Junior
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Ivandré Paraboni
Examinador Externo
Universidade de São Paulo

São Luís
2025

Dedicatória: Aos meus pais por todo sacrífico que fizeram para me tornar quem eu sou hoje.

Agradecimentos

Ao longo desta trajetória, tive o privilégio de contar com apoio, aprendizado e desafios que contribuíram imensamente para o meu crescimento acadêmico e pessoal. Por isso, faço questão de expressar minha gratidão a todos que, de alguma forma, contribuíram direta ou indiretamente para a realização deste trabalho.

Ao Prof. Dr. Ariel Soares Teles, minha profunda gratidão por sua orientação ao longo do mestrado. Agradeço pela paciência, confiança e dedicação, bem como pelo valioso direcionamento na condução desta pesquisa.

Aos ilustres membros da banca examinadora, Prof. Dr. Ariel Soares Teles, Prof. Dr. Antonio de Abreu Batista Junior e Prof. Dr. Ivandré Paraboni, expresso minha sincera gratidão pelo tempo dedicado à análise deste trabalho, bem como pelas valiosas contribuições e sugestões, que certamente enriqueceram esta pesquisa.

Aos meus queridos pais, Vânia de Oliveira Cavalcanti Azevedo e Werbeth Menezes de Azevedo, minha eterna gratidão por todo o sacrifício que fizeram para que eu me tornasse quem sou hoje e por sempre me incentivarem a estudar. Agradeço pelo apoio incondicional, pelos conselhos, pela força e, acima de tudo, pela amizade inigualável que sempre me fortaleceu ao longo desta jornada.

Aos amigos que tive a felicidade de conhecer durante este mestrado e àqueles que sempre estiveram ao meu lado, em especial Uhilma, Adonias Caetano e Rosy, meus sinceros agradecimentos pela amizade, apoio e incentivo ao longo dessa jornada.

Aos colegas da UFMA e do LSDI, que sempre me apoiaram e contribuíram de maneira inestimável para o desenvolvimento deste trabalho, sou profundamente grato pela colaboração e pelo companheirismo.

"A persistência é o caminho do êxito." Charles Chaplin.

Resumo

O suicídio continua sendo um grave problema de saúde pública em todo o mundo, e a identificação precoce da ideação suicida é fundamental para a prevenção de casos fatais. Neste contexto, o presente trabalho teve como objetivo aprimorar o sistema Boamente, uma ferramenta baseada em inteligência artificial desenvolvida para apoiar profissionais de saúde mental na detecção e monitoramento da ideação suicida em textos não clínicos escritos em português brasileiro. O Boamente opera por meio da coleta de textos dos usuários via um teclado virtual em dispositivos móveis, que os envia a uma plataforma web para análise automática. Para tornar o sistema mais preciso e confiável, foram conduzidos dois estudos complementares. No primeiro estudo, buscou-se melhorar o desempenho do modelo de classificação do Boamente por meio da aplicação de técnicas de aprendizado semi-supervisionado. Novos dados foram coletados e diferentes abordagens foram testadas, resultando em uma evolução da acurácia entre 2,39% e 4,30% em relação ao modelo base, com destaque para o método *self-learning*, que apresentou o melhor desempenho geral. No segundo estudo, propôs-se a expansão da arquitetura do Boamente com a incorporação de um módulo gerador de explicações, utilizando engenharia de prompt com *Large Language Models*. A proposta teve como foco aumentar a interpretabilidade do sistema, permitindo que os profissionais de saúde mental compreendam as razões pelas quais um texto foi classificado como contendo ou não ideação suicida. Para isso, foi realizada uma avaliação quantitativa com diferentes LLMs, na qual o modelo Qwen 2.5 (14B) obteve o melhor desempenho em AUC (0,9898), enquanto os modelos Qwen 2.5 de 3B e 7B apresentaram os melhores valores de recall, métrica crítica para evitar falsos negativos em contextos clínicos. Em paralelo, conduziu-se uma avaliação qualitativa com especialistas de diferentes áreas (ciência da computação, linguística e psicologia), que consideraram as explicações geradas pelo modelo LLaMA 3.1 (8B) como as mais coerentes e claras. Os resultados obtidos demonstram que a integração de métodos de aprendizado semi-supervisionado com LLMs explicativos pode elevar significativamente a qualidade do Boamente, tanto em termos de desempenho técnico quanto de confiabilidade percebida pelos profissionais que o utilizam. Além disso, a geração de explicações tem o potencial de fortalecer a confiança dos usuários no sistema de inteligência artificial, prevenindo uma aceitação cega das predições e promovendo uma tomada de decisão mais crítica e embasada. Este trabalho contribui, portanto, para o desenvolvimento de sistemas de mecanismos de inteligência artificial mais transparente e eficientes às demandas da saúde mental, especialmente no contexto da prevenção ao suicídio no Brasil.

Palavras-chave: Ideação suicida; Aprendizado semi-supervisionado; Modelos de linguagem generativos; Engenharia de prompt; Saúde mental; Explicabilidade de modelos; Processamento de linguagem natural.

Abstract

Suicide remains a serious public health issue worldwide, and the early identification of suicidal ideation is essential for preventing fatal outcomes. In this context, the present work aimed to enhance the Boamente system, an artificial intelligence–based tool developed to support mental health professionals in detecting and monitoring suicidal ideation in non-clinical texts written in Brazilian Portuguese. Boamente operates by collecting user texts via a virtual keyboard on mobile devices, which are then sent to a web platform for automatic analysis. To improve the system's accuracy and reliability, two complementary studies were conducted. The first study aimed to improve Boamente's classification model performance through the application of semi-supervised learning techniques. New data were collected, and different approaches were tested, resulting in an accuracy improvement ranging from 2.39% to 4.30% compared to the baseline model. Notably, the self-learning method achieved the best overall performance. These results highlight the potential of combining labeled and unlabeled data when training models for sensitive tasks such as suicidal ideation detection. The second study proposed expanding Boamente's architecture by incorporating an explanation generation module, using prompt engineering with Large Language Models (LLMs). This addition aimed to increase the system's interpretability, enabling mental health professionals to understand the reasons why a text was classified as containing or not containing suicidal ideation. For this purpose, a quantitative evaluation was conducted using various LLMs, in which the Qwen 2.5 model (14B) achieved the best AUC score (0.9898), while the 3B and 7B versions of Qwen 2.5 achieved the highest recall values a critical metric for minimizing false negatives in clinical contexts. In parallel, a qualitative assessment was conducted with experts from different fields (computer science, linguistics, and psychology), who rated the explanations generated by the LLaMA 3.1 (8B) model as the most coherent and clear. The results demonstrate that integrating semi-supervised learning methods with explanatory LLMs can significantly improve the quality of Boamente, both in terms of technical performance and perceived trustworthiness among professionals. Furthermore, explanation generation has the potential to strengthen user trust in the AI system, preventing blind acceptance of predictions and promoting more critical and informed decision-making. This work, therefore, contributes to the development of more transparent and effective artificial intelligence systems to meet the demands of mental health care, particularly in the context of suicide prevention in Brazil.

Keywords: Suicide ideation; Semi-supervised learning; Generative language models; Prompt engineering; Mental health; Model explainability; Natural language processing.

Lista de ilustrações

Lista de tabelas

Lista de Siglas

AET Análise de Emoção em Texto.

AM Aprendizado de Máquina.

AmI Ambient Intelligence.

AP Aprendizado Profundo.

API *Application Programming Interface*.

ASSL Aprendizado Semi-Supervisionado.

AUC *Area Under the Curve*.

BERT *Bidirectional Encoder Representations from Transformers*.

BiLSTM *Bidirectional Long Short-Term Memory*.

BoW *Bag of Words*.

BrWaC Web brasileira como Corpus.

CNN *Convolutional neural network*.

DBLS *Deep Bayesian Label Smoothing*.

DPMH *Digital Phenotyping of Mental Health*.

FD Fenotipagem Digital.

GPT *Generative Pre-trained Transformer*.

GPT-2 *Generative Pre-trained Transformer - 2*.

GPT-3 *Generative Pre-trained Transformer - 3*.

GPT-4 *Generative Pre-trained Transformer - 4*.

GPU *Graphics Processing Unit*.

IA Inteligência Artificial.

IAG Inteligência Artificial Generativa.

IS Ideação Suicida.

ITM *Instruct Models*.

LLaMA *Large Language Model Meta AI*.

LLMs *Large Language Models*.

OMS Organização Mundial da Saúde.

PEFT *Parameter-Efficient Fine-Tuning*.

PLN Processamento de Linguagem Natural.

PSMs Profissionais de Saúde Mental.

PT-BR Português Brasileiro.

QloRA *Quantized Low-Rank Adaptation*.

RLHF *Reinforcement Learning from Human Feedback*.

ROC *Receiver Operating Characteristic*.

SCS *Stratified Confidence Sampling*.

SVM *Support Vector Machine*.

Tf-Idf *Term Frequency-Inverse Document Frequency*.

TM Transtorno Mental.

Sumário

1 Introdução

1.1 Contexto de Pesquisa

A saúde mental diz respeito ao estado de equilíbrio emocional, psicológico e social de um indivíduo, influenciando diretamente a forma como ele pensa, sente, age e se relaciona com os outros. Esse estado de bem-estar pode ser sustentado por práticas saudáveis, como a manutenção de um sono reparador, a adoção de uma alimentação equilibrada e a realização regular de atividades físicas. Além disso, estratégias eficazes de enfrentamento do estresse e de apoio emocional também desempenham um papel fundamental nesse equilíbrio (FU *et al.*, 2020).

No entanto, esse equilíbrio é dinâmico e pode ser profundamente afetado por fatores biopsicossociais, especialmente pela presença de um Transtorno Mental (TM). Quando a saúde mental é comprometida, os impactos tendem a ser amplos e significativos, refletindo-se na qualidade das interações sociais, no desempenho acadêmico e profissional, na tomada de decisões e na capacidade de adaptação às exigências do cotidiano (DORAN; KINCHIN, 2019; FU *et al.*, 2020). A presença de um TM pode restringir a autonomia do indivíduo, dificultar sua integração social e limitar seu potencial de desenvolvimento pessoal. Em muitos casos, o sofrimento psíquico se torna invisível aos olhos da sociedade, sendo negligenciado ou estigmatizado, o que agrava ainda mais sua manifestação e suas consequências.

Além disto, a falta de acompanhamento profissional qualificado, como o oferecido por psicólogos e psiquiatras, eleva de forma significativa o risco de suicídio entre indivíduos diagnosticados com algum TM (BERTOLOTE; MELLO-SANTOS; BOTEGA, 2010; RENAUD *et al.*, 2022). Entre os diversos transtornos, a depressão destaca-se como o mais frequentemente associado a comportamentos suicidas, sendo amplamente identificada em relatos de tentativas e, sobretudo, em casos de suicídio consumado (HANDLEY *et al.*, 2018; WEISS *et al.*, 2022).

O suicídio caracteriza-se como um ato intencional e consciente de pôr fim à própria vida. Trata-se de uma ação deliberada, realizada por um indivíduo que compreende ou presume com elevado grau de certeza, que sua conduta resultará em morte (BERTOLOTE, 2016). Esse fenômeno é resultado de uma interação multifacetada entre fatores genéticos, biológicos, psicológicos, sociais, históricos e culturais (O'CONNOR; KIRTLEY, 2018; CASSORLA, 2021). No entanto, Cassorla (2021) adverte que atribuir "causas" específicas ao suicídio é uma abordagem inadequada, pois os fatores identificados atuam apenas como gatilhos ou facilitadores do comportamento suicida, e não como determinantes diretos.

O comportamento suicida pode se manifestar de diferentes maneiras, abrangendo desde pensamentos sobre a morte até atos concretos que visam a própria eliminação. Suas manifestações incluem a ideação suicida, o planejamento, as tentativas e, por fim, o suicídio consumado (BERTOLOTE, 2016; BOONIAM *et al.*, 2020). A Ideação Suicida (IS) refere-se a pensamentos recorrentes, desejos ou preocupações envolvendo a própria morte, podendo englobar o desenvolvimento de planos, a redação de cartas de despedida e até mesmo manifestações públicas, como postagens em redes sociais (BERTOLOTE, 2016; HARMER *et al.*, 2020).

A IS pode ser classificada em duas categorias: passiva, quando se restringe ao desejo de morrer sem intenção ou plano concreto; e ativa, quando há planejamento e intenção deliberada de realizar o ato suicida (BOONIAM *et al.*, 2020; WASTLER *et al.*, 2023). Estudos apontam que transtornos depressivos e o transtorno hipomaníaco estão associados a ambas as formas de IS (BOONIAM *et al.*, 2020). Ademais, embora a depressão seja considerada um forte preditor da IS, evidências sugerem que ela, isoladamente, não prediz diretamente as tentativas de suicídio, ainda que esteja presente em muitos casos de ambos os comportamentos (MAY; KLONSKY, 2016; RIBEIRO *et al.*, 2018; WEISS *et al.*, 2022).

No contexto da prevenção ao suicídio, tecnologias digitais, como smartphones e dispositivos vestíveis (wearables), têm se mostrado promissoras ao fornecer dados relevantes sobre o ambiente, o comportamento e os estados psicológicos dos indivíduos (MOHR; SHILTON; HOTOPF, 2020). Essas tecnologias possibilitam a aplicação da Fenotipagem Digital (FD) (TOROUS *et al.*, 2016), uma abordagem inovadora que realiza a coleta e o monitoramento contínuo de informações por meio da detecção ativa e/ou passiva de dados comportamentais (MENDES *et al.*, 2022).

A detecção ativa envolve a participação direta do usuário, por meio do preenchimento de questionários digitais, registros de humor, avaliações cognitivas e outras formas de autorrelato estruturado. Por outro lado, a detecção passiva ocorre de maneira automática e não intrusiva, utilizando sensores embarcados em smartphones (como GPS, acelerômetro e padrões de digitação) e dispositivos vestíveis (como monitores de sono e de frequência cardíaca) para captar dados em tempo real sobre hábitos e rotinas do indivíduo.

Com base nessas informações, torna-se possível identificar padrões associados à IS, incluindo alterações no sono, níveis de atividade física, variabilidade da frequência cardíaca, padrões de sociabilidade e conteúdo textual digitado (MENDES *et al.*, 2022; MOURA *et al.*, 2020). Dessa forma, a FD oferece suporte valioso para profissionais de saúde mental, contribuindo para a tomada de decisões clínicas mais assertivas e permitindo a implementação de estratégias terapêuticas personalizadas voltadas à saúde comportamental (MOHR; SHILTON; HOTOPF, 2020; MELCHER; HAYS; TOROUS, 2020).

Nesse contexto, a ferramenta móvel Boamente foi desenvolvida com o objetivo

de auxiliar os Profissionais de Saúde Mental (PSMs) na prevenção do suicídio, por meio da identificação automática de manifestações textuais indicativas de IS em pacientes em acompanhamento psicológico ou psiquiátrico (DINIZ *et al.*, 2022). Trata-se de uma solução baseada nos princípios da FD, voltada para o monitoramento contínuo da IS por meio de dispositivos móveis, como smartphones e tablets.

O sistema opera de forma passiva, capturando os textos digitados pelo usuário a partir de um teclado virtual personalizado. Essas informações são então enviadas para uma plataforma web, onde são processadas e classificadas quanto à presença ou ausência de indícios de IS. Sentenças avaliadas como negativas são imediatamente descartadas, enquanto aquelas com possíveis indícios suicidas são encaminhadas a um mecanismo de inferência, que utiliza um modelo de Inteligência Artificial (IA) para realizar a classificação detalhada.

Importante destacar que, após a inferência, os textos analisados são descartados definitivamente, garantindo que nenhum conteúdo textual seja armazenado de forma permanente. Os PSMs têm acesso exclusivamente aos resultados das classificações por meio de dashboards interativos disponíveis na plataforma, o que permite o acompanhamento contínuo e em tempo real do estado emocional dos pacientes (DINIZ *et al.*, 2022).

1.2 Caracterização do Problema

A identificação e a prevenção eficazes da IS e das tentativas de suicídio ainda constituem um desafio substancial, em razão da persistente estigmatização social e da complexidade multifatorial dos riscos envolvidos. Um dos principais entraves é a ausência de marcadores clínicos ou psicológicos validados que possibilitem a detecção precoce e confiável da propensão ao comportamento suicida, o que limita as abordagens exclusivamente médicas e evidencia a necessidade de estratégias multidisciplinares (INSEL, 2014; RENAUD *et al.*, 2022; BARUA *et al.*, 2024).

O estigma relacionado ao suicídio refere-se à incongruência entre a identidade genuína do indivíduo e os rótulos socialmente atribuídos, frequentemente pejorativos (DEFLEUR, 1964; LEãO; LUSSI, 2021). Indivíduos que expressam pensamentos ou comportamentos suicidas são, muitas vezes, injustamente classificados como pessoas “em busca de atenção” ou consideradas “malucas”, o que contribui para o reforço de estereótipos negativos e práticas discriminatórias. Essa estigmatização representa uma barreira crítica à prevenção, pois tende a inibir a busca por ajuda profissional e o apoio social necessário (LEãO; LUSSI, 2021; ZOU; TANG; BIE, 2022).

Diversas abordagens têm sido investigadas com o objetivo de aprimorar a prevenção do suicídio e a detecção da IS em pacientes (JI *et al.*, 2021; BRACISZEWSKI, 2021). Tradicionalmente, a avaliação do risco suicida é realizada por meio de instrumentos psico-

lógicos e clínicos, como entrevistas presenciais, autorrelatos e questionários padronizados. Esses métodos dependem da interação direta entre o paciente e o profissional de saúde mental, exigindo confiança mútua e comunicação clara.

Além das avaliações psicométricas, diversas estratégias clínicas têm sido empregadas como apoio na identificação de fatores de risco associados à ideação e comportamento suicida. Entre elas, destacam-se análises da frequência cardíaca em repouso, exames de neuroimagem como a ressonância magnética funcional que pode revelar padrões de ativação neural relacionados a estímulos sobre vida e morte, além da análise do tom de voz (SCHERER; PESTIAN; MORENCY, 2013), da variabilidade da frequência cardíaca (SIKANDER *et al.*, 2016) e de investigações psicológicas voltadas à compreensão de aspectos comportamentais e cognitivos associados ao risco de suicídio (CASSORLA, 2021; BERTOLOTE, 2016; O'CONNOR; NOCK, 2014).

Apesar de sua relevância, as abordagens tradicionais para a identificação e prevenção do suicídio apresentam limitações importantes, como baixa capacidade de generalização, viés nas avaliações e elevados custos operacionais (BERNERT *et al.*, 2020; BARUA *et al.*, 2024). A limitação na generalização diz respeito à dificuldade de adaptação desses métodos a diferentes contextos socioculturais, especialmente em populações marginalizadas, áreas rurais ou comunidades com condições socioeconômicas adversas. Instrumentos desenvolvidos e validados em ambientes urbanos, por exemplo, podem não capturar adequadamente os fatores de risco em regiões com acesso restrito a serviços de saúde, onde a expressão do sofrimento psíquico pode assumir formas distintas, como a somatização de sintomas depressivos.

Adicionalmente, essas práticas exigem profissionais altamente capacitados, supervisão constante e infraestrutura adequada, o que compromete sua escalabilidade para grandes populações (BRACISZEWSKI, 2021). Em locais com baixa cobertura de serviços especializados, como áreas remotas, a prevenção do suicídio demanda estratégias colaborativas que envolvam tanto profissionais da atenção básica quanto lideranças comunitárias, capazes de reconhecer e interpretar sinais de sofrimento emocional dentro de contextos culturais específicos (HANDLEY *et al.*, 2018).

Nesse contexto, o Boamente se destaca como uma ferramenta promissora para auxiliar os PSMs no monitoramento remoto e na identificação de sinais de IS. No entanto, sua efetiva adoção por PSMs está condicionada a diversos fatores, especialmente à transparência e compreensão das predições realizadas pelo sistema (PRADHAN; VARSHNEY; SINGH, 2023; ADADI; BERRADA, 2020; KUMAR *et al.*, 2023).

A incorporação de *Large Language Models* (LLMs) representa uma alternativa promissora para apoiar os PSMs no enfrentamento dos desafios relacionados à detecção de IS. Esses modelos têm demonstrado elevado potencial na área da saúde mental e na identificação de risco suicida (HE *et al.*, 2023; QI *et al.*, 2023; LEVKOVICH; ELYOSEPH,

2023). Os LLMs são arquiteturas baseadas em redes neurais do tipo *Transformer*, capazes de processar grandes volumes de dados textuais e identificar padrões linguísticos complexos. Essa capacidade não apenas potencializa a acurácia das predições, mas também permite a geração de explicações mais ricas e compreensíveis, contribuindo para o aumento da transparência e da confiança nos sistemas baseados em IA (ZHAO *et al.*, 2023; ZHAO *et al.*, 2024; KALYAN, 2024).

Apesar do potencial dos LLMs, sua aplicação na saúde pública ainda levanta preocupações importantes. Estudos indicam que é necessário aprofundar a validação empírica desses modelos para assegurar que sua utilização resulte, de fato, em benefícios clínicos (ARENDT *et al.*, 2023). Algumas avaliações apontam que os LLMs podem subestimar o risco de tentativas de suicídio em comparação com a análise realizada por profissionais da saúde (ELYOSEPH; LEVKOVICH, 2023). Além disso, a adoção desses sistemas enfrenta desafios técnicos e operacionais, como a dependência de engenharia de prompt, ou seja, a formulação cuidadosa de instruções, das quais os modelos dependem para gerar respostas relevantes e seguras (CHENG *et al.*, 2023). Outro ponto crítico refere-se à possibilidade de os LLMs produzirem respostas que, embora bem estruturadas linguisticamente, podem conter imprecisões ou informações incorretas, uma vez que esses modelos priorizam a coerência textual em detrimento da veracidade dos conteúdos.

1.3 Relevância do Trabalho

Indivíduos com IS costumam expressar seus sentimentos por meio de publicações em redes sociais (KEMP, 2020; HAQUE *et al.*, 2022; ALDHYANI *et al.*, 2022). Além disso, pessoas com TM frequentemente compartilham seus estados emocionais ou discutem questões relacionadas à saúde mental por meio de dispositivos móveis, utilizando plataformas digitais para publicar textos, fotos, vídeos (KEMP, 2020; ZHANG *et al.*, 2022). Nesse contexto, a identificação precoce de sinais de alerta ou fatores de risco, a partir da análise das mensagens digitadas em dispositivos móveis, pode representar uma estratégia complementar eficaz para apoiar psicólogos e psiquiatras na prevenção do suicídio (ALDHYANI *et al.*, 2022).

A prevenção do suicídio é possível, sobretudo entre pessoas que expressam sentimentos intensos de desesperança, impotência, inutilidade ou percepções de aprisionamento (TURECKI; BRENT, 2016; WEISS *et al.*, 2022). A detecção precoce de sinais de sofrimento psíquico, como a IS, tem se mostrado eficaz na redução de comportamentos autodestrutivos. Ao permitir intervenções oportunas, sejam elas clínicas, psicossociais ou de suporte, essa abordagem pode evitar o agravamento do quadro emocional e reduzir o risco de progressão para planejamento ou tentativa de suicídio (NUTTING *et al.*, 2005; HARKAVY-FRIEDMAN, 2006; DUGAS *et al.*, 2012).

Como discutido na Seção 1.2, embora as abordagens tradicionais sejam eficazes na detecção da IS, elas apresentam limitações relevantes, como a exigência de interação presencial e o risco de viés nas avaliações. Esse risco está relacionado, em parte, à dependência de autorrelatos, que podem ser afetados por vieses idiossincráticos, isto é, distorções individuais na forma como uma pessoa percebe e relata seus próprios sentimentos e comportamentos, frequentemente influenciadas por seu estado emocional momentâneo. Em contraste, métodos baseados em entrevistas estruturadas tendem a ser menos suscetíveis a esse tipo de viés, pois se baseiam em indicadores concretos para classificar eventos (LIU; MILLER, 2014).

Adicionalmente, alguns instrumentos clínicos não são viáveis em contextos como departamentos de emergência, onde o tempo e os recursos são limitados. Muitas dessas abordagens também carecem de validade clínica robusta, o que pode resultar em um elevado número de falsos positivos e falsos negativos (CZYZ, 2015). Nesse cenário, estratégias digitais de prevenção ao suicídio, como o monitoramento remoto de pacientes via sistema Boamente, surgem como soluções complementares promissoras. Elas têm o potencial de ampliar o alcance dos cuidados em saúde mental, melhorar o acesso a intervenções convencionais e fortalecer os sistemas já existentes, ao mesmo tempo em que simplificam processos clínicos (BRACISZEWSKI, 2021).

Diante do exposto, torna-se essencial o desenvolvimento de tecnologias que auxiliem os PSMs no acompanhamento remoto de seus pacientes, incluindo a identificação de casos críticos, como aqueles com alto risco de suicídio. Ferramentas baseadas em IA têm demonstrado resultados promissores nesse contexto, oferecendo alternativas capazes de superar algumas das limitações dos métodos clínicos tradicionais (RABANI; KHAN; KHANDAY, 2020; BARUA *et al.*, 2024). Sistemas de IA aplicados à saúde digital apresentam o potencial de apoiar os profissionais na tomada de decisões mais embasadas, automatizar tarefas rotineiras e, consequentemente, contribuir para a melhoria dos desfechos clínicos dos pacientes (PRADHAN; VARSHNEY; SINGH, 2023).

Com uso de soluções de FD, PSMs podem receber avaliações em tempo real do risco de recaída com base nas alterações dos dados físicos, sociais e digitais capturados, alertando a equipe clínica que poderá planejar e aplicar intervenções até mesmo digitais (VAIDYAM; HALAMKA; TOROUS, 2019). O aprimoramento do Boamente, uma solução de *Digital Phenotyping of Mental Health* (DPMH) integrada com técnicas avançadas de IA, poderá apoiar os profissionais de saúde no cuidado dos pacientes, nas suas decisões clínicas ou terapêuticas, e até mesmo monitorar remotamente o nível de risco de suicídio dos seus pacientes (SAWHNEY *et al.*, 2021). Portanto, através do Boamente, os PSMs podem monitorar o estado psicológico dos seus pacientes em tempo real via análise de emoções relacionados à IS em textos, permitindo apoiar eficientemente as suas decisões clínicas e aplicar intervenções de saúde mais adequadas (TELES *et al.*, 2017; DINIZ *et al.*,

2022).

O uso de soluções baseadas em FD permite que os PSMs recebam avaliações em tempo real sobre o risco de recaída, com base em alterações detectadas em dados físicos, sociais e digitais capturados continuamente. Essas informações possibilitam o envio de alertas à equipe clínica, permitindo o planejamento e a implementação de intervenções oportunas, inclusive por meios digitais (VAIDYAM; HALAMKA; TOROUS, 2019). Nesse contexto, o aprimoramento do Boamente, uma solução de DPMH integrada a técnicas avançadas de IA pode oferecer suporte valioso aos profissionais de saúde, tanto no acompanhamento remoto quanto na tomada de decisões clínicas e terapêuticas, inclusive no monitoramento contínuo do risco de suicídio (SAWHNEY *et al.*, 2021). Por meio da análise de emoções e conteúdos textuais relacionados à IS, o Boamente permite que os PSMs monitorem em tempo real o estado psicológico de seus pacientes, oferecendo apoio efetivo às decisões clínicas e viabilizando intervenções de saúde mental mais precisas e personalizadas (TELES *et al.*, 2017; DINIZ *et al.*, 2022).

Por fim, o Boamente é um sistema aplicado à saúde, no qual o profissional de saúde o utiliza como critério adicional na tomada de decisões relacionadas com a vida. Então é importante que ele seja compreensível em relação a como e por quais razões a IA do Boamente identificou ou não IS nas sentenças recebidas (ADADI; BERRADA, 2018; ADADI; BERRADA, 2020).

1.4 Hipótese de Pesquisa

Essa dissertação de mestrado possui duas hipóteses de pesquisa:

1. O uso de Aprendizado Semi-Supervisionado (ASSL) pode aprimorar o desempenho do modelo Boamente na identificação de ideação suicida em textos escritos em Português Brasileiro (PT-BR);

2. Os LLMs podem ser utilizados não apenas para classificar textos quanto à presença de IS, mas também para gerar explicações interpretáveis que auxiliem os PSMs na compreensão dos resultados das predições, promovendo maior transparência na utilização do Boamente no contexto clínico.

1.5 Objetivos

1.5.1 Geral

Este trabalho tem como objetivo explorar métodos de ASSL e também propor uma nova arquitetura para o sistema Boamente, visando melhorar sua capacidade de identificar

textos com IS. Além disso, busca-se utilizar LLMs para realizar a classificação dos textos e gerar explicações sobre os motivos pelos quais um determinado texto apresenta ou não sinais de IS. Dessa forma, pretende-se contribuir para uma maior transparência do sistema e para o aumento da confiança dos PSMs no uso da ferramenta.

1.5.2 Específicos

Os objetivos específicos deste trabalho de mestrado são:

- Investigar o impacto de diferentes abordagens de ASSL na melhoria da identificação de IS no sistema Boamente;
- Avaliar o desempenho dos métodos semi-supervisionados testados, considerando métricas relevantes para o contexto clínico;
- Realizar o *fine-tuning* de LLMs de propósito geral, com o objetivo de adaptá-los ao domínio da saúde mental e ao contexto linguístico do português;
- Avaliar o desempenho dos LLMs na tarefa de classificação de IS;
- Desenvolver e aplicar engenharia de prompt, utilizando a abordagem *few-shot*, para a geração de explicações interpretáveis de IS;
- Submeter as explicações geradas à avaliação de especialistas das áreas de Computação, Letras e Psicologia, a fim de verificar sua clareza, coerência e utilidade no apoio à tomada de decisão clínica.

1.6 Organização do Trabalho

Este texto de dissertação de mestrado está estruturado em seis capítulos, os quais estão organizados da seguinte maneira:

- O **Capítulo 2** apresenta os fundamentos teóricos que sustentam este trabalho. São abordadas a importância da detecção precoce de sinais de risco de suicídio em ambientes digitais, introduzindo o conceito de FD e técnicas de análise de emoções em textos. Também apresenta a ferramenta Boamente e discute os fundamentos do ASSL, dos LLMs e da geração automática de explicações (Inteligência Artificial Generativa (IAG)) no contexto desta pesquisa.
- O **Capítulo 3** revisa trabalhos relacionados à proposta desta pesquisa, incluindo o uso de LLMs na saúde, a aplicação de métodos de ASSL para melhorar modelos preditivos e estudos que investigam o impacto linguístico nos resultados gerados por modelos de Aprendizado de Máquina (AM) utilizados pelo Boamente.

- O **Capítulo 4** apresenta o primeiro estudo realizado, que investiga a aplicação de métodos de ASSL para a detecção de IS em textos não clínicos escritos em PT-BR. O estudo explora diferentes abordagens de ASSL com base no modelo *BERTimbau*;
- O **Capítulo 5** apresenta o segundo estudo, que investiga o uso de LLMs e técnicas de engenharia de prompt em LLMs para a classificação e explicação de textos com IS em PT-BR;
- O **Capítulo 6** apresenta as considerações finais deste trabalho, destacando as principais contribuições para as áreas de Ciência da Computação e IA. São discutidos os impactos e avanços proporcionados pelo estudo realizado ao longo do mestrado, assim como suas limitações. O capítulo também propõe direções para pesquisas futuras que possam dar continuidade ou ampliar os resultados obtidos. Por fim, são listadas as publicações científicas decorrentes desta dissertação de mestrado.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos fundamentais relacionados à saúde mental e ao comportamento suicida, com o objetivo de contextualizar o leitor e facilitar a compreensão da problemática abordada neste estudo de mestrado. Em seguida, é realizada uma breve introdução à análise de emoções em textos, destacando seus principais aspectos e distinguindo-a da análise de sentimentos. Na sequência, descreve-se a ferramenta Boamente, abordando sua proposta, funcionalidades e arquitetura. O capítulo também explora os métodos de aprendizado semi-supervisionado, discutindo suas principais abordagens, diferenças e aqueles selecionados para aplicação neste trabalho. Por fim, são apresentados os Grandes Modelos de Linguagem Generativos, explicando-se o que são, como funcionam e suas aplicações no contexto desta pesquisa.

2.1 Saúde Mental e Suicidio

De acordo com a Organização Mundial da Saúde (OMS) (OMS, 2021), a saúde mental é um direito básico de todo indivíduo e um fator indispensável para o bem-estar e a qualidade de vida do individuo. Ela abrange uma ampla gama de experiências, que vão desde um estado de equilíbrio e plenitude até condições de intenso sofrimento emocional. Além disso, exerce uma influência significativa na habilidade do ser humano de interagir, desempenhar suas funções e alcançar seu potencial (RIERA-SERRA *et al.*, 2024).

O suicídio, definido como o ato intencional de tirar a própria vida, resulta de uma complexa interação de fatores que se desenvolvem ao longo da existência do indivíduo de maneiras únicas e imprevisíveis (CASSORLA, 2021). Ele constitui um grave problema de saúde pública global, com causas multifatoriais que incluem desde aspectos macroeconômicos e influência midiática até componentes biológicos, psicológicos (especialmente relacionados a vivências precoces), genéticos, ambientais e socioculturais (TURECKI; BRENT, 2016; BALDACARA *et al.*, 2022).

O processo suicida geralmente segue uma progressão que se inicia com pensamentos sobre morte e autodestruição IS, que podem ser vagos e ocasionais. Esses pensamentos tendem a se tornar mais consistentes e recorrentes, evoluindo para a elaboração de planos concretos (plano suicida) até culminar no ato em si, que pode ser fatal (suicídio) ou não (tentativa de suicídio) (BERTOLOTE, 2016). Essa fase final é denominada comportamento suicida, englobando desde manifestações menos graves até atos letais (BOTEGA *et al.*, 2005).

O Brasil está entre os dez países com os maiores números absolutos de suicídios.

Entre adolescentes e jovens adultos, o suicídio figura como uma das principais causas de mortalidade (BOTEGA, 2014), especialmente por se tratar de uma fase crítica para o desenvolvimento de competências emocionais e sociais essenciais (World Health Organization, 2022). Nesse período da vida, o comportamento suicida tende a ser fortemente influenciado por fatores como adversidades psicossociais, conflitos familiares, traços de personalidade e dificuldades cognitivas (BREZO; PARIS; TURECKI, 2006; ORRI *et al.*, 2019). Além disso, o uso precoce de substâncias psicoativas agrava ainda mais os riscos à saúde mental dessa população.

A IS compreende um espectro de pensamentos que variam desde desejos passivos de morte até planos ativos de autoextermínio, incluindo a elaboração de cartas ou mensagens de despedida (HARMER *et al.*, 2020; BOONIAM *et al.*, 2020). Esta condição pode manifestar-se de forma passiva, caracterizada por um mero desejo de deixar de existir, ou ativa, quando acompanhada de intenção e planejamento concretos (NEELEMAN; de Graaf; VOLLEBERGH, 2004).

Historicamente, as notas suicidas eram o principal meio de expressão dos sentimentos associados à ideação suicida, tendo sido amplamente analisadas sob enfoques linguísticos e psicológicos (ALADAĞ *et al.*, 2018). No entanto, com a ascensão das plataformas digitais, como Facebook (SHOIB *et al.*, 2022), Instagram (PICARDO *et al.*, 2020), Twitter e Reddit (MONSELISE; YANG, 2022), esse cenário passou por transformações significativas. Atualmente, indivíduos com tendências suicidas frequentemente expressam sentimentos de desesperança e pensamentos autodestrutivos por meio dessas redes sociais (WONGKOBLAP; VADILLO; CURCIN, 2017).

Essas manifestações digitais, embora muitas vezes menos explícitas do que aquelas observadas nas fases avançadas do processo suicida, têm despertado um interesse científico crescente. Diversos estudos têm se dedicado ao desenvolvimento de métodos eficazes para identificar padrões de IS nas redes sociais, com o objetivo de possibilitar intervenções precoces (JI *et al.*, 2021). Tradicionalmente, a detecção de IS baseia-se em abordagens clínicas e psicológicas, como entrevistas estruturadas e questionários padronizados (GELDER; LOPEZ-IBOR; ANDREASEN, 2003). No entanto, tais métodos apresentam limitações importantes, incluindo baixa escalabilidade, suscetibilidade a vieses e altos custos operacionais (BERNERT *et al.*, 2020).

2.2 Fenotipagem Digital de Saúde Mental

De acordo com (BUFANO *et al.*, 2023), FD refere-se ao monitoramento contínuo e em tempo real de comportamentos, estados psicológicos e fisiológicos de indivíduos por meio de dispositivos digitais, como smartphones, relógios inteligentes (wearables) e sensores. Essa abordagem tem como objetivo avaliar e gerenciar condições de saúde mental,

utilizando tecnologias que capturam dados do cotidiano dos usuários.

Os dados coletados pela FD são classificados, em geral, como ativos e passivos. Os dados ativos são obtidos diretamente por meio da interação do usuário, como respostas a questionários, autorrelatos ou uso voluntário de aplicativos de saúde mental. Já os dados passivos são capturados automaticamente, sem necessidade de intervenção direta do usuário, e incluem variáveis como padrões de sono e locomoção, frequência de chamadas, tempo de uso de aplicativos, dinâmica de digitação, voz e linguagem usada em mensagens de texto ou postagens em redes sociais. Essa coleta constante e discreta permite monitorar alterações sutis no comportamento que podem estar associadas ao surgimento ou agravamento de condições de saúde mental (INSEL, 2017).

Ambientes inteligentes projetados para compreender as características e necessidades de um indivíduo, com o objetivo de oferecer suporte de forma discreta, personalizada e eficaz no cotidiano, descrevem o conceito de Ambient Intelligence (AmI), do inglês *Ambient Intelligence*). Essa abordagem integra tecnologias que permitem ao ambiente adaptar-se dinamicamente ao comportamento dos usuários, promovendo maior conforto, segurança e bem-estar (SADRI, 2011).

Entre as tecnologias fundamentais para a implementação da AmI, destaca-se a computação pervasiva, fortemente relacionada à computação ubíqua no que diz respeito à mobilidade e à integração invisível de dispositivos tecnológicos ao cotidiano. A computação pervasiva pode ser entendida como o esforço para embutir e interconectar sistemas computacionais em objetos comuns do dia a dia, como smartphones, smartwatches, veículos, móveis e até mesmo roupas. Tais sistemas distinguem-se por operarem em rede e por serem projetados para desempenhar exclusivamente as funções específicas às quais se destinam, de forma autônoma e contextual (BIMPAS *et al.*, 2024).

Aplicações baseadas em AmI, por meio do uso de dispositivos ubíquos como smartphones e wearables, têm viabilizado o monitoramento passivo e contínuo de comportamentos humanos em tempo real. Essa abordagem tem ganhado destaque no desenvolvimento de ferramentas de apoio à tomada de decisão clínica, especialmente no contexto da psicoterapia, ao possibilitar o rastreamento de estados emocionais e distúrbios relacionados à saúde mental (MOURA *et al.*, 2022; MENDES *et al.*, 2022). Nesse cenário, tal aplicação constitui-se como um dos principais pilares da FD, ao integrar dados comportamentais e fisiológicos com o objetivo de promover intervenções mais precisas e personalizadas.

O principal objetivo da FD é construir um retrato dinâmico, contínuo e contextualizado do bem-estar mental de um indivíduo, por meio da análise integrada de dados coletados digitalmente. Essa abordagem visa possibilitar a detecção precoce de transtornos mentais, como depressão, ansiedade, transtorno bipolar e IS, contribuindo para intervenções mais precisas e oportunas (JAIN *et al.*, 2015; TOROUS *et al.*, 2016; D'ALFONSO, 2020; MOURA *et al.*, 2022).

2.3 Análise de Emoção em Textos

A Análise de Emoção em Texto (AET) é uma subárea do Processamento de Linguagem Natural (PLN) que visa identificar, classificar e interpretar emoções específicas (como alegria, tristeza, raiva, medo, surpresa ou nojo, dentre outras) expressas em dados textuais. Diferentemente da análise de sentimento, que se concentra na polaridade (positivo/negativo/neutro), a AET busca capturar nuances emocionais mais complexas e categorizá-las conforme modelos psicológicos estabelecidos (EKMAN, 1992; PLUTCHIK, 1980). Seu objetivo é extrair significados afetivos para aplicações como monitoramento de saúde mental, personalização de serviços e compreensão de comportamentos coletivos. A Figura 1 mostra alguns tipos de classifação de emoções.

Figura 1 – Tipos de classificação de emoções.

Fonte: adaptado de Alanazi *et al.* (2022)

Essa tarefa envolve a extração semântica de sinais emocionais contidos em palavras, frases ou contextos discursivos, e pode ser realizada por meio de abordagens baseadas em regras, dicionários de emoções (como NRC Emotion Lexicon) ou modelos de AM e redes neurais (MOHAMMAD; TURNEY, 2013; FELBO *et al.*, 2017). O objetivo é permitir a compreensão automatizada de estados afetivos em textos produzidos por humanos, como postagens em redes sociais, diários, e-mails, comentários e outros registros digitais.

A AET é amplamente aplicada em áreas como monitoramento de bem-estar psicológico, detecção de transtornos mentais, experiência do usuário, e interação humano-computador, sendo particularmente relevante em contextos sensíveis como a identificação de IS, onde certas emoções recorrentes podem atuar como sinais de alerta precoce (CALVO *et al.*, 2017; LIN *et al.*, 2014).

2.4 A Ferramenta Boamente

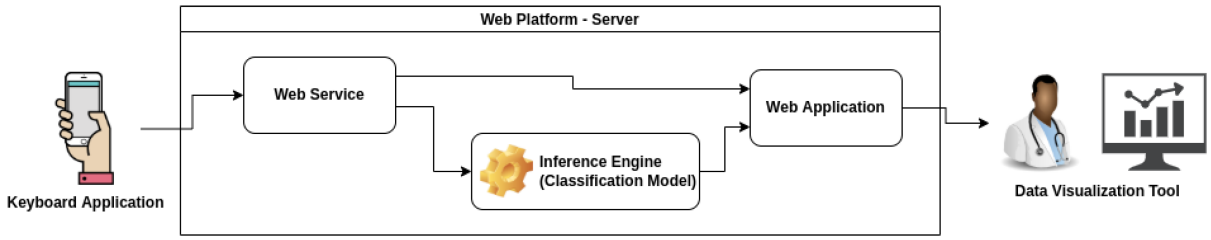
Os métodos de prevenção ao suicídio abrangem um conjunto diversificado de estratégias voltadas para a identificação, o monitoramento e a intervenção em situações de risco. Entre essas estratégias, o monitoramento contínuo de indicadores associados à IS destaca-se como uma ferramenta essencial. Com os avanços tecnológicos e computacionais, emergiu o conceito de *just-in-time intervention* (intervenção no momento certo) (COPPERSMITH *et al.*, 2022), que ganhou relevância por sua abordagem proativa e personalizada. Essa metodologia baseia-se na aplicação de algoritmos e ferramentas analíticas capazes de rastrear, em tempo real, sinais comportamentais e emocionais de indivíduos em situação

de vulnerabilidade. O objetivo é viabilizar intervenções precisas e contextualmente apropriadas no exato momento em que a pessoa enfrenta uma crise, potencializando a eficácia do suporte oferecido e reduzindo o risco de desfechos graves.

O Boamente (DINIZ *et al.*, 2022) é uma ferramenta projetada para enfrentar o desafio de monitorar a IS em indivíduos em situação de risco. A solução combina um aplicativo de teclado virtual para dispositivos móveis, responsável pela coleta passiva de dados textuais durante a digitação dos usuários, com uma plataforma web integrada que centraliza o processamento e análise dessas informações. Os dados capturados são enviados de forma segura para a plataforma, onde são analisados por meio de técnicas de PLN e modelos de (Aprendizado Profundo (AP)), como o *BERTimbau Large*. Utilizando os princípios da FD, o Boamente busca identificar padrões linguísticos e indicadores sutis associados à IS. Essa abordagem permite não apenas a detecção precoce de sinais de risco, mas também viabiliza intervenções oportunas as chamadas *just-in-time interventions* realizadas por profissionais de saúde mental, ampliando as possibilidades de prevenção ao suicídio.

Na Figura 2, temos a primeira arquitetura usada no Boamente, o sistema inicia com um teclado virtual desenvolvido para dispositivos Android, utilizando a linguagem Java. Esse teclado substitui o teclado padrão do dispositivo e é responsável por capturar os textos digitados pelo usuário de forma passiva, sem armazená-los localmente. Para garantir a privacidade do usuário, apenas frases contendo duas ou mais palavras são enviadas para a plataforma web, acompanhadas de um identificador único (UUID) e um registro de data e hora (timestamp). Além disso, o teclado foi projetado sem funções como autocompletar ou acesso ao clipboard, assegurando que os textos analisados sejam exclusivamente originados pelo usuário, sem interferências externas.

Figura 2 – Primeira arquitetura do Boamente.



Fonte: Diniz *et al.* (2022)

Os dados coletados pelo teclado são recebidos por um serviço web construído com o *framework* FastAPI, que utiliza o protocolo SSL para garantir a segurança na transmissão das informações. Nesta etapa, é realizada uma verificação inicial para identificar a presença de termos associados a suicídio, como "suicídio"ou "querer morrer". Caso nenhum termo relevante seja detectado, o texto é imediatamente descartado, preservando a privacidade do usuário. Se termos relacionados forem encontrados, o texto é encaminhado ao mecanismo

de inferência para uma análise mais detalhada.

O núcleo do sistema consiste em um modelo de AP baseado no *BERTimbau Large*, uma versão especializada do *Bidirectional Encoder Representations from Transformers* (BERT) para o PT-BR. O modelo foi treinado utilizando tweets previamente anotados por psicólogos, que passaram por etapas de pré-processamento, incluindo limpeza de URLs, remoção de stopwords e tokenização. Para lidar com o desequilíbrio entre as classes "positiva"(com ideação suicida) e "negativa"(sem ideação suicida), foi aplicada a técnica *SMOTE (Synthetic Minority Over-sampling Technique)*. Os resultados da classificação são exibidos em uma plataforma web desenvolvida com o *framework* Laravel, destinada a PSMs. O acesso é controlado por um sistema baseado em funções (ACL), garantindo que apenas usuários autorizados visualizem os dados. Os dashboards fornecem uma visão abrangente e individualizada, incluindo gráficos de tendências temporais e análises comparativas entre pacientes. Essa ferramenta permite que os profissionais monitorem a evolução dos casos e identifiquem padrões, auxiliando na tomada de decisões clínicas de forma mais informada e oportuna.

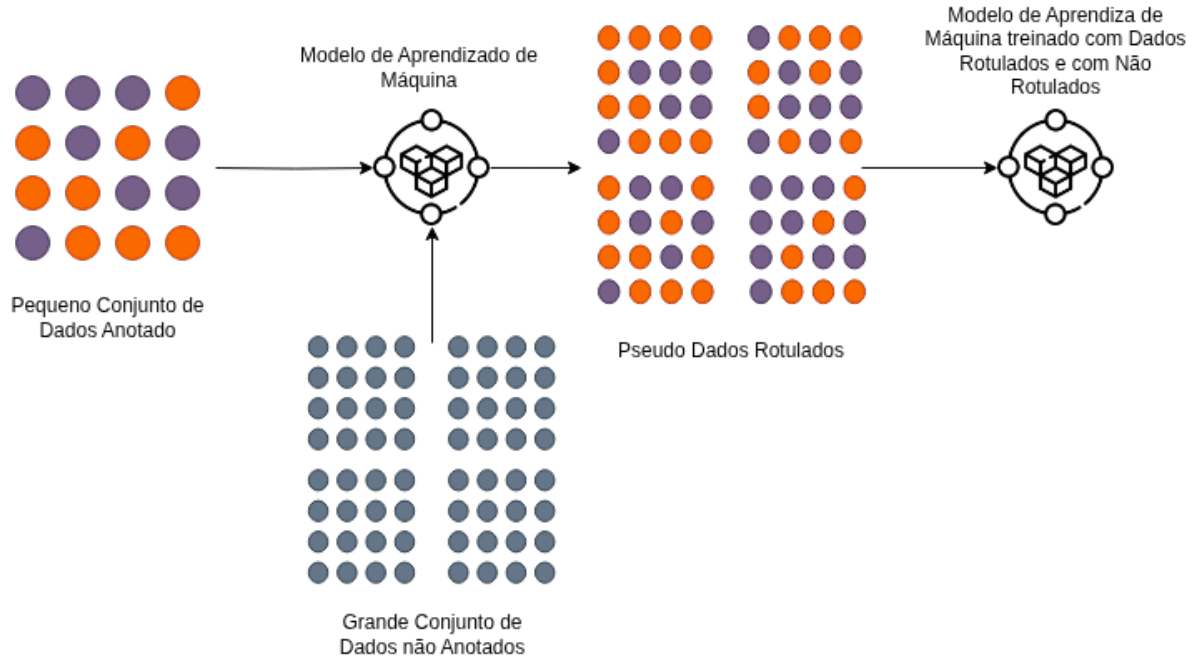
2.5 Aprendizado Semi-supervisionado

O ASSL ocupa uma posição intermediária entre os paradigmas supervisionado e não supervisionado, combinando elementos de ambos. Entre suas principais abordagens estão a classificação semissupervisionada, que utiliza um conjunto misto de dados rotulados e não rotulados para treinar classificadores. Esse tipo de aprendizado busca aprimorar o desempenho em relação às abordagens puramente supervisionadas ou não supervisionadas. É uma estratégia relevante tanto do ponto de vista prático, para o desenvolvimento de algoritmos mais eficazes, quanto do ponto de vista teórico, para compreender os mecanismos de aprendizado em humanos e máquinas. Sua utilidade é particularmente evidente em cenários onde há escassez de dados rotulados, que são geralmente caros e trabalhosos para serem obtidos (CHAPELLE; SCHOLKOPF; ZIEN, 2009; ZHU; GOLDBERG, 2022).

A Figura 3 ilustra a estrutura básica de um dos métodos utilizados neste trabalho, o *self-learning* (AMINI *et al.*, 2023). Esse método segue uma arquitetura conceitual baseada em três princípios fundamentais: a utilização simultânea de dados rotulados e não rotulados, a propagação de informações entre os exemplos, e o refinamento iterativo do modelo a partir das previsões realizadas sobre os dados não rotulados.

Além dos métodos que iremos apresentar a seguir, o ASSL engloba uma variedade de outras abordagens teóricas e aplicadas, como *multi-view learning* (MUSLEA; MINTON; KNOBLOCK, 2002), *transductive learning* (SHI *et al.*, 2018) e *generative models* (KINGMA *et al.*, 2014), cada uma com suas particularidades e domínios de aplicação preferenciais. No entanto, para os propósitos deste trabalho, optou-se por focar nos métodos de *self-learning*,

Figura 3 – Estrutura do método *Self-Learning*.



Fonte: Autor

co-training, *boosting*, *label propagation* e *weighting*. A escolha desses métodos específicos também reflete uma estratégia deliberada de cobrir diferentes princípios de funcionamento dentro do ASSL. Enquanto o *self-learning* e o *co-training* representam abordagens baseadas em iteração e consenso, o *boosting* traz o paradigma de combinação de modelos, e o *label propagation* com *weighting* exemplificam técnicas baseadas em grafos e ponderação. Essa diversidade permite não apenas uma comparação mais abrangente entre diferentes estratégias, mas também favorece a identificação de complementaridades metodológicas, aprofundando a análise realizada e contribuindo para a obtenção de resultados mais consistentes e cientificamente embasados no contexto do problema estudado.

O *self-learning* (AMINI *et al.*, 2023) é uma abordagem clássica de ASSL que explora a capacidade de um modelo supervisionado generalizar a partir de dados não rotulados. Inicialmente, o modelo é treinado com um conjunto de dados rotulados, aprendendo os padrões discriminativos associados às classes de interesse. Em seguida, esse modelo é utilizado para realizar predições sobre dados não rotulados, atribuindo a eles pseudo-rótulos baseados em sua própria confiança nas classificações realizadas. Tipicamente, apenas as amostras com maior probabilidade preditiva, ou seja, aquelas em que o modelo demonstra maior certeza, são selecionadas para compor o novo conjunto rotulado. Esses exemplos pseudo-rotulados são então incorporados ao conjunto original de treinamento, e o modelo é re-treinado com a base expandida. Esse processo pode ser repetido de forma iterativa, promovendo o refinamento progressivo da função de decisão à medida que mais dados não rotulados são integrados com rótulos presumidos. A ideia central é que, a cada iteração, o modelo se torne mais robusto, melhorando sua capacidade de generalização e reduzindo a

dependência de grandes volumes de dados rotulados — que são geralmente caros e difíceis de obter.

Embora o *self-learning* possa envolver múltiplas iterações, ele também pode ser aplicado de maneira não iterativa, com uma única etapa de pseudo-rotulagem, especialmente em cenários com restrições de tempo ou recursos computacionais. No entanto, a eficácia do método depende fortemente da qualidade dos pseudo-rótulos gerados. Caso o modelo inicial tenha baixa acurácia, há risco de propagação de erros durante o processo, o que pode comprometer o desempenho final. Assim, estratégias como limiares de confiança, validação cruzada ou uso de comitês de modelos são frequentemente empregadas para mitigar esse risco.

O método *co-training* (LANG *et al.*, 2022) é uma estratégia de ASSL que se fundamenta na colaboração entre dois (ou mais) modelos de AM, cada um treinado com vistas complementares, ou subconjuntos distintos dos dados disponíveis. A premissa central da abordagem é que, ao explorar diferentes representações da mesma instância (por exemplo, características lexicais e sintáticas em um texto), os modelos podem aprender aspectos distintos mas complementares do problema, resultando em um processo de aprendizagem mais robusto. Inicialmente, ambos os modelos são treinados separadamente utilizando um conjunto limitado de dados rotulados. Em seguida, cada modelo realiza predições sobre o conjunto de dados não rotulados e identifica, com base em critérios de alta confiança, instâncias que acredita ter classificado corretamente. Esses exemplos pseudo-rotulados são então compartilhados entre os modelos: o que foi rotulado com confiança por um modelo é incorporado ao conjunto de treinamento do outro. Esse processo de rotulagem cruzada é repetido iterativamente, permitindo que os modelos aprendam progressivamente com os exemplos rotulados pelo outro, expandindo assim o conjunto de dados rotulados com supervisão mínima.

O método *boosting* (CHEN *et al.*, 2023) é uma técnica de AM amplamente utilizada que visa construir um modelo preditivo robusto a partir da combinação de diversos modelos fracos, ou seja, classificadores que, individualmente, apresentam desempenho ligeiramente superior ao acaso. A principal característica do *boosting* é seu processo iterativo, no qual cada novo modelo é treinado para corrigir os erros cometidos pelos modelos anteriores. A cada iteração, o algoritmo ajusta os pesos atribuídos às instâncias de treinamento, de forma que exemplos mal classificados recebam maior importância na próxima rodada de aprendizado. Essa reponderação contínua permite que o algoritmo concentre seus esforços nos exemplos mais difíceis, promovendo um aprendizado mais eficaz e direcionado. Com o avanço das iterações, os modelos fracos são sequencialmente combinados, geralmente por meio de uma ponderação dos votos ou predições, resultando em um classificador final com desempenho superior. Ao reduzir simultaneamente o viés e a variância do modelo, o *boosting* melhora significativamente a capacidade de generalização da solução aprendida.

O processo prossegue até que um critério de parada seja atingido, como um número máximo de iterações ou a convergência do erro. Entre suas variantes mais conhecidas estão o *AdaBoost*, o *Gradient Boosting* e o *XGBoost*, amplamente utilizados em tarefas de classificação e regressão com dados rotulados limitados.

O método *label propagation* (ISCEN *et al.*, 2019) é uma abordagem de ASSL fundamentada na estrutura de grafos para a inferência de rótulos em dados não rotulados. Inicialmente, constrói-se um grafo de vizinhança, no qual os nós representam as instâncias de dados e as arestas refletem a similaridade entre elas, geralmente definida a partir de representações vetoriais extraídas por redes neurais. A propagação dos rótulos é realizada ao longo das conexões desse grafo, partindo dos dados rotulados para os não rotulados, de modo que os exemplos próximos compartilhem informação supervisionada. A inferência ocorre com base na premissa de que dados similares tendem a possuir rótulos semelhantes. Assim, os rótulos dos exemplos conhecidos são propagados iterativamente para seus vizinhos mais próximos, e essa informação supervisionada se difunde gradualmente pelo grafo. Os rótulos inferidos são, então, incorporados ao conjunto de treinamento para refinar o modelo. Um aspecto crucial dessa abordagem é a consideração da incerteza associada aos rótulos propagados, que pode variar conforme a distância entre os exemplos rotulados e não rotulados ou a ambiguidade presente nas representações. Essa incerteza influencia a confiança das predições e pode ser utilizada para ponderar a contribuição de cada instância no treinamento subsequente, tornando o processo mais robusto em cenários com dados escassamente rotulados.

O método *weighting* (CHEN *et al.*, 2019) representa uma estratégia para aprimorar a classificação semi-supervisionada baseada em grafos por meio da atribuição de pesos às amostras, considerando sua relevância no processo de propagação de rótulos. Essa abordagem utiliza índices de dificuldade de agrupamento, calculados a partir de múltiplos agrupamentos (clusterings), para estimar a confiabilidade de cada amostra no contexto do grafo. Especificamente, amostras rotuladas localizadas próximas às fronteiras de decisão entre diferentes classes são consideradas mais informativas e, portanto, recebem maior peso no processo de inferência. Ao valorizar as instâncias com maior potencial discriminativo, ou seja, aquelas que melhor representam a transição entre classes, o método promove uma propagação de rótulos mais eficaz, especialmente em regiões ambíguas do espaço de características. Isso contribui para uma melhoria substancial na acurácia da classificação, tornando o processo mais robusto em cenários com rótulos escassos e fronteiras complexas entre categorias.

2.6 Grandes Modelos de Linguagem Generativos

Os LLMs são modelos de IA projetados para gerar texto de forma autônoma com base em uma entrada inicial fornecida, como um prompt ou uma sequência de palavras (YANG *et al.*, 2023). Esses modelos são baseados em redes neurais profundas e treinados com vastos volumes de dados textuais para entender padrões, sintaxe, semântica e contextos linguísticos, permitindo-lhes produzir texto coerente e fluido em uma variedade de estilos, tópicos e formatos (LI *et al.*, 2024).

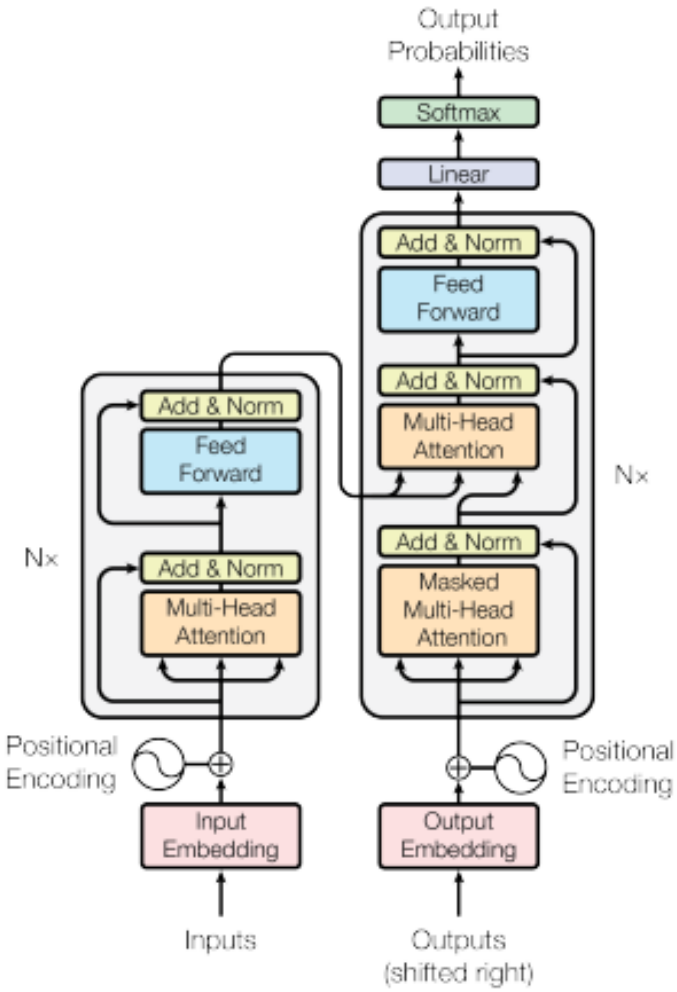
Os LLMs, como o *Generative Pre-trained Transformer* (GPT), utilizam a arquitetura Transformer (VASWANI *et al.*, 2023), que se destaca por sua capacidade de capturar dependências de longo alcance dentro do texto. Durante o treinamento, esses modelos aprendem a prever a probabilidade da próxima palavra em uma sequência, o que lhes permite criar texto original com base no contexto dado. Esses modelos podem ser aplicados em diversas áreas, como chatbots, geração de conteúdo automatizado, tradução automática, sumarização de texto, assistentes virtuais e até na criação de código de programação. Sua característica "generativa" significa que, ao contrário de modelos discriminativos, que são projetados para classificar ou fazer previsões, os LLMs são projetados para criar novas informações baseadas nas entradas fornecidas.

Conforme a Figura 4 podemos observar que a arquitetura é composta por um codificador e um decodificador, cada um contendo múltiplas camadas idênticas com subcamadas de atenção multi-head (que processa a informação em diferentes espaços de representação) e redes neurais *feed-forward*, intercaladas com conexões residuais e normalização de camada. Essa estrutura paralelizável elimina a dependência do processamento sequencial típico de RNNs/LSTMs, permitindo treinamento mais eficiente em grandes volumes de dados e capturando tanto dependências locais quanto globais no texto, o que a torna particularmente adequada para modelos de linguagem em grande escala.

A integração de LLMs na saúde mental tem gerado avanços significativos, com aplicações que abrangem desde intervenções terapêuticas até sistemas de monitoramento contínuo. Um estudo pioneiro conduzido por (HEINZ *et al.*, 2025) demonstrou o potencial dessas tecnologias ao testar o chatbot *Therabot*, uma solução baseada em IAG com diálogos terapêuticos, que apresentou eficácia clínica na redução de sintomas de depressão, ansiedade e risco de transtornos alimentares em um ensaio clínico randomizado com 210 participantes. Além dos resultados clínicos positivos, o estudo evidenciou altos níveis de engajamento, satisfação dos usuários e estabelecimento de uma aliança terapêutica comparável à observada em terapias conduzidas por humanos, indicando que LLMs bem treinados podem representar uma alternativa viável e escalável no suporte à saúde mental.

A evolução da arquitetura *Transformer* permitiu o surgimento de três paradigmas principais nos LLMs, cada um adaptado a demandas específicas do processamento de

Figura 4 – Arquitetura *Transformer*.



Fonte: Vaswani *et al.* (2023)

linguagem natural. O primeiro paradigma, codificadores puros, emprega apenas o módulo codificador do *Transformer*, como no modelo BERT. Esses sistemas são especializados em análise contextual bidirecional, sendo amplamente utilizados em tarefas como classificação de sentimentos e reconhecimento de entidades nomeadas. O segundo paradigma, decodificadores puros, exemplificado por modelos como *Generative Pre-trained Transformer - 3* (GPT-3) e *Large Language Model Meta AI* (LLaMA), utiliza exclusivamente o módulo decodificador, otimizando a geração autoregressiva de texto para aplicações como tradução automática e criação de conteúdo. Por fim, a arquitetura codificador-decodificador, representada pelo *T5 (Text-To-Text Transfer Transformer)*, combina ambos os módulos para mapear sequências de entrada em saída, sendo ideal para tarefas de transformação direta, como resposta a perguntas e reformulação de frases (RAFFEL *et al.*, 2023; YANG *et al.*, 2024).

Essas variações arquiteturais não apenas definem as capacidades funcionais dos modelos, mas também influenciam sua eficiência computacional e adaptabilidade. Enquanto

codificadores puros destacam-se em compreensão contextual, decodificadores puros excelam em geração criativa, e modelos híbridos oferecem versatilidade para problemas complexos que exigem interpretação e síntese simultâneas. Essa segmentação reflete a maturidade do campo, onde a escolha da arquitetura é estratégica e depende diretamente dos requisitos da aplicação final (SHAO *et al.*, 2024).

Contudo, os desafios técnicos e éticos permanecem substanciais. Riscos como alucinações de modelos, onde sistemas geram recomendações não validadas ou informações errôneas, são amplificados em contextos sensíveis como saúde mental (GUO *et al.*, 2024a). Mais recentemente, surgiram modelos de linguagem que combinam diferentes tipos de dados, como texto, imagem, áudio e vídeo, permitindo interações mais completas e eficientes. Modelos como o *Generative Pre-trained Transformer - 4* (GPT-4) e o *Gemini* são exemplos dessa nova geração multimodal, que amplia o uso dos LLMs para uma variedade ainda maior de aplicações (RAO *et al.*, 2025). A pesquisa nessa área continua avançando, buscando tanto melhorar o desempenho técnico desses modelos quanto lidar com questões práticas e éticas que surgem com seu uso em situações do dia a dia (DAS; AMINI; WU, 2025).

3 Trabalhos Relacionados

Neste capítulo, são apresentados os principais trabalhos relacionados à detecção de IS. O primeiro grupo de trabalhos relacionados foca no uso de técnicas de ASSL. O segundo grupo de trabalhos explora a aplicação de LLMs na identificação de riscos de suicídio, destacando o potencial dessas ferramentas para gerar análises mais precisas e interpretáveis. Por fim, apresentamos uma análise comparativa dos estudos apresentados nessa dissertação de mestrado com os trabalhos relacionados.

3.1 Aprendizado Semi-supervisionado no Desenvolvimento de Modelos para Prevenção ao Suicídio

Para este primeiro grupo de trabalhos relacionados, consideramos estudos que utilizaram técnicas de ASSL no processo de desenvolvimento de modelos para detectar IS. No contexto de prevenção ao suicídio, as técnicas de ASSL são utilizadas para enfrentar a escassez de dados rotulados e o desbalanceamento entre as classes.

Squires et al. Squires *et al.* (2024) propuseram uma abordagem inovadora para a detecção de IS em redes sociais utilizando a técnica de *Deep Bayesian Label Smoothing* (DBLS), visando incorporar a incerteza inerente aos rótulos em tarefas de classificação de saúde mental. O estudo teve como objetivo melhorar a acurácia na classificação de postagens de redes sociais quanto ao nível de risco de suicídio, considerando a subjetividade envolvida na rotulagem de dados. Para isso, os autores empregaram uma arquitetura baseada em *Convolutional neural network* (CNN), adaptada para incorporar a suavização de rótulos não uniforme, baseada em inferência bayesiana via *MC Dropout* (dropout ativado na inferência para estimar incerteza com múltiplas previsões). O modelo foi treinado para classificar as postagens no dataset *Reddit C-SSRS* em cinco categorias distintas de risco, conforme a escala *Columbia Suicide Severity Rating Scale*: IS (IS — pensamentos sobre cometer suicídio, com ou sem plano), comportamento suicida (atos preparatórios ou comportamentos que indicam intenção de se suicidar), tentativa de suicídio (ato autoinfligido com alguma intenção de morrer), indicador de suicídio (sinais indiretos como escrever cartas de despedida ou distribuir pertences), e apoio (busca ativa por ajuda ou suporte emocional). Os resultados demonstraram que a aplicação do DBLS melhora significativamente o desempenho do modelo, elevando a acurácia de 43% para 52%.

Tank et al. Tank *et al.* (2024) propuseram o *Su-RoBERTa*, um modelo baseado no *RoBERTa* (LIU *et al.*, 2019) e ajustado para a tarefa de previsão de risco de suicídio

baseada em postagens nas redes sociais, com o objetivo de demonstrar a viabilidade de modelos menores e mais eficientes para essa aplicação. O estudo utilizou uma abordagem semi-supervisionada, combinando 500 postagens rotuladas e 1.500 não rotuladas do dataset do *IEEE BigData 2024 Suicide Risk Detection Challenge*. O modelo foi treinado para classificar postagens em quatro níveis de risco: Indicador, Ideação, Comportamento e Tentativa. A classe "Indicador" refere-se a sinais indiretos de sofrimento emocional que podem sugerir risco suicida, como expressões de desesperança, isolamento, sem menção direta à morte. A classe "Ideação" envolve pensamentos sobre suicídio, com ou sem planos definidos, sendo expressa por frases como "não vejo mais sentido em viver" ou "penso em desaparecer". A classe "Comportamento" abrange ações preparatórias associadas à IS, como procurar métodos, planejar o ato ou escrever cartas de despedida. Por fim, a classe "Tentativa" representa casos em que houve uma tentativa real de suicídio, ou seja, uma ação autoinfligida com alguma intenção de morrer, independentemente do resultado. Para mitigar o desbalanceamento das classes, especialmente a classe minoritária "Tentativa", os autores realizaram aumento de dados utilizando o modelo *Generative Pre-trained Transformer - 2* (GPT-2), gerando amostras sintéticas para equilibrar as categorias. A estratégia envolveu duas iterações de pseudo-rotulagem e re-treinamento do *RoBERTa* com os novos dados, resultando em um modelo computacionalmente eficiente que alcançou um F1 ponderado (ponderada pelo número de amostras de cada classe) de 69,84%, destacando-se como uma solução promissora para aplicações em dispositivos de baixa capacidade computacional.

Caicedo et al. Acuña Caicedo, Gómez Soriano e Melgar Sasieta (2022) propuseram um método semi-supervisionado baseado em *bootstrapping* para a anotação de mensagens com potencial risco de suicídio em redes sociais, visando expandir o *Life Corpus* (um corpus bilíngue para detectar IS) e melhorar a detecção desse tipo de risco. O estudo utilizou uma abordagem de expansão progressiva do corpus, inicialmente treinando um classificador *Support Vector Machine* (SVM) com características de *Bag of Words* (BoW) e *Term Frequency-Inverse Document Frequency* (Tf-Idf), que foi iterativamente realimentado com novos exemplos não rotulados coletados do *subreddit r/SuicideWatch*. A tarefa consistiu na classificação binária de mensagens em categorias de risco e não risco de suicídio, permitindo aumentar o *Life Corpus* de 102 para 273 amostras, com uma concordância entre anotadores (especialistas da área) de 0,86 e atingindo um *macro F1-score* de até 0,80. O trabalho destaca a viabilidade de métodos semi-supervisionados para reduzir o esforço manual na criação de corpus de saúde mental e potencializar sistemas de detecção de IS.

Lovitt et al. Lovitt *et al.* (2024) propuseram um *framework* semi-supervisionado para a avaliação do risco de suicídio em postagens de redes sociais, visando mitigar as limitações decorrentes da escassez e do desbalanceamento de dados rotulados. O estudo empregou o modelo *RoBERTa*, que demonstrou desempenho superior em comparação com outros modelos transformes como BERT e *BiomedBERT* (GU *et al.*, 2020), sendo utilizado

como base para a geração e incorporação de pseudo-rótulos. O *framework* proposto consiste na classificação de postagens extraídas do *subreddit r/SuicideWatch* em quatro níveis de risco de suicídio: Indicador, Ideação, Comportamento e Tentativa, conforme as categorias definidas no desafio *Suicide Detection on Social Media BigData Cup*. Para lidar com o desbalanceamento das classes, os autores desenvolveram o método *Stratified Confidence Sampling* (SCS), que seleciona pseudo-rótulos com base na confiança estratificada por classe. Os resultados mostraram que a utilização de dados pseudo-rotulados, especialmente após validação parcial, elevou significativamente o desempenho do modelo, com destaque para o aumento do *F1-score* para a classe menos representada.

3.2 LLMs Generativos na Prevenção ao Suicídio

Sistemas baseados em LLMs têm demonstrado capacidades notáveis na geração de diferentes modalidades de dados a partir de comandos humanos. Essas habilidades incluem resultados promissores em diversas áreas da saúde, como pesquisa médica, diagnóstico, tratamento, cuidados com pacientes, sumarização de textos médicos e recomendações personalizadas (SAI *et al.*, 2024). No contexto da saúde mental, os LLMs também têm mostrado um potencial significativo (GUO *et al.*, 2024b), especialmente em iniciativas voltadas à prevenção do suicídio (HOLMES *et al.*, 2025).

Alguns estudos investigaram o uso de IS para aprimorar o desempenho de modelos preditivos. Ghanadian, Nejadgholi e Osman (2024), diante da escassez e da sensibilidade dos dados reais sobre suicídio, propuseram uma estratégia de geração de dados sintéticos baseada em fatores sociais extraídos da literatura em psicologia. Métodos de geração de dados sintéticos oferecem uma alternativa econômica para complementar dados reais e auxiliar na detecção de IS em textos. Modelos baseados no BERT, quando treinados com dados sintéticos através do aumento de dados, demonstraram desempenho eficaz na detecção de IS. Além disso, ao combinar 30% do conjunto de dados real com os dados sintéticos, observou-se uma melhoria substancial no desempenho dos modelos.

Outro estudo, conduzido por Qorich e Ouazzani (2024), propôs uma versão combinada dos modelos CNN e *Bidirectional Long Short-Term Memory* (BiLSTM), denominada *C-BiLSTM*, utilizando incorporação tripla de palavras *(triple word embedding)* para detectar IS em postagens do *Reddit*. Os autores também avaliaram modelos baseados em BERT e GPT-2. Além disso, apresentaram uma análise combinatória envolvendo CNN, BiLSTM e *embeddings* pré-treinados como *GloVe*, *Word2Vec* e *FastText*. O modelo *C-BiLSTM* alcançou uma acurácia de 94,5%, enquanto o GPT-2 obteve 97,69%.

Com o avanço dos LLMs, novas abordagens têm explorado o potencial dessas ferramentas para compreender e interpretar manifestações de IS em ambientes digitais. Nesse contexto, Bauer *et al.* (2024) utilizaram LLMs para compreender o uso da linguagem

natural em discussões no *Reddit* sobre temas relacionados ao suicídio. Os autores analisaram 2,9 milhões de postagens em 30 subreddits, identificando sentimentos de desconexão, sobrecarga, desesperança, desespero, resignação e trauma na comunidade ligada ao suicídio. Foi observada uma sobreposição linguística significativa entre postagens do *r/SuicideWatch* e outros subreddits, como *r/Depression*, *r/BPD* e *r/SocialAnxiety*. Os achados sugerem que a combinação de técnicas analíticas e de interpretabilidade pode oferecer novas compreensões sobre os conteúdos emocionais e temáticos compartilhados por indivíduos em risco de suicídio nessa plataforma.

Além da análise de grandes volumes de dados, LLMs também têm sido utilizados para gerar explicações interpretáveis sobre o risco de suicídio. Stern *et al.* (2024) utilizaram LLMs para gerar explicações em linguagem natural com base em postagens do *Reddit*, comparando os textos gerados com anotações feitas por especialistas em psicologia. As métricas *BLEU* e *ROUGE* foram empregadas para avaliar a qualidade das explicações, tendo como referência aquelas produzidas por especialistas humanos. O modelo *MentalBERT* foi ajustado como linha de base para classificar os níveis de risco de IS (nenhum, baixo, médio ou alto), enquanto modelos como *LLaMA 2 13B*, *WizardLM 13B*, *SpeechlessLM 13B* e *LLaMA 2 70B* foram avaliados na tarefa de geração de explicações. Os resultados mostraram que as previsões e justificativas fornecidas pelos LLMs estavam bem alinhadas com a percepção dos especialistas, destacando o potencial desses modelos em oferecer análises mais transparentes, confiáveis e humanamente compreensíveis.

3.3 Comparação com os Trabalhos Relacionados

Esta pesquisa de mestrado realiza contribuições que estão organizadas e apresentadas em dois capítulos desta dissertação. No Capítulo 4, exploramos métodos de ASSL com o objetivo de melhorar o desempenho do modelo original do Boamente, enfrentando desafios de escassez de dados rotulados e o desbalanceamento das classes. No Capítulo 5, propomos uma nova arquitetura para o Boamente, marcada pela integração de dois modelos complementares: um modelo classificador e um modelo explicador.

Em linha do que foi apresentado na Seção 3.1, no Estudo 1 adotamos técnicas de ASSL que possibilitaram o treinamento dos modelos a partir de dados inicialmente não rotulados, ampliando substancialmente o tamanho e a diversidade do conjunto de dados de treinamento. Essa abordagem foi especialmente vantajosa frente à limitação do volume de dados rotulados disponíveis no Boamente. Com a incorporação de dados não rotulados, o modelo pôde aprender padrões a partir de uma base mais ampla e variada, resultando em melhorias significativas no seu desempenho e na sua capacidade de generalização para a tarefa de análise de sentimentos e classificação de textos no idioma PT-BR.

Diferentemente dos estudos apresentados na Seção 3.2 e de trabalhos anteriores

do grupo de pesquisa do Projeto Boamente (DINIZ *et al.*, 2022; de Oliveira *et al.*, 2022; OLIVEIRA; BESSA; TELES, 2024; AZEVEDO; OLIVEIRA; TELES, 2024; OLIVEIRA *et al.*, 2025), realizamos no Estudo 2 uma análise comparativa do desempenho de diferentes versões ajustadas de pequenos LLMs generativos (especificamente Qwen, Mistral, Phi e LLaMA) na tarefa de detecção de IS em textos escritos em PT-BR. Esses modelos passaram por um processo de ajuste fine em um conjunto de dados rotulado para a detecção de IS, e suas versões ajustadas foram comparadas quanto à eficácia e precisão nessa tarefa. Essa análise oferece contribuições relevantes sobre a aplicabilidade e desempenho desses modelos para a detecção de risco de suicídio em PT-BR.

Além disso, o Estudo 2 vai além das métricas de desempenho ao investigar a qualidade das explicações textuais geradas por pequenos LLMs (modelos de linguagem com menor número de parâmetros). Para garantir uma avaliação rigorosa e confiável, contamos com a colaboração de três profissionais especializados: um psicólogo, um cientista da computação e um linguista com expertise em PT-BR. Suas percepções e análises fornecem uma base sólida para compreender as forças e limitações desses modelos no que se refere à geração de explicações.

4 Estudo 1: Identificação de Ideação Suicida usando métodos de Aprendizado Semi-Supervisionado

Esse capítulo apresenta um primeiro estudo que objetivou melhorar o desempenho do modelo do sistema Boamente para a detecção de IS em textos em PT-BR, utilizando métodos de ASSL para ampliar o conjunto de dados rotulados e aprimorar a generalização do modelo. O capítulo é organizado em seções que abordam a coleta e preparação dos dados, o treinamento dos modelos com métodos semi-supervisionados, e a avaliação de desempenho, com ênfase na comparação dos métodos e nas métricas obtidas. Além disso, são discutidos os resultados e a análise do impacto das técnicas aplicadas, visando aprimorar a eficácia da ferramenta na detecção de IS em textos escritos em PT-BR.

4.1 Materiais e Métodos

Para a realização deste estudo, foram seguidas as etapas ilustradas na Figura 5, as quais são detalhadas nas seções subsequentes: coleta e preparação dos dados, treinamento dos modelos e avaliação de desempenho. As implementações foram realizadas utilizando a linguagem *Python* e as bibliotecas *Pandas*, *NumPy*, *PyTorch* e *Matplotlib*. O ambiente de desenvolvimento utilizado foi o *Google Colaboratory*, que proporcionou uma infraestrutura robusta composta por processador *Intel(R) Xeon(R)*, 65 GB de memória RAM e uma *Graphics Processing Unit* (GPU) *NVIDIA T4* com 22,5 GB de memória dedicada.

Figura 5 – Etapas da Metodologia.

4.1.1 Coleta e Preparação dos Dados

O conjunto inicial de dados utilizado foi o original do Boamente, o qual passou por um rigoroso processo de revisão por pares. Nessa etapa, psicólogos especializados avaliaram as sentenças e as categorizaram em duas classes: contendo ou não IS. Inicialmente, foram coletados 5,699 tweets. Após a coleta, três psicólogos, com diferentes abordagens terapêuticas, terapia cognitivo-comportamental, teoria psicanalítica e teoria humanística,

foram convidados a realizar a anotação dos dados. Para garantir a imparcialidade e minimizar possíveis vieses, os profissionais foram cuidadosamente selecionados com base nas suas abordagens distintas. Cada sentença foi classificada como positiva (indicando IS) ou negativa (não indicando IS). Sentenças com discordância entre os psicólogos foram excluídas, resultando em um conjunto de dados final composto por 3,788 amostras rotuladas.

Os dados não rotulados foram coletados da rede social *Reddit*, mais especificamente de três *subreddits* (i.e., comunidades temáticas dentro da plataforma) relacionadas à IS: *Ansiedade e Depressão*, *Transtornos Mentais* e *Desabafos*. A coleta dos dados foi realizada por meio da utilização da biblioteca *Praw*, que permite acessar a *Application Programming Interface* (API) do *Reddit*, facilitando a extração dos posts e comentários dessas comunidades.

A coleta gerou um total de 11,240 novas amostras, representando quase quatro vezes o tamanho do conjunto de dados original. O conjunto de dados coletado é composto por quatro colunas: *ID*, *Subreddit*, *Título* e *Descrição*. A coluna *ID* foi utilizada para identificar e remover exemplos duplicados. Em seguida, as colunas *ID*, *Subreddit* e *Título* foram descartadas, mantendo-se apenas o texto presente na coluna *Descrição* do post. Além disso, os dados passaram por um processo de limpeza, no qual links, emojis e caracteres especiais foram removidos. Após essa limpeza, os dados foram consolidado em um novo conjunto de dados. A Figura 6 ilustra de forma quantitativa a composição dos dados no conjunto Boamente e no conjunto extraído do *Reddit*.

No conjunto de dados original, a coleta foi realizada com base em palavras-chave específicas. Em contraste, na coleta realizada a partir do *Reddit*, os dados foram obtidos de forma livre, ou seja, sem a aplicação de filtros por palavras-chave ou termos específicos. A Figura 7 apresenta a nuvem de palavras do conjunto de dados do Boamente, enquanto a Figura 8 exibe a nuvem de palavras gerada a partir da coleta no *Reddit*. Ao comparar ambas as figuras, é possível perceber diferenças marcantes na frequência e na variedade dos termos utilizados em cada conjunto, refletindo a distinta natureza das coletas.

4.1.2 Treinamento dos Modelos com Métodos Semi-supervisionados

Inicialmente, os dados rotulados e não rotulados foram carregados e submetidos a um pré-processamento. Em seguida, o modelo *BERTimbau Large* (SOUZA; NOGUEIRA; LOTUFO, 2020) e seu tokenizador foram inicializados, com a configuração do ambiente computacional para utilização da GPU. Para o treinamento, foram definidos o otimizador *AdamW*, a função de perda *CrossEntropyLoss* e o número de épocas, estabelecido em três.

O modelo *BERTimbau* (SOUZA; NOGUEIRA; LOTUFO, 2020) foi treinado em uma vasta quantidade de texto em português pelo Web brasileira como Corpus (BrWaC)

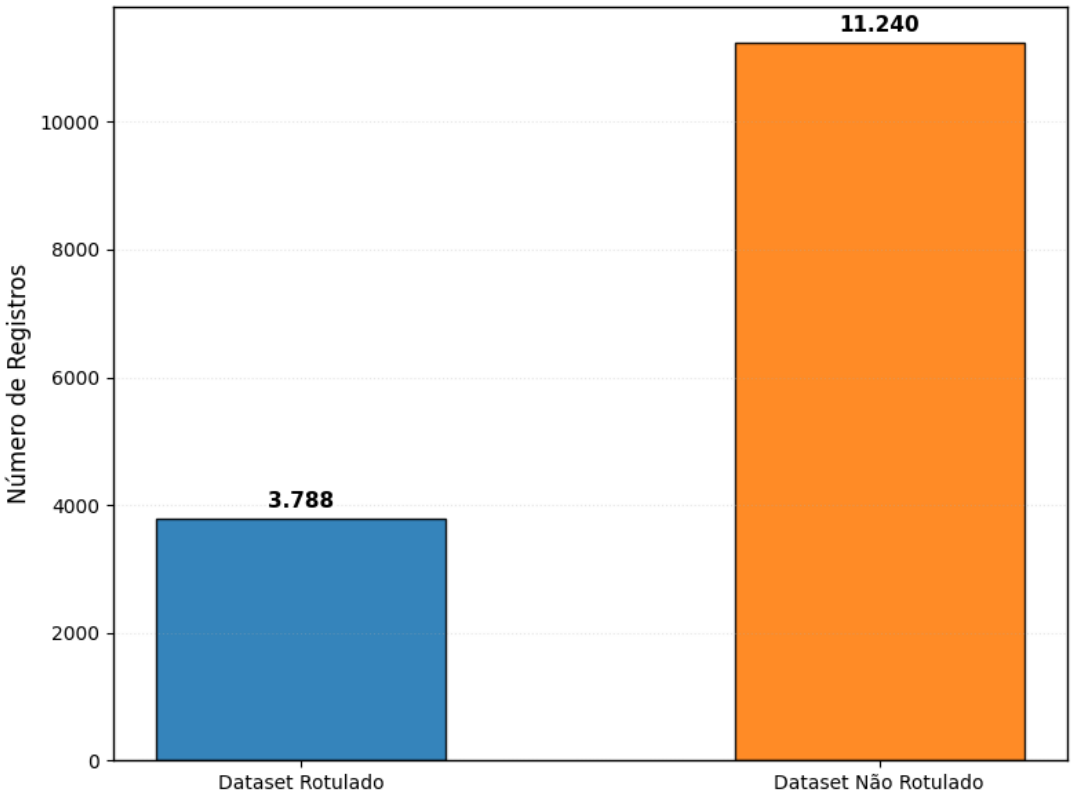
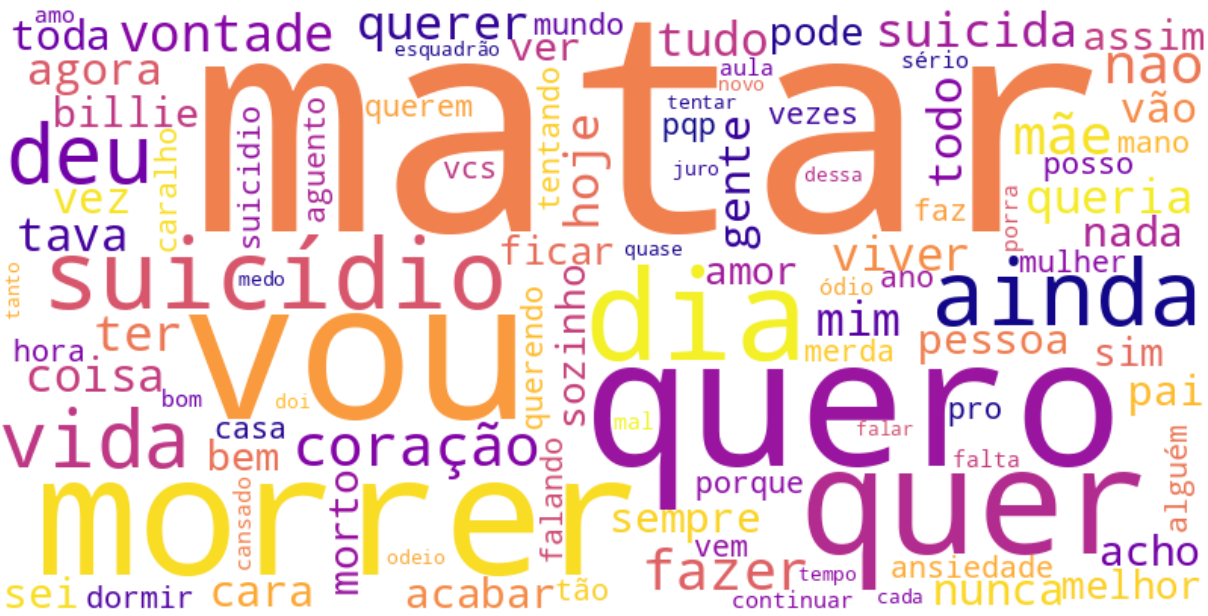


Figura 6 – Quantidade de exemplos nos conjuntos de dados rotulado e não rotulado.

Figura 7 – Nuvem de palavras do Boamente Original.



Fonte: Autor

(WAGNER *et al.*, 2018), possuindo uma compreensão aprimorada da estrutura linguística e dos padrões semânticos, permitindo uma representação mais rica dos dados. Essa representação contextualizada dos textos é baseada em *word embeddings*, sendo vetores numéricos que capturam o significado das palavras ao considerar o contexto em que elas aparecem. *Word embeddings* permitem que palavras com significados semelhantes fiquem

Figura 8 – Nuvem de palavras do conjunto de dados não rotulado, coletado do *Reddit*.

Fonte: Autor

próximas no espaço vetorial, facilitando a identificação de sinônimos e relações semânticas. Essa técnica é fundamental para tarefas de classificação, em que nuances e contextos sutis podem influenciar as decisões do modelo.

Utilizamos cinco métodos de ASSL: *self-learning*, *co-training*, *label propagation*, *weighting* e *boosting*. No *self-learning* e no *co-training* os principais hiperparâmetros usados foram o número de épocas definido em 3, o número de *folds* configurado em 5, e o número do *batch size*, que foi 8. O uso do *batch size* igual a 8 significa que durante cada iteração de treinamento do modelo, 8 exemplos de dados são processados simultaneamente. Isso implica que o modelo receberá 8 exemplos de entrada, calculará as saídas correspondentes para cada exemplo e, em seguida, atualizará os pesos com base na média dos gradientes desses 8 exemplos. Os três métodos restantes recorreram a um hiperparâmetro a mais, que foi o *learning rate* com o valor de *2e-5*, o qual controla o tamanho das atualizações dos pesos do modelo durante o treinamento.

Além disso, foi realizada uma etapa de propagação de rótulos para os dados não rotulados, utilizando diferentes limiares de confiança. No método de *self-learning*, foram considerados três limiares: 0,7, 0,8 e 0,9. Já no *co-training*, foi adotado apenas o limiar de 0,9. Apenas os exemplos cuja predição atingiu ou superou esses valores foram atribuídos com pseudo-rótulos e incorporados ao conjunto de dados rotulado. O número final de amostras rotuladas e descartadas foi contabilizado, assim como a distribuição das classes nos dados rotulados. Para uma análise mais detalhada, a distribuição dos níveis de confiança das predições foi visualizada por meio de histogramas, tanto para o *self-learning* quanto para o *co-training*, permitindo avaliar a confiabilidade dos rótulos atribuídos por cada abordagem.

4.1.3 Avaliação e Métricas

A avaliação dos métodos de aprendizado semi-supervisionado, *self-learning* e *co-training*, foi conduzida por meio da técnica de validação cruzada estratificada (*Stratified K-Fold Cross-Validation*) com K=5. Nessa abordagem, novas amostras rotuladas são progressivamente incorporadas ao conjunto de treinamento com base nos limiares de confiança previamente definidos. Para os demais métodos supervisionados, foi adotada a estratégia de validação *Hold-out*, com uma divisão de 80% para treino e 20% para teste.

Foram desenvolvidas classes personalizadas, como *CustomDataset* e *TextDataset*, para manipular os textos e seus respectivos rótulos, utilizando o tokenizador do *BERTimbau*. Ao final de cada época de treinamento, o modelo realiza a predição dos rótulos para os dados não rotulados, registrando tanto os rótulos atribuídos quanto os níveis de confiança obtidos, especialmente na última época.

A avaliação do desempenho dos modelos foi realizada com base em cinco métricas: *acurácia*, *precisão*, *recall*, *F1-score* e *Area Under the Curve* (AUC). Para os métodos com validação cruzada, a média das métricas ao longo das cinco dobras foi calculada, proporcionando uma estimativa mais robusta e confiável do desempenho geral dos modelos.

4.2 Resultados

Inicialmente, foi realizada uma análise exploratória dos dados rotulados utilizados no treinamento do modelo. A Figura 9 apresenta a distribuição dos textos de acordo com seus respectivos níveis de confiança. Observa-se que a maior parte das amostras concentra-se em valores superiores a 0,9, o que indica uma alta certeza nas predições do modelo. Com base nessa observação, foram definidos os limiares de confiança de 0,7, 0,8 e 0,9 como critérios para a atribuição de pseudo-rótulos neste estudo.

A Tabela 1 apresenta os resultados parciais obtidos pelos métodos empregados neste trabalho. Para facilitar a análise comparativa, os melhores valores de cada métrica estão destacados em negrito.

A Figura 10 apresenta o gráfico comparativo da métrica de *recall* entre os resultados obtidos neste estudo e o valor correspondente do modelo original do Boamente. Essa métrica é especialmente relevante, pois, no contexto de identificação de IS, é fundamental minimizar os falsos negativos, ou seja, casos em que textos com sinais de ideação não são identificados pelo modelo. Os resultados demonstram que, na maioria dos cenários, os métodos propostos superam o desempenho do modelo original, reforçando a eficácia das abordagens baseadas em aprendizado semi-supervisionado.

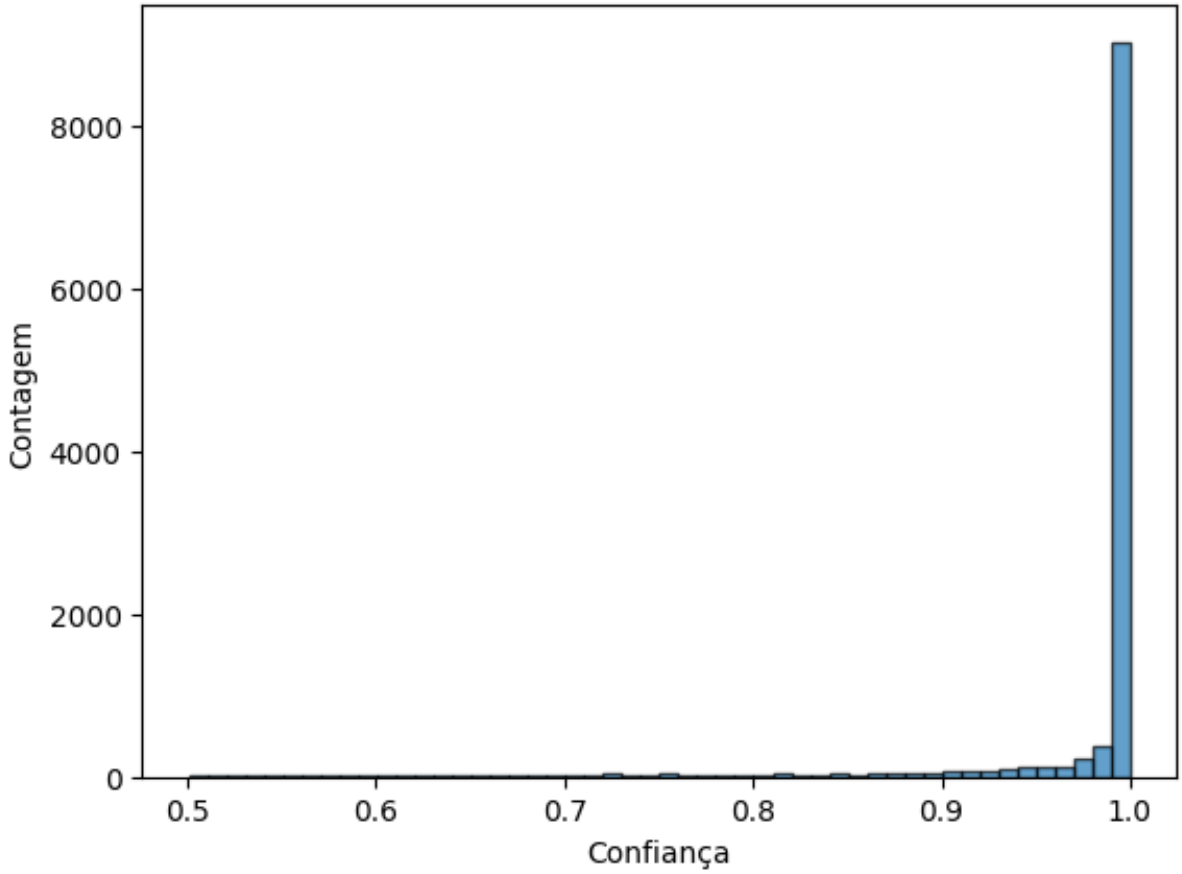


Figura 9 – Histograma que apresenta a distribuição dos níveis de confiança atribuídos às amostras do conjunto de dados não rotulado.

Tabela 1 – Desempenho dos modelos com diferentes métodos usados.

Método	Acurácia	Precisão	*Recall*	F1-score	AUC
Boamente	0,9555	0,9612	0,9530	0,9544	0,9545
Self-Learning (Confiança 0,9)	0,9960	**0,9954**	0,9909	0,9931	0,9945
Self-Learning (Confiança 0,8)	**0,9966**	**0,9954**	**0,9927**	**0,9941**	**0,9954**
Self-Learning (Confiança 0,7)	0,9945	0,9892	0,9918	0,9905	0,9937
Co-Training (Confiança 0,9)	0,9813	0,9798	0,9545	0,9666	0,9733
Label Propagation	0,9783	0,9721	0,9517	0,9616	0,9705
Weigthing	0,9852	0,9784	0,9708	0,9745	0,9809
Boosting	0,9830	0,9740	0,9809	0,9709	0,9860

4.3 Discussão

Neste estudo, comparamos a eficácia de diferentes métodos de ASSL para aprimorar o desempenho na classificação de um conjunto de dados não rotulado. Os resultados obtidos visam explorar a aplicação de técnicas avançadas de PLN e AM para abordar um problema crítico de saúde mental. O objetivo foi oferecer novas abordagens para a detecção de IS, proporcionando suporte a indivíduos em situações de vulnerabilidade e aprimorando a

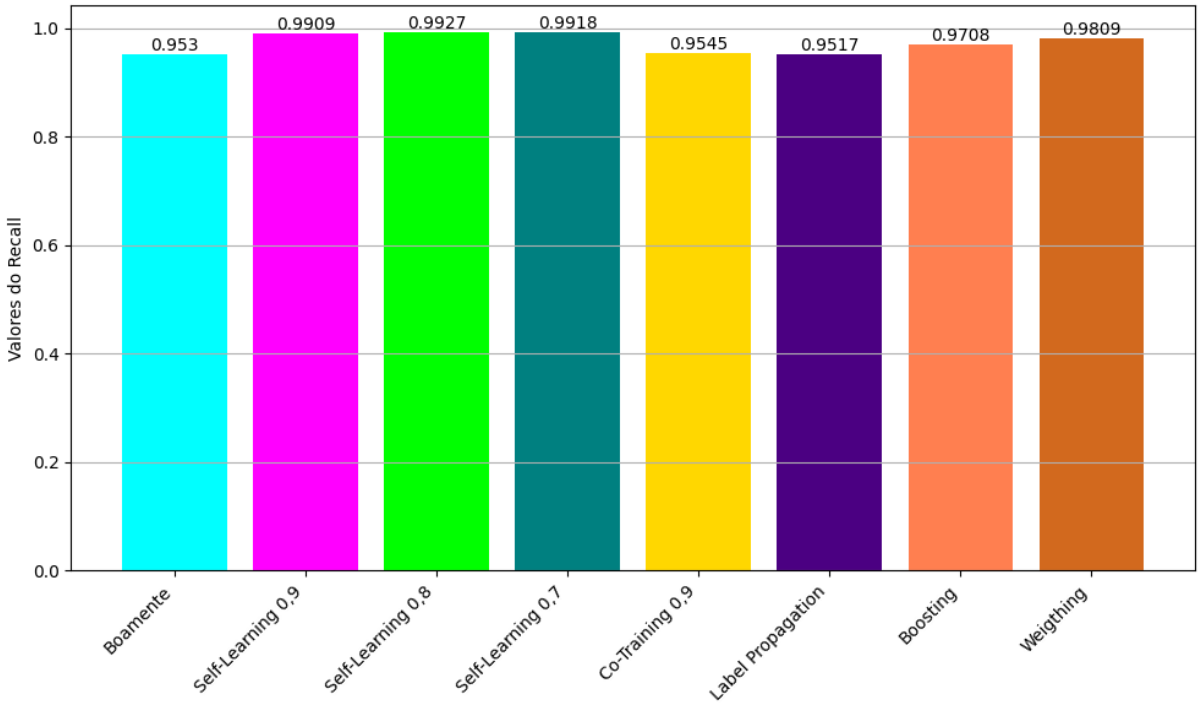


Figura 10 – Comparação dos resultados obtidos para a métrica de *recall* entre os diferentes métodos avaliados e o modelo original do Boamente.

tomada de decisões pelos PSMs. Dessa forma, esperamos contribuir para o desenvolvimento de estratégias de prevenção e intervenção mais eficazes.

Os resultados mostram que o método *self-learning* obteve o melhor desempenho, seguido pelos métodos *weighting*, *boosting*, *co-training* e *label propagation*. Vale ressaltar que a maioria dos métodos avaliados superaram o desempenho do modelo original do Boamente.

As técnicas de ASSL permitiram que os modelos fossem treinados com dados não rotulados, ampliando o tamanho e a diversidade do conjunto de dados de treinamento. Isso foi especialmente benéfico considerando a limitação do conjunto de dados rotulado do Boamente. Ao aplicar essas técnicas, conseguimos aumentar a quantidade e a variedade de dados utilizados, o que resultou em um melhor desempenho e maior capacidade de generalização do modelo na tarefa de AET em textos escritos em PT-BR. Com base nos resultados, observamos uma melhoria significativa no desempenho do modelo do Boamente. Acreditamos que, com esses avanços, os PSMs podem ter maior confiança na ferramenta proposta para a identificação de IS, permitindo também um monitoramento mais preciso dos pacientes.

5 Estudo 2: LLMs Generativos para Detectar e Explicar Ideação Suicida em Textos

Este capítulo apresenta o segundo estudo desenvolvido neste trabalho de mestrado, cujo objetivo foi propor uma nova arquitetura do sistema Boamente por meio da integração de LLMs, com a finalidade de detectar e explicar a presença de IS em textos escritos em PT-BR. O estudo buscou não apenas aumentar a eficácia da detecção de IS, mas também fornecer explicações textuais que justifiquem cada previsão realizada pelo modelo. Neste capítulo, são descritos os materiais utilizados, o processo de fine-tuning dos modelos, a engenharia de prompt aplicada para a geração das explicações, os métodos de avaliação quantitativa e qualitativa empregados, assim como os resultados obtidos com os modelos testados.

5.1 Materiais e Métodos

Este estudo estende a primeira versão da arquitetura Boamente (DINIZ *et al.*, 2022) ao integrar LLMs para detectar IS em textos e fornecer explicações para casos positivos e negativos. Portanto, a arquitetura proposta neste trabalho (Figura 11) não apenas classifica textos não clínicos com base na presença ou ausência de IS, mas também gera explicações textuais para suas previsões. Nessa arquitetura aprimorada, o texto de entrada é primeiro processado por um modelo classificador, que determina se o texto contém IS. Em seguida, um modelo explicador gera uma justificativa textual para a classificação. Os resultados, incluindo tanto a classificação quanto sua explicação, são disponibilizados por meio de uma aplicação web para visualização por profissionais de saúde mental.

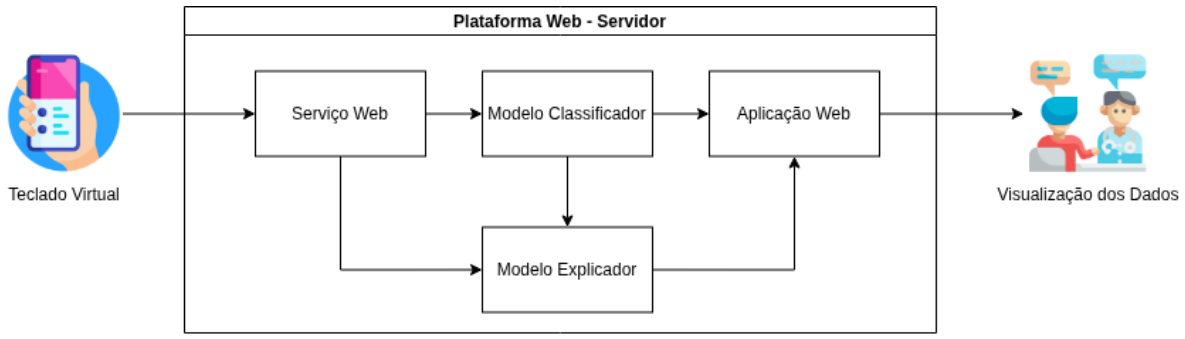


Figura 11 – Nova arquitetura proposta do Boamente.

A seguir, detalhamos o processo de fine-tuning dos LLMs para atuarem como modelos classificadores na arquitetura proposta, bem como sua avaliação quantitativa.

Além disso, apresentamos a estratégia empregada para que os LLMs gerem explicações textuais e descrevemos uma avaliação qualitativa de uma amostra dessas justificativas.

5.1.1 Fine-tuning LLMs para Classificação

O conjunto de dados Boamente (DINIZ *et al.*, 2022; TELES *et al.*, 2023) foi utilizado para o fine-tuning dos LLMs. Ele contém 3,788 textos rotulados em duas classes: positivos (1,097 amostras) e negativos (2,691 amostras) para IS. As anotações foram feitas por três psicólogos com abordagens distintas: cognitivo-comportamental, psicanalítica e humanista. Para melhorar a qualidade dos dados, foi aplicada uma etapa de pré-processamento para remover elementos como hashtags, links e caracteres especiais. Diferentemente do estudo original do Boamente, nenhuma técnica de balanceamento de classes (como oversampling ou undersampling) foi aplicada neste segundo estudo. Essa decisão foi tomada para avaliar o desempenho do modelo sob a distribuição natural do conjunto de dados. A Figura 12 ilustra a distribuição dos dados, incluindo o conjunto completo e a divisão estratificada entre treino e teste.

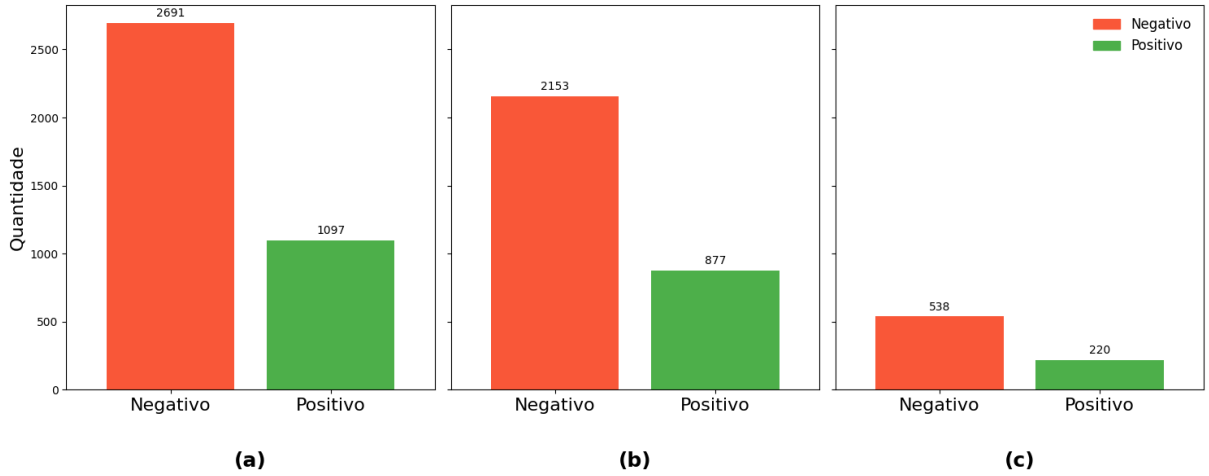


Figura 12 – Distribuição de classes no conjunto de dados: (a) conjunto completo, (b) conjunto de treinamento (4 folds - 80%) e (c) conjunto de teste (1 fold - 20%).

O fine-tuning dos LLMs foi realizado em uma máquina com as seguintes especificações: *Intel Core i9* de 13ª geração, 128 GB de RAM, SSD de 2TB e GPU *NVIDIA GeForce RTX 4090*. O treinamento dos modelos foi conduzido usando a biblioteca *Unsloth* (HAN; TEAM, 2023), que permite o fine-tuning eficiente de LLMs em GPU com suporte a 4 bits. Para otimizar o processo, empregamos a técnica *Quantized Low-Rank Adaptation* (QloRA), que combina quantização em 4 bits com adaptações de uma *low-rank matrix* às matrizes de pesos originais. Essa abordagem reduz significativamente o consumo de memória e os custos computacionais, mantendo o desempenho do modelo (DETTMERS *et al.*, 2023). O fine-tuning foi implementado usando o *framework Parameter-Efficient Fine-Tuning* (PEFT).

Diversos *Instruct Models* (ITM) foram testados para avaliar seu desempenho na classificação de IS. ITM são ajustados em conjuntos de dados de instrução-resposta, permitindo que sigam melhor prompts humanos e se alinhem à intenção do usuário. Fine-tuning em um ITM (em vez de um modelo base) é preferível porque exige menos dados, melhora a generalização, reduz custos computacionais e garante saídas mais seguras e confiáveis. Os modelos testados incluíram *Qwen 2.5* (14B, 7B, 1.5B, 500M), *Qwen 2* (7B, 3B, 1.5B), *Mistral Nemo* (12B), *LLaMA 3.1* (8B), *LLaMA 3* (8B), *Phi-3 Medium* (14B) e *Phi-3 Mini* (3.8B).

A configuração do *QLoRA* seguiu os hiperparâmetros padrão da biblioteca *Unsloth*. O modelo foi treinado por uma época, com tamanho de lote *(batch size)* de 32 e passos de acumulação de gradiente definidos como 1. O processo de otimização teve 76 passos de treinamento no total, com taxa de aprendizagem de 1e-4 e 10 passos de aquecimento (warmup), facilitando uma adaptação gradual dos pesos para melhorar a estabilidade da convergência. Além disso, um decaimento de peso *(weight decay)* de 0.01 foi aplicado para mitigar *overfitting*, enquanto um agendador de taxa de aprendizagem em cosseno foi usado para regular suavemente a dinâmica de aprendizagem. O otimizador *AdamW* de 8 bits foi empregado para melhorar a eficiência computacional.

O prompt utilizado no fine-tuning dos LLMs para classificação foi projetado para produzir uma resposta simples ("sim" ou "não"), indicando a presença ou ausência de IS, conforme abaixo.

Prompt zero-shot para LLMs classificadores.

Esta é uma frase.
{FRASE}

Se a frase contém ideação suicida, responda com "Sim". Caso contrário, responda com "Não".

SOLUÇÃO
A resposta correta é: "{}"

5.1.2 Avaliação dos LLMs

Para avaliar o desempenho dos modelos para classificação, foi realizada uma avaliação quantitativa usando as métricas como precisão, acurácia, *recall*, *F1-Score* e AUC. A curva *Receiver Operating Characteristic* (ROC) também foi utilizada para avaliar o desempenho dos modelos. *Recall* e AUC foram priorizados por serem especialmente relevantes no contexto de detecção de IS. O *recall* foi enfatizado para minimizar falsos negativos, garantindo que textos contendo IS fossem corretamente identificados, algo crucial em aplicações de saúde mental, onde deixar passar um caso positivo pode ter

consequências graves. A AUC foi usada para avaliar a capacidade geral do modelo de distinguir entre as duas classes em diferentes configurações de limiar. O ajuste fino e a validação do modelo seguiram uma abordagem de validação cruzada de 5 folds, com divisão estratificada entre treino e teste em cada fold (Figura 12).

5.1.3 Engenharia de Prompt para Geração de Explicações

Para gerar explicações, selecionamos dois exemplos do conjunto de dados (um positivo e um negativo) e usamos o GPT-4o para explicar seus respectivos rótulos. Após revisar essas explicações, as incorporamos em um prompt *few-shot* e o fornecemos como entrada aos três modelos com melhor desempenho (um de cada família de LLM) na etapa de classificação. O objetivo foi permitir que os ITM gerassem suas próprias explicações, visto que recebiam o resultado da classificação (positivo ou negativo para IS) do LLM classificador. A seguir, apresentamos o prompt *few-shot* utilizado nesse processo.

Prompt *few-shot* para LLMs explicadores.

Você é um especialista em saúde mental.
Sua tarefa é analisar a {presença/ausência} de ideação suicida em textos.

Baseie sua explicação em evidências linguísticas e psicológicas, destacando padrões emocionais, escolha de palavras e estrutura do discurso. Seja objetivo e conciso em sua análise.

Exemplo 1 - Texto com ideação suicida:
Texto: exemplo 1
Explicação: explicação 1

Exemplo 2 - Texto sem ideação suicida:
Texto: exemplo 2
Explicação: explicação 2

Agora, analise o seguinte texto e forneça uma explicação similar aos exemplos acima:
Texto: “texto”
Explicação: “{}”

5.1.4 Avaliação de LLMs Explicadores

Para avaliar os modelos explicadores, recrutamos três profissionais: um cientista da computação, um linguista de PT-BR e um psicólogo. Cada profissional trouxe uma perspectiva distinta à avaliação: o cientista da computação avaliou a coerência técnica e a consistência das explicações geradas; o linguista avaliou a clareza linguística e a legibilidade

dos textos gerados; e o psicólogo verificou se as explicações eram clinicamente relevantes e significativas. Eles analisaram 10 amostras de texto (cinco de cada classe) do conjunto Boamente, juntamente com as explicações geradas pelos três LLMs explicadores avaliados.

A avaliação foi conduzida por meio de um formulário do *Google Forms*, no qual os profissionais atribuíram notas às explicações em uma escala de 0 a 5, sendo 0 para uma explicação de baixa qualidade e 5 para uma excelente. Para cada amostra, eles puderam justificar sua avaliação, explicando o raciocínio por trás das notas. Todos os participantes receberam detalhes sobre o estudo. A pesquisa recebeu aprovação ética do Comitê de Ética em Pesquisa Envolvendo Seres Humanos da Universidade Federal do Delta do Parnaíba (número de aprovação 5.787.194).

5.2 Resultados

A Figura 13 apresenta as matrizes de confusão dos modelos avaliados. Observa-se que alguns modelos, como o *Qwen 2.5 (14B)* e o *Qwen 2.5 (7B)*, alcançaram resultados satisfatórios, exibindo um baixo número de erros. Em contraste, os modelos das famílias Phi e Mistral demonstraram desempenho significativamente inferior.

A Tabela 2 apresenta os resultados obtidos na avaliação dos LLMs classificadores, destacando as diferenças nas métricas de desempenho. Os valores em negrito representam os melhores resultados para *recall* e AUC. Os maiores valores de *recall* foram alcançados por *Qwen 2.5 (7B)* e *Qwen 2 (3B)*, enquanto as maiores AUCs foram obtidas por *Qwen 2.5 (14B)*, *Qwen 2.5 (7B)* e *Qwen 2 (3B)*.

Tabela 2 – Resultados de desempenho dos LLMs classificadores avaliados.

Modelo	Acurácia	Precisão	Recall	F1-score	AUC
Qwen 2.5 (14B)	0,9591	0,9522	0,9045	0,9277	**0,9898**
Qwen 2.5 (7B)	0,9538	0,8936	**0,9545**	0,9231	**0,9890**
Qwen 2 (3B)	0,9591	0,9091	**0,9545**	0,9313	**0,9858**
Qwen 2 (7B)	0,9551	0,9115	0,9364	0,9238	**0,9770**
Qwen 2 (1.5B)	0,9485	0,9249	0,8955	0,9099	**0,9769**
Qwen 2.5 (1.5B)	0,9393	0,8750	0,9227	0,8982	**0,9769**
Qwen 2.5 (500M)	0,9380	0,8844	0,9045	0,8944	**0,9759**
Mistral Nemo (12B)	0,7850	0,5882	0,8636	0,6998	0,8751
LLama 3.1 (8B)	0,7639	0,5635	0,8273	0,6703	0,8489
LLama 3 (8B)	0,7784	0,5942	0,7455	0,6613	0,8441
Phi-3 Medium (14B)	0,6992	0,4892	0,8273	0,6149	0,7936
Phi-3 Mini (3.8B)	0,6675	0,4437	0,5727	0,5000	0,6813

A Figura 14 mostra as curvas ROC dos modelos avaliados, ilustrando sua capacidade

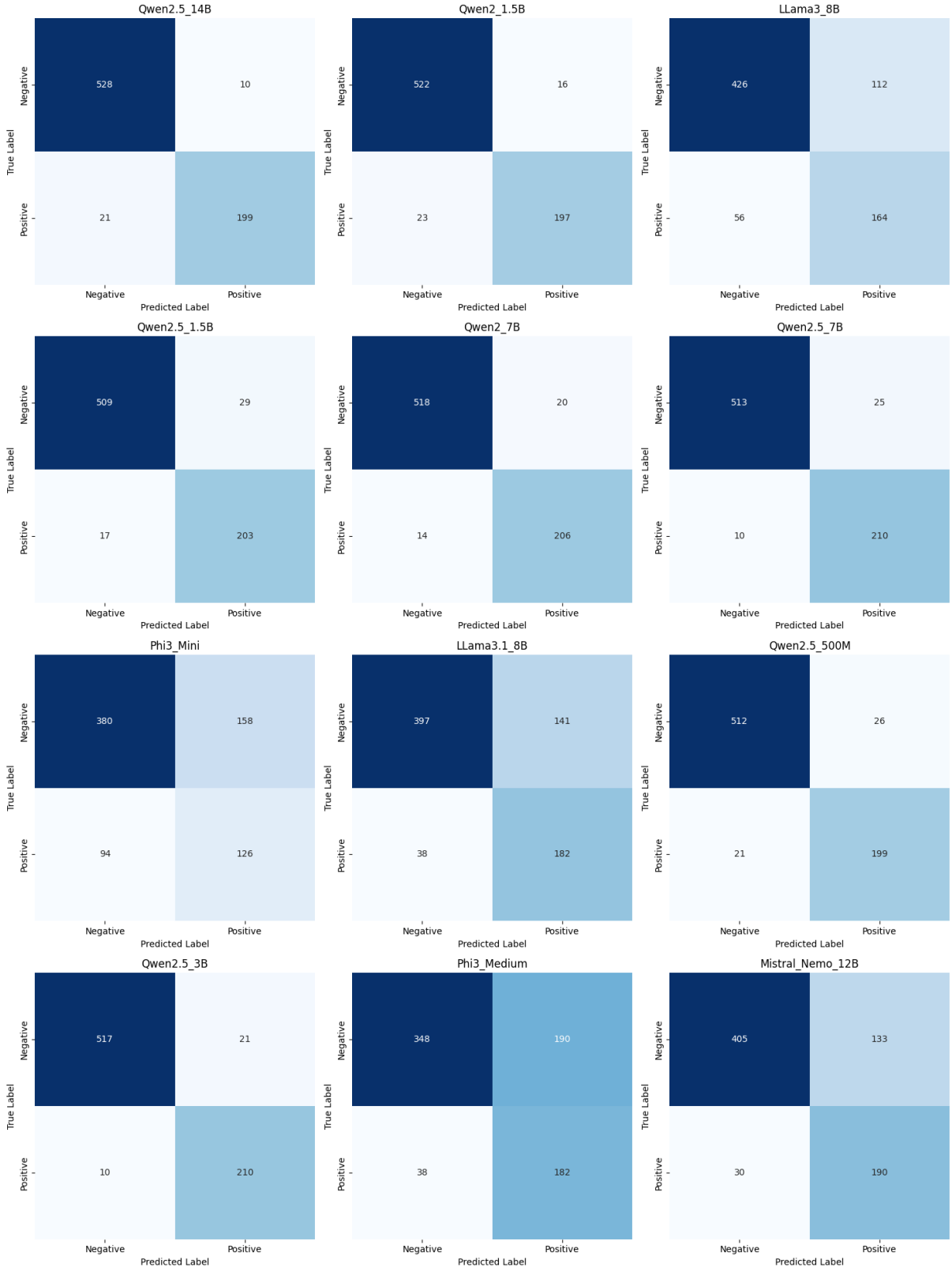


Figura 13 – Matrizes de Confusão dos Modelos.

de distinguir entre textos com e sem IS. Em nossa análise, consideramos apenas os melhores modelos de cada família. O modelo *Qwen 2.5 (14B)* obteve o melhor desempenho, com AUC de 0,9898, indicando uma discriminação quase ideal. Além disso, os modelos da

família *Qwen* apresentaram valores de AUC próximos, conforme destacado na Tabela 2.

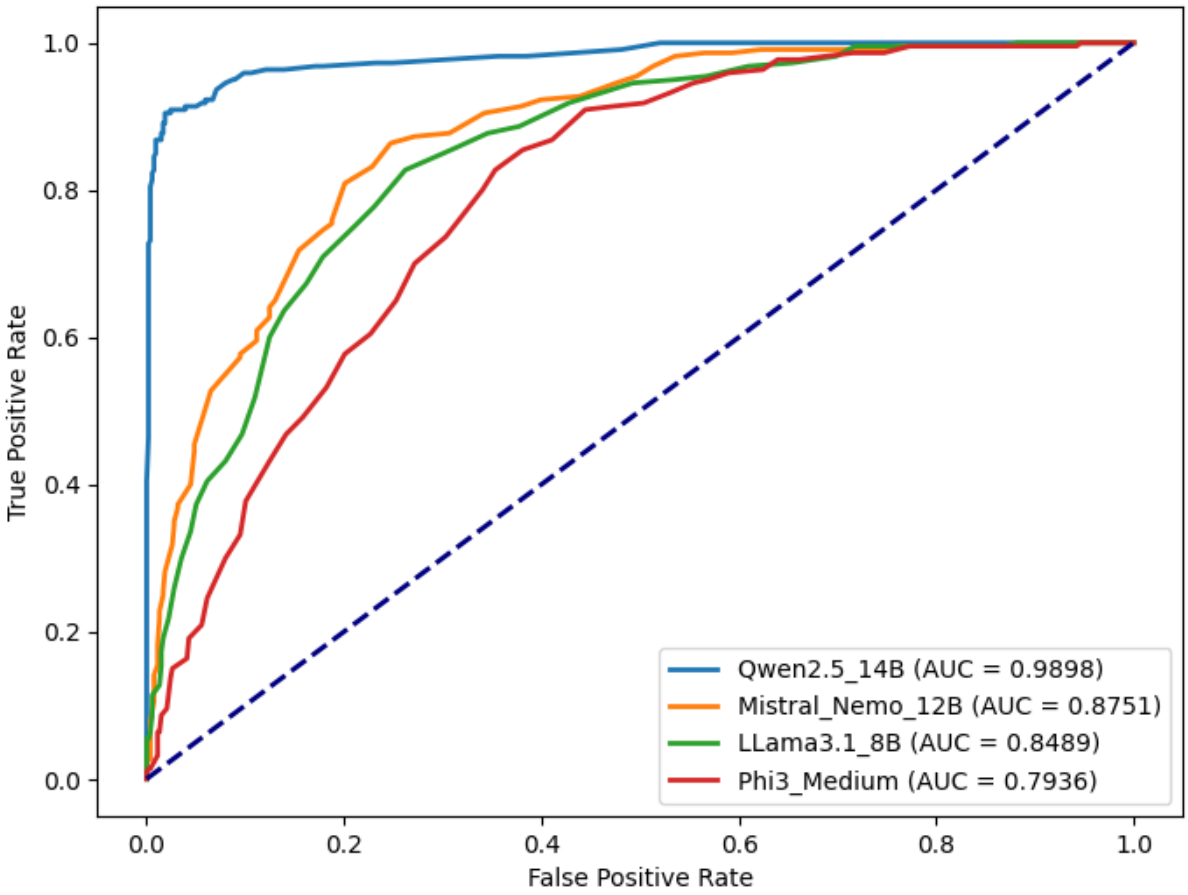


Figura 14 – Curvas ROC dos melhores modelos de cada família avaliados.

Na análise qualitativa, selecionamos apenas os modelos com melhor desempenho de cada família (AUC> 0,80): *Qwen 2.5 (14B)*, *Mistral Nemo (12B)* e LLaMA 3.1 (8B). Para avaliar sua efitividade, calculamos as notas médias (atribuídas pelos profissionais a cada explicação) para cada modelo, conforme mostrado na Figura 15.

Ao analisar as avaliações individuais e justificativas dos profissionais, o cientista da computação atribuiu a nota mais alta (4,2) ao LLaMA 3.1 (8B). Ele classificou o *Qwen 2.5 (14B)* como moderado (3,6) e deu a nota mais baixa (3,0) ao *Mistral Nemo (12B)*, sugerindo possível falta de consistência em suas explicações. Ele também apontou que a ausência de contexto nas amostras pode ter levado o modelo a gerar respostas mais genéricas.

Do ponto de vista do linguista, tanto o LLaMA 3.1 (8B) quanto o *Qwen 2.5 (14B)* receberam a mesma nota (3,9), indicando fluência e legibilidade similares, enquanto o *Mistral Nemo (12B)* obteve um desempenho ligeiramente inferior (3,7). Assim como o cientista da computação, o linguista também identificou o uso excessivo de expressões coloquiais como uma limitação.

Focando na aplicabilidade clínica das explicações, o psicólogo avaliou tanto o *Mistral*

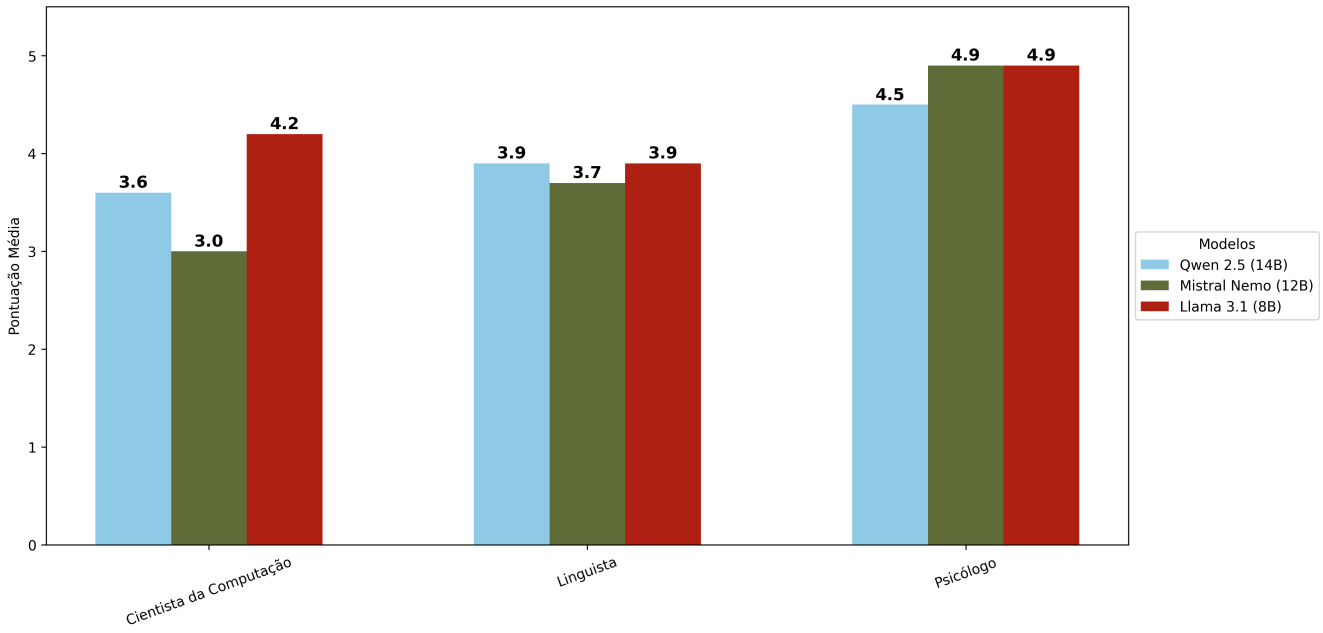


Figura 15 – Notas médias atribuídas pelos profissionais.

Nemo (12B) quanto o LLaMA 3.1 (8B) com a mesma nota (4,9), indicando que ambos os modelos forneceram justificativas relevantes e significativas para explicar a IS. Curiosamente, o *Qwen 2.5 (14B)* recebeu uma nota um pouco menor (4,5), apesar do forte desempenho em outras áreas. Isso sugere que, embora o *Qwen 2.5 (14B)* tenha demonstrado robustez técnica, algumas de suas explicações podem ter carecido da profundidade psicológica necessária para interpretação clínica. O psicólogo também observou que certos termos apresentavam inconsistências sintáticas, o que, em alguns casos, poderia afetar a qualidade geral das explicações.

5.3 Discussão

5.3.1 Análise de Performance do Modelo

Como parte da análise quantitativa, os resultados mostram que os modelos da família Qwen apresentaram o melhor desempenho geral, com *Qwen 2.5 (7B)* e *Qwen 2 (3B)* alcançando os maiores valores de recall, a métrica mais relevante para nosso estudo. Vale destacar que um dos maiores modelos testados, o *Qwen 2.5 (14B)*, obteve a maior AUC, atingindo quase 99% de eficácia. Contudo, apesar desse alto valor, seu recall foi inferior ao do Qwen 2.5 (7B), um modelo com metade de seu tamanho. Isso sugere que o *Qwen 2.5 (14B)* falhou em identificar corretamente alguns casos positivos, que são críticos para o sistema Boamente.

Neste estudo, utilizamos quatro modelos como baseline: *BERTimbau Large*, *BERTimbau Base*, GPT-4, (Microsoft Copilot) e GPT-3.5 (ChatGPT), explorados em nossos

trabalhos anteriores em (DINIZ *et al.*, 2022; OLIVEIRA; BESSA; TELES, 2024). Em (DINIZ *et al.*, 2022), realizamos o fine-tuning dos modelos *BERTimbau Large* e *BERTimbau Base* usando o conjunto de dados Boamente e avaliamos os modelos com validação cruzada de 5 folds. Em (OLIVEIRA; BESSA; TELES, 2024), avaliamos GPT-4 e GPT-3.5, dois modelos de linguagem de propósito geral, usando apenas 100 amostras do conjunto Boamente (50 de cada classe). Embora reconheçamos que essa abordagem de validação é limitada em comparação com a validação cruzada (devido à sua dependência de um conjunto de teste pequeno e fixo que pode não representar plenamente a distribuição geral dos dados ou capturar sua variabilidade), ainda consideramos valioso comparar e analisar os resultados.

Como apresentado na Tabela 3, nossos dois melhores modelos (AUC acima de 0,98) superaram os modelos baseados em BERT. Mesmo sem aplicar técnicas de balanceamento de classes, como fizemos no treinamento dos modelos BERT, nossos modelos apresentaram desempenho promissor, demonstrando robustez ao lidar com os dados em sua distribuição natural. Além disso, os resultados mostram que nossos LLMs com fine-tuning performaram de forma comparável a LLMs generativos comerciais, como ChatGPT e Copilot, apesar de terem menos parâmetros.

Tabela 3 – Comparação dos nossos melhores modelos com os melhores de (DINIZ *et al.*, 2022; OLIVEIRA; BESSA; TELES, 2024).

Modelo	Recall	AUC	Tipo de Validação
BERTimbau Large (DINIZ *et al.*, 2022)	0,9530	0,9545	Cross-validation
BERTimbau Base (DINIZ *et al.*, 2022)	0,9450	0,9450	Cross-validation
GPT-4 (OLIVEIRA; BESSA; TELES, 2024)	**0,9800**	0,9800	100 Amostras
GPT-3.5 (OLIVEIRA; BESSA; TELES, 2024)	0,8600	0,8100	100 Amostras
Qwen 2.5 (14B)	0,9045	**0,9898**	Cross-validation
Qwen 2.5 (7B)	0,9545	**0,9890**	Cross-validation

5.3.2 Análise Qualitativa

As justificativas dos participantes para algumas amostras indicam que a mesma explicação pode ser interpretada de maneira diferente. Observamos que o LLaMA 3.1 (8B) recebeu a pontuação geral mais alta, indicando que suas explicações foram consideradas de qualidade superior. No entanto, é importante destacar que tanto o psicólogo quanto o linguista atribuíram pontuações similares a outros modelos, sugerindo que a qualidade das explicações não foi exclusivamente superior em um único modelo.

Sob a perspectiva do psicólogo, a maioria das explicações foi clara e compreensível, embora algumas amostras geradas pelo *Qwen 2.5 (7B)* apresentassem problemas de sintaxe que poderiam dificultar a interpretação. O linguista reconheceu que os modelos

produziram explicações de alta qualidade, mas apontou que amostras contendo termos excessivamente coloquiais tenderam a gerar justificativas mais genéricas e com menor profundidade analítica. Para o cientista da computação, a falta de contexto nas amostras foi um fator crítico, pois poderia levar o modelo a classificar incorretamente um texto como contendo ou não IS. Segundo ele, fornecer contexto adicional ajudaria a produzir explicações mais precisas e bem fundamentadas.

Portanto, esses achados destacam que diferentes perspectivas profissionais influenciaram a avaliação dos modelos, demonstrando a importância de uma análise multidisciplinar na interpretação dos resultados. A diversidade de interpretações sobre as justificativas apresentadas pelos participantes aponta que uma mesma explicação pode ser recebida de forma distinta conforme o avaliador. Os dados demonstram que o LLaMA 3.1 (8B) obteve a maior média de pontuação, sinalizando que suas respostas foram percebidas como qualitativamente superiores. Contudo, é relevante mencionar que psicólogos e linguistas distribuíram notas comparáveis a outros sistemas, indicando que a excelência das explicações não se concentrou em um único modelo.

Sob o olhar do psicólogo, a maior parte das respostas foi considerada coerente e acessível, embora amostras do *Qwen 2.5 (7B)* tenham exibido inconsistências gramaticais capazes de comprometer a clareza. Já o linguista valorizou a qualidade geral das explicações, mas criticou o uso excessivo de registros informais em certos casos, associando essa característica a análises superficiais e pouco detalhadas. Para o cientista da computação, a ausência de contextualização nas entradas emergiu como uma limitação central, já que aumentaria o risco de classificações equivocadas sobre a presença (ou não) de IS. Ele defendeu que informações complementares enriqueceriam a precisão das explicações geradas.

Esses resultados evidenciam como vieses disciplinares moldaram a percepção de qualidade, reforçando a necessidade de integrar múltiplas expertises na avaliação de sistemas de IA. A convergência parcial entre as críticas sugere que, embora modelos como o LLaMA 3.1 se destaquem, a busca por explicações ideais requer equilibrar critérios técnicos, linguísticos e cognitivos.

6 Considerações Finais

Os objetivos deste trabalho de mestrado foram alcançados. Inicialmente, buscou-se investigar o impacto de métodos de ASSL na melhoria do desempenho do modelo Boamente, e os resultados evidenciaram ganhos significativos em métricas cruciais, especialmente no recall, fundamental para a detecção de IS. Além disso, investigou-se o uso de LLMs tanto na tarefa de classificação, com o objetivo de avaliar seu desempenho preditivo, quanto na geração de explicações por meio de técnicas de engenharia de prompt. A análise quantitativa, voltada para as métricas de desempenho, e a análise qualitativa, realizada por profissionais especializados evidenciaram que os modelos avaliados, além de apresentarem desempenho competitivo na identificação da IS, também foram capazes de gerar justificativas claras, coerentes e relevantes. Esses resultados reforçam o potencial dos LLMs não apenas como classificadores eficazes, mas também como ferramentas promissoras para aumentar a interpretabilidade e a confiabilidade do sistema junto aos PSMs. Dessa forma, confirma-se a hipótese de que a aplicação de técnicas semi-supervisionadas e a incorporação de LLMs podem aprimorar tanto o desempenho quanto a transparência de modelos voltados à identificação de sinais de IS.

6.1 Contribuições

6.1.1 Científicas

As contribuições científicas deste trabalho abrangem dois eixos complementares. Primeiramente, explorou-se o uso de técnicas de ASSL com foco em AET para aprimorar o desempenho do sistema Boamente na detecção de IS em textos não clínicos em PT-BR. A utilização de dados rotulados e não rotulados resultou em melhorias significativas na acurácia dos modelos, demonstrando o potencial dessa abordagem em contextos sensíveis como a saúde mental. Em paralelo, propôs-se uma nova arquitetura baseada em LLMs, integrando não apenas a classificação, mas também a geração de explicações por meio de engenharia de prompt, um aspecto ainda pouco explorado na literatura. Diferentemente de abordagens tradicionais que se limitam à classificação binária, este estudo demonstrou que é possível desenvolver soluções que conciliem desempenho preditivo com interpretabilidade. Dessa forma, o trabalho contribui para o avanço do estado da arte em PLN aplicado à saúde mental, ao oferecer uma ferramenta mais robusta, transparente e alinhada às necessidades dos PSMs no contexto da prevenção ao suicídio.

Adicionalmente, os experimentos conduzidos reforçam achados importantes para a comunidade acadêmica: comprovou-se que LLMs de código aberto com tamanho moderado,

próximo a 8 bilhões de parâmetros, quando adaptados ao domínio da saúde mental por meio de fine-tuning e engenharia de prompt, podem alcançar desempenho competitivo e gerar explicações coerentes, respectivamente. Esse resultado estimula futuras pesquisas a explorarem modelos abertos e acessíveis para aplicação em contextos sensíveis e socialmente relevantes, como o suporte à prevenção do suicídio.

6.1.2 Tecnológicas

Este trabalho apresenta avanços tecnológicos significativos em duas frentes complementares. Primeiramente, no Estudo 1, foi proposto o aprimoramento do modelo de classificação do sistema Boamente por meio da aplicação de técnicas de ASSL, com ênfase na AET. Essa abordagem permitiu a incorporação de dados não rotulados ao processo de treinamento, resultando em melhorias expressivas na acurácia do modelo base, ao mesmo tempo em que manteve a sensibilidade necessária para a detecção de casos críticos de IS. Tal estratégia reforça o potencial do uso de métodos semi-supervisionados em contextos sensíveis, como a saúde mental, ao permitir a expansão eficiente de bases de dados com custos reduzidos de anotação.

Em paralelo, no Estudo 2, foi desenvolvida e avaliada uma nova arquitetura para o Boamente, incorporando LLMs não apenas na tarefa de classificação, mas também na geração de explicações por meio de engenharia de prompt. Essa inovação tornou o sistema mais interpretável, ao fornecer justificativas textuais compreensíveis para cada predição. A integração bem-sucedida de LLMs com módulos explicativos representa um avanço tecnológico relevante, promovendo soluções éticas, auditáveis e eficazes, respectivamente.

Outro ponto de destaque tecnológico é o uso de dados provenientes de redes sociais, com curadoria e pré-processamento cuidadosos, garantindo que os modelos fossem treinados sob condições realistas. A adaptação de IS de código aberto por meio de técnicas de fine-tuning e engenharia de prompt permitiu a especialização dos modelos para o domínio da saúde mental, respeitando suas nuances linguísticas e contextuais. Essa especialização foi essencial para que os modelos atingissem desempenho competitivo e fossem capazes de gerar explicações coerentes e compreensíveis para profissionais da área.

Além disso, a arquitetura desenvolvida foi projetada com potencial de integração futura a aplicações práticas voltadas a PSMs. As estratégias de prompt e os módulos implementados fornecem um arcabouço reutilizável e adaptável para pesquisadores e desenvolvedores que desejam empregar LLMs em contextos sensíveis, como a análise de comportamentos de risco suicida. Assim, este trabalho contribui não apenas com uma solução técnica funcional e avaliável, mas também com diretrizes replicáveis para o desenvolvimento de tecnologias responsáveis, éticas e alinhadas às demandas reais da sociedade.

6.1.3 Sociais

O uso de ASSL neste trabalho representa uma contribuição social significativa ao possibilitar o aprimoramento de modelos de detecção de IS mesmo em cenários com escassez de dados rotulados, realidade comum em domínios sensíveis como a saúde mental. Ao aproveitar textos não rotulados com segurança e eficácia, a abordagem adotada amplia o alcance de soluções baseadas em IA, tornando-as mais acessíveis e viáveis para contextos onde a rotulagem manual é limitada por questões éticas, financeiras ou operacionais. Dessa forma, promove-se a construção de tecnologias mais inclusivas, com maior cobertura e potencial de impacto em populações vulneráveis, contribuindo para a redução de desigualdades no acesso a ferramentas de apoio à saúde mental.

Este trabalho oferece uma contribuição significativa ao propor uma ferramenta baseada em IA para auxiliar na identificação de sinais de IS em textos escritos em PT-BR. A aplicação de LLMs nesse contexto sensível pode apoiar PSMs na triagem e monitoramento de pacientes, ampliando o alcance e a eficácia de estratégias preventivas. Em cenários nos quais o acesso a acompanhamento psicológico é limitado, ferramentas como a desenvolvida neste estudo podem servir como apoio inicial, ajudando a identificar casos que demandam atenção urgente.

Além disso, ao tratar de um tema sensível com responsabilidade técnica e ética, este trabalho contribui para o fortalecimento do debate sobre o papel da IA em problemas de saúde mental. Ao demonstrar que é possível unir precisão e interpretabilidade, a solução proposta representa um passo importante rumo ao uso consciente e empático da tecnologia em prol do bem-estar social.

6.2 Limitações

6.2.1 Estudo 1

A primeira limitação diz respeito ao tamanho do conjunto de dados não rotulado utilizado neste estudo. Embora tenhamos conseguido aplicar técnicas semi-supervisionadas com sucesso, o volume de dados poderia ser ampliado para melhorar ainda mais o desempenho e a generalização dos modelos. Enquanto a coleta original do Boamente foi realizada no *Twitter*, optamos pelo uso do *Reddit* nesta etapa, por ser atualmente uma das poucas alternativas viáveis. A escolha se deu devido à limitação imposta pela API do *Twitter*, que passou a exigir uma conta premium paga, inviabilizando a coleta em larga escala naquela plataforma.

Além disso, o estudo se restringiu à aplicação de um conjunto específico de técnicas semi-supervisionadas. Outras abordagens, como métodos baseados em grafos mais avançados ou técnicas recentes de pseudo-rotulação dinâmica, poderiam ter sido exploradas.

Outro ponto relevante é a ausência de um ajuste fino exaustivo dos hiperparâmetros de cada método testado. A fim de garantir uma comparação justa entre os modelos, optou-se por manter os mesmos hiperparâmetros, o que, embora metodologicamente coerente, pode ter limitado a performance de alguns algoritmos.

6.2.2 Estudo 2

No segundo estudo, voltado tanto à classificação quanto à geração de explicações com LLMs, uma das principais limitações está na abordagem adotada: utilizamos engenharia de prompt para geração das explicações, sem realizar fine-tuning específico dos LLMs para a detecção de IS. Embora os resultados obtidos tenham sido positivos, reconhece-se que a construção e o uso de um conjunto de dados anotado, tanto com rótulos quanto com explicações de alta qualidade, poderiam especializar ainda mais os modelos, potencialmente elevando a acurácia da classificação e a precisão, relevância e contextualização das justificativas geradas.

Além disso, a avaliação qualitativa das explicações foi realizada com base em uma amostra reduzida de participantes. Apesar de cuidadosamente selecionados, os avaliadores representam um universo limitado, tanto em termos de número quanto de diversidade de perfis profissionais e culturais. Essa limitação pode impactar a generalização dos resultados, uma vez que percepções sobre clareza, coerência e utilidade das explicações podem variar significativamente entre profissionais de diferentes áreas, níveis de experiência e contextos de atuação.

6.3 Trabalhos Futuros

Os resultados obtidos neste trabalho abrem diversas possibilidades de aprofundamento e expansão, tanto no campo do ASSL, quanto no uso de LLMs para tarefas de classificação e geração de explicações aplicadas à identificação de IS.

6.3.1 Estudo 1

Para o aprimoramento do primeiro estudo, futuras pesquisas podem focar na ampliação do conjunto de dados não rotulados, de modo a aproveitar melhor o potencial das técnicas semi-supervisionadas. A integração de dados provenientes de múltiplas plataformas sociais, desde que eticamente obtidos, pode aumentar a diversidade e a representatividade das amostras, ampliando a capacidade de generalização dos modelos treinados.

Além disso, seria interessante investigar o uso de algoritmos mais avançados de ASSL, como métodos baseados em pseudo-rotulação dinâmica, ou modelos baseados em grafos que considerem as relações latentes entre exemplos rotulados e não rotulados. Paralelamente, a

realização de uma busca mais exaustiva por hiperparâmetros ideais, utilizando abordagens como *grid search* ou *bayesian optimization*, pode contribuir significativamente para o desempenho final dos modelos.

6.3.2 Estudo 2

No segundo estudo, uma direção promissora envolve o fine-tuning supervisionado de LLMs para a tarefa específica de explicação. A construção de um conjunto de dados anotado com justificativas humanas, preferencialmente validadas por especialistas da área da saúde mental, pode tornar o modelo mais fiel às expectativas clínicas e éticas do domínio.

Outra linha relevante é a aplicação de técnicas de *Reinforcement Learning from Human Feedback* (RLHF) para aprimorar a geração de explicações pelos modelos. Essa abordagem permitiria alinhar as justificativas geradas aos critérios de clareza, sensibilidade e relevância definidos por especialistas, por meio de feedback iterativo durante o treinamento. Além disso, o RLHF pode ajudar a desenvolver modelos de raciocínio mais robustos, capazes de fornecer explicações mais contextualizadas e que reflitam a complexidade dos fenômenos relacionados à ideação suicida, aumentando a confiança e a utilidade prática das respostas geradas. O avanço constante no desenvolvimento de LLMs também abre oportunidades para explorar versões mais recentes e especializadas, como o LLaMA 3, Mistral 7B e 13B, Qwen 3.0 e outros modelos emergentes. Esses modelos aprimorados têm o potencial de proporcionar ganhos expressivos tanto no desempenho das tarefas de classificação quanto na geração de explicações mais coerentes, detalhadas e alinhadas ao contexto clínico, ampliando ainda mais a eficácia do sistema para a detecção e monitoramento da ideação suicida.

Por fim, o estudo 2 já incorpora uma arquitetura modular em que diferentes instâncias de modelos de linguagem atuam de forma cooperativa: um módulo é responsável pela classificação da presença ou ausência de IS, enquanto outro gera explicações empáticas e informativas. Essa estrutura modular tem se mostrado eficaz para aumentar a interpretabilidade e a utilidade do sistema. Futuramente, é possível expandir essa abordagem para arquiteturas multiagente mais complexas, incluindo, por exemplo, um terceiro agente dedicado à validação da coerência e consistência das saídas geradas. Tal evolução pode fortalecer ainda mais a robustez, a confiabilidade e a transparência dos sistemas de IA aplicados ao apoio na saúde mental, promovendo soluções cada vez mais alinhadas às necessidades clínicas e éticas do campo.

6.4 Publicações

Durante o desenvolvimento deste mestrado, um artigo científico foi publicado em um periódico, e outro foi aceito para publicação em uma conferência.

- **1º Estudo - Publicado**: AZEVEDO, João Pedro Cavalcanti; DE OLIVEIRA, Adonias Caetano; TELES, Ariel Soares. Identificação de ideação suicida em textos usando aprendizado semi-supervisionado. Journal of Health Informatics, Brasil, v. 16, n. Especial, 2024. DOI: 10.59681/2175-4411.v16.iEspecial.2024.1321.
- **2º Estudo - Aceito**: AZEVEDO, João Pedro Cavalcanti; OLIVEIRA, A. C.; da Silva e Silva, Francisco José; Coutinho, Luciano Reis; Teles, Ariel Soares. Harnessing Generative LLMs to Detect and Explain Suicidal Ideation in Brazilian Portuguese Texts. In: 38th IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2025), 2025, Madrid. Proceedings of the 38th IEEE International Symposium on Computer-Based Medical Systems, 2025.

Referências

Acuña Caicedo, R. W.; Gómez Soriano, J. M.; Melgar Sasieta, H. A. Bootstrapping semi-supervised annotation method for potential suicidal messages. *Internet Interventions*, v. 28, p. 100519, 2022. ISSN 2214-7829. Citado na página 37.

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018. Citado na página 21.

ADADI, A.; BERRADA, M. Explainable ai for healthcare: From black box to interpretable models. In: BHATEJA, V.; SATAPATHY, S. C.; SATORI, H. (Ed.). *Embedded Systems and Artificial Intelligence*. Singapore: Springer Singapore, 2020. p. 327–337. ISBN 978-981-15-0947-6. Citado 2 vezes nas páginas 18 e 21.

ALADAĞ, A. E. *et al.* Detecting suicidal ideation on forums: Proof-of-concept study. *J Med Internet Res*, v. 20, n. 6, p. e215, Jun 2018. ISSN 1438-8871. Citado na página 25.

ALANAZI, S. A. *et al.* Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques. *International Journal of Environmental Research and Public Health*, v. 19, n. 15, 2022. ISSN 1660-4601. Citado na página 27.

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*, v. 19, n. 19, 2022. ISSN 1660-4601. Citado na página 19.

AMINI, M.-R. *et al. Self-Training: A Survey*. 2023. Citado 2 vezes nas páginas 29 e 30.

ARENDT, F. *et al.* Chatgpt, artificial intelligence, and suicide prevention. *Crisis*, v. 44, n. 5, p. 367–370, 2023. PMID: 37737578. Citado na página 19.

AZEVEDO, J. P. C.; OLIVEIRA, A. C. de; TELES, A. S. Identificação de ideação suicida em textos usando aprendizado semi-supervisionado. *Journal of Health Informatics*, v. 16, n. Especial, nov. 2024. Citado na página 40.

BALDACARA, L. *et al.* Epidemiology of suicides in brazil: a systematic review. *GLOBAL PSYCHIATRY ARCHIVES*, v. 5, n. 1, p. 10–25, 2022. ISSN 2754-9380. Citado na página 24.

BARUA, P. D. *et al.* Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: a review. *Cognitive Neurodynamics*, v. 18, n. 1, p. 1–22, Feb 2024. ISSN 1871-4099. Citado 3 vezes nas páginas 17, 18 e 20.

BAUER, B. *et al.* Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: Linguistic analysis of reddit posts. *JMIR Ment Health*, v. 11, p. e57234, May 2024. ISSN 2368-7959. Citado na página 38.

BERNERT, R. A. *et al.* Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health*, v. 17, n. 16, 2020. ISSN 1660-4601. Citado 2 vezes nas páginas 18 e 25.

BERTOLOTE, J. M. *O suicídio e sua prevenção*. São Paulo: Editora Unesp, 2016. Citado 4 vezes nas páginas 15, 16, 18 e 24.

BERTOLOTE, J. M.; MELLO-SANTOS, C. d.; BOTEGA, N. J. Detecção do risco de suicídio nos serviços de emergência psiquiátrica. *Brazilian Journal of Psychiatry*, Associação Brasileira de Psiquiatria, v. 32, p. S87–S95, Oct 2010. ISSN 1516-4446. Citado na página 15.

BIMPAS, A. *et al.* Leveraging pervasive computing for ambient intelligence: A survey on recent advancements, applications and open challenges. *Computer Networks*, v. 239, p. 110156, 2024. ISSN 1389-1286. Citado na página 26.

BOONIAM, S. *et al.* Predictors of passive and active suicidal ideation and suicide attempt among older people: A study in tertiary care settings in thailand. *Neuropsychiatric Disease and Treatment*, Dove Medical Press, v. 16, p. 3135–3144, 2020. Citado 2 vezes nas páginas 16 e 25.

BOTEGA, N. J. Comportamento suicida: epidemiologia. *Psicologia USP*, Instituto de Psicologia da Universidade de São Paulo, v. 25, n. 3, p. 231–236, Sep 2014. ISSN 0103-6564. Citado na página 25.

BOTEGA, N. J. *et al.* Suicidal behavior in the community: prevalence and factors associated with suicidal ideation. *Brazilian Journal of Psychiatry*, Associação Brasileira de Psiquiatria, v. 27, n. 1, p. 45–53, Mar 2005. ISSN 1516-4446. Citado na página 24.

BRACISZEWSKI, J. M. Digital technology for suicide prevention. *Advances in Psychiatry and Behavioral Health*, v. 1, n. 1, p. 53–65, 2021. ISSN 2667-3827. Advances in Psychiatry and Behavioral Health. Citado 3 vezes nas páginas 17, 18 e 20.

BREZO, J.; PARIS, J.; TURECKI, G. Personality traits as correlates of suicidal ideation, suicide attempts, and suicide completions: a systematic review. *Acta Psychiatrica Scandinavica*, v. 113, n. 3, p. 180–206, 2006. Citado na página 25.

BUFANO, P. *et al.* Digital phenotyping for monitoring mental disorders: Systematic review. *J Med Internet Res*, v. 25, p. e46778, Dec 2023. ISSN 1438-8871. Citado na página 25.

CALVO, R. *et al.* Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, p. 1–37, 01 2017. Citado na página 27.

CASSORLA, R. M. S. *Estudos sobre suicídio: psicanálise e saúde mental*. [S.l.]: Editora Blucher, 2021. Citado 3 vezes nas páginas 15, 18 e 24.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. *Semi-Supervised Learning*. [S.l.]: MIT Press, 2009. Citado na página 29.

CHEN, X. *et al.* Weighted samples based semi-supervised classification. *Applied Soft Computing*, v. 79, p. 46–58, 2019. ISSN 1568-4946. Citado na página 32.

CHEN, Y. *et al. Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data*. 2023. Citado na página 31.

CHENG, S.-W. *et al.* The now and future of chatgpt and gpt in psychiatry. *Psychiatry and Clinical Neurosciences*, v. 77, n. 11, p. 592–596, 2023. Citado na página 19.

COPPERSMITH, D. D. *et al.* Just-in-time adaptive interventions for suicide prevention: Promise, challenges, and future directions. *Psychiatry*, Routledge, v. 85, n. 4, p. 317–333, 2022. Citado na página 27.

CZYZ, E. K. *Prediction of Suicidal Behavior Risk among Adolescents Seen in Psychiatric Settings.* Tese (Doutorado) — niversity of Michigan, 2015. Doctor of Philosophy (Psychology) . Citado na página 20.

DAS, B. C.; AMINI, M. H.; WU, Y. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 57, n. 6, fev. 2025. ISSN 0360-0300. Citado na página 35.

de Oliveira, A. C. *et al.* How can machine learning identify suicidal ideation from user's texts? towards the explanation of the boamente system. *Procedia Computer Science*, v. 206, p. 141–150, 2022. ISSN 1877-0509. International Society for Research on Internet Interventions 11th Scientific Meeting. Citado na página 40.

DEFLEUR, M. L. Stigma: Notes on the management of spoiled identity. *Social Forces*, Oxford University Press, v. 43, n. 1, p. 127–128, 1964. ISSN 00377732, 15347605. Citado na página 17.

DETTMERS, T. *et al.* Qlora: efficient finetuning of quantized llms. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS '23). Citado na página 49.

DINIZ, E. J. S. *et al.* Boamente: A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. *Healthcare*, v. 10, n. 4, 2022. ISSN 2227-9032. Citado 9 vezes nas páginas x, 17, 20, 21, 28, 40, 48, 49 e 56.

DORAN, C. M.; KINCHIN, I. A review of the economic impact of mental illness. *Australian Health Review*, v. 43, n. 1, p. 43–48, 2019. Citado na página 15.

DUGAS, E. *et al.* Early predictors of suicidal ideation in young adults. *The Canadian Journal of Psychiatry*, v. 57, n. 7, p. 429–436, 2012. PMID: 22762298. Citado na página 19.

D'ALFONSO, S. Ai in mental health. *Current Opinion in Psychology*, v. 36, p. 112–117, 2020. ISSN 2352-250X. Cyberpsychology. Citado na página 26.

EKMAN, P. An argument for basic emotions. *Cognition & Emotion*, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992. Citado na página 27.

ELYOSEPH, Z.; LEVKOVICH, I. Beyond human expertise: the promise and limitations of chatgpt in suicide risk assessment. *Frontiers in Psychiatry*, v. 14, 2023. ISSN 1664-0640. Citado na página 19.

FELBO, B. *et al.* Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2017. p. 1615–1625. Citado na página 27.

FU, W. *et al.* Psychological health, sleep quality, and coping styles to stress facing the covid-19 in wuhan, china. *Translational Psychiatry*, v. 10, n. 1, p. 225, Jul 2020. ISSN 2158-3188. Citado na página 15.

GELDER, M. G.; LOPEZ-IBOR, J. J.; ANDREASEN, N. C. *New Oxford Textbook of Psychiatry, Volume 2*. [S.l.]: Oxford university press, 2003. Citado na página 25.

GHANADIAN, H.; NEJADGHOLI, I.; OSMAN, H. A. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*, v. 12, p. 14350–14363, 2024. Citado na página 38.

GU, Y. *et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. 2020. Citado na página 37.

GUO, T. *et al. Large Language Model based Multi-Agents: A Survey of Progress and Challenges*. 2024. Citado na página 35.

GUO, Z. *et al.* Large language models for mental health applications: Systematic review. *JMIR Ment Health*, v. 11, p. e57400, Oct 2024. ISSN 2368-7959. Citado na página 38.

HAN, M. H. D.; TEAM, U. *Unsloth*. 2023. Citado na página 49.

HANDLEY, T. *et al.* The challenges of predicting suicidal thoughts and behaviours in a sample of rural australians with depression. *International Journal of Environmental Research and Public Health*, v. 15, n. 5, 2018. ISSN 1660-4601. Citado 2 vezes nas páginas 15 e 18.

HAQUE, R. *et al.* A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. *Technologies*, v. 10, n. 3, 2022. ISSN 2227-7080. Citado na página 19.

HARKAVY-FRIEDMAN, J. M. Can early detection of psychosis prevent suicidal behavior? *American Journal of Psychiatry*, v. 163, n. 5, p. 768–770, 2006. Citado na página 19.

HARMER, B. *et al.* Suicidal ideation. 2020. Citado 2 vezes nas páginas 16 e 25.

HE, K. *et al. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics*. 2023. Citado 2 vezes nas páginas 18 e 19.

HEINZ, M. V. *et al.* Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, v. 2, n. 4, p. AIoa2400802, 2025. Citado na página 33.

HOLMES, G. *et al.* Applications of large language models in the field of suicide prevention: Scoping review. *Journal of Medical Internet Research*, v. 27, 2025. ISSN 1438-8871. Citado na página 38.

INSEL, T. R. The nimh research domain criteria (rdoc) project: Precision medicine for psychiatry. *American Journal of Psychiatry*, v. 171, n. 4, p. 395–397, 2014. PMID: 24687194. Citado na página 17.

INSEL, T. R. Digital phenotyping: Technology for a new science of behavior. *JAMA*, v. 318, n. 13, p. 1215–1216, 10 2017. ISSN 0098-7484. Citado na página 26.

ISCEN, A. *et al. Label Propagation for Deep Semi-supervised Learning*. 2019. Citado na página 32.

JAIN, S. H. *et al.* The digital phenotype. *Nature Biotechnology*, v. 33, n. 5, p. 462–463, May 2015. ISSN 1546-1696. Citado na página 26.

JI, S. *et al.* Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, v. 8, n. 1, p. 214–226, 2021. Citado 2 vezes nas páginas 17 e 25.

KALYAN, K. S. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, v. 6, p. 100048, 2024. ISSN 2949-7191. Citado na página 19.

KEMP, S. *Digital 2020: 3.8 Billion People Use Social Media*. 2020. Disponível em: https://wearesocial.com/uk/blog/2020/01/digital-2020-3-8-billion-people-use-social-media/. Acessado em: 16 de maio de 2024. Citado na página 19.

KINGMA, D. P. *et al. Semi-Supervised Learning with Deep Generative Models*. 2014. Citado na página 29.

KUMAR, A. *et al.* Overview of ai based e- healthcare system. *EPRA International Journal of Multidisciplinary Research (IJMR)*, v. 9, n. 6, p. 182–186, Jun. 2023. Citado na página 18.

LANG, H. *et al. Co-training Improves Prompt-based Learning for Large Language Models*. 2022. Citado na página 31.

LEVKOVICH, I.; ELYOSEPH, Z. Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: Vignette study. *JMIR Ment Health*, v. 10, p. e51232, Sep 2023. ISSN 2368-7959. Citado 2 vezes nas páginas 18 e 19.

LEãO, A.; LUSSI, I. A. d. O. Estigmatização: consequências e possibilidades de enfrentamento em centros de convivência e cooperativas. *Interface - Comunicação, Saúde, Educação*, UNESP, v. 25, p. e200474, 2021. ISSN 1414-3283. Citado na página 17.

LI, L. *et al. A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs)*. 2024. Citado na página 33.

LIN, H. *et al.* User-level psychological stress detection from social media using deep neural network. In: ACM. *Proceedings of the 22nd ACM international conference on Multimedia*. [S.l.], 2014. p. 507–516. Citado na página 27.

LIU, R. T.; MILLER, I. Life events and suicidal ideation and behavior: A systematic review. *Clinical Psychology Review*, v. 34, n. 3, p. 181–192, 2014. ISSN 0272-7358. Citado na página 20.

LIU, Y. *et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. Citado na página 36.

LOVITT, M. *et al. Suicide Risk Assessment on Social Media with Semi-Supervised Learning*. 2024. Citado na página 37.

MAY, A. M.; KLONSKY, E. D. What distinguishes suicide attempters from suicide ideators? a meta-analysis of potential factors. *Clinical Psychology: Science and Practice*, Wiley-Blackwell Publishing Ltd., United Kingdom, v. 23, n. 1, p. 5–20, 2016. Citado na página 16.

MELCHER, J.; HAYS, R.; TOROUS, J. Digital phenotyping for mental health of college students: a clinical review. *BMJ Ment Health*, Royal College of Psychiatrists, v. 23, n. 4, p. 161–166, 2020. ISSN 1362-0347. Citado na página 16.

MENDES, J. P. M. *et al.* Sensing apps and public data sets for digital phenotyping of mental health: Systematic review. *J Med Internet Res*, v. 24, n. 2, p. e28735, Feb 2022. ISSN 1438-8871. Citado 2 vezes nas páginas 16 e 26.

MOHAMMAD, S. M.; TURNEY, P. D. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, Wiley Online Library, v. 29, n. 3, p. 436–465, 2013. Citado na página 27.

MOHR, D. C.; SHILTON, K.; HOTOPF, M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *npj Digital Medicine*, v. 3, n. 1, p. 45, Mar 2020. ISSN 2398-6352. Citado na página 16.

MONSELISE, M.; YANG, C. C. "i'm always in so much pain and no one will understand" - detecting patterns in suicidal ideation on reddit. In: *Companion Proceedings of the Web Conference 2022*. New York, NY, USA: Association for Computing Machinery, 2022. (WWW '22), p. 686–691. ISBN 9781450391306. Citado na página 25.

MOURA, I. *et al.* Towards identifying context-enriched multimodal behavioral patterns for digital phenotyping of human behaviors. *Future Generation Computer Systems*, v. 131, p. 227–239, 2022. ISSN 0167-739X. Citado na página 26.

MOURA, I. *et al.* Mental health ubiquitous monitoring supported by social situation awareness: A systematic review. *Journal of Biomedical Informatics*, v. 107, p. 103454, 2020. ISSN 1532-0464. Citado na página 16.

MUSLEA, I.; MINTON, S.; KNOBLOCK, C. A. Active+ semi-supervised learning= robust multi-view learning. In: CITESEER. *ICML*. [S.l.], 2002. v. 2, p. 435–442. Citado na página 29.

NEELEMAN, J.; de Graaf, R.; VOLLEBERGH, W. The suicidal process; prospective comparison between early and later stages. *Journal of Affective Disorders*, v. 82, n. 1, p. 43–52, 2004. ISSN 0165-0327. Citado na página 25.

NUTTING, P. A. *et al.* Improving detection of suicidal ideation among depressed patients in primary care. *The Annals of Family Medicine*, The Annals of Family Medicine, v. 3, n. 6, p. 529–536, 2005. ISSN 1544-1709. Citado na página 19.

O'CONNOR, R. C.; KIRTLEY, O. J. The integrated motivational–volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, The Royal Society, v. 373, n. 1754, p. 20170268, 2018. Citado na página 15.

O'CONNOR, R. C.; NOCK, M. K. The psychology of suicidal behaviour. *The Lancet Psychiatry*, v. 1, n. 1, p. 73–85, 2014. ISSN 2215-0366. Citado na página 18.

OLIVEIRA, A. C. d.; BESSA, R. F.; TELES, A. S. Comparative analysis of bert-based and generative large language models for detecting suicidal ideation: a performance evaluation study. *Cadernos de Saúde Pública*, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 40, n. 10, p. e00028824, 2024. ISSN 0102-311X. Citado 3 vezes nas páginas x, 40 e 56.

OLIVEIRA, A. C. de *et al.* Effect of explainable artificial intelligence on trust of mental health professionals in an ai-based system for suicide prevention. *IEEE Access*, v. 13, p. 60987–61005, 2025. Citado na página 40.

OMS, O. M. d. S. *Suicide worldwide in 2019: global health estimates*. [S.l.]: World Health Organization, 2021. World Health Organization. Citado na página 24.

ORRI, M. *et al.* Pathways of association between childhood irritability and adolescent suicidality. *Journal of the American Academy of Child & Adolescent Psychiatry*, v. 58, n. 1, p. 99–107.e3, 2019. ISSN 0890-8567. Citado na página 25.

PICARDO, J. *et al.* Suicide and self-harm content on instagram: A systematic scoping review. *PLOS ONE*, Public Library of Science, v. 15, n. 9, p. 1–16, 09 2020. Citado na página 25.

PLUTCHIK, R. A general psychoevolutionary theory of emotion. In: PLUTCHIK, R.; KELLERMAN, H. (Ed.). *Emotion: Theory, research, and experience*. New York: Academic Press, 1980. v. 1, p. 3–33. Citado na página 27.

PRADHAN, D.; VARSHNEY, A.; SINGH, S. A critical review on challenges and limitations in artificial intelligence-based e-health applications. v. 5, p. 001–005, 05 2023. Citado 2 vezes nas páginas 18 e 20.

QI, H. *et al. Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media*. 2023. Citado 2 vezes nas páginas 18 e 19.

QORICH, M.; OUAZZANI, R. E. Advanced deep learning and large language models for suicide ideation detection on social media. *Progress in Artificial Intelligence*, v. 13, n. 2, p. 135–147, Jun 2024. ISSN 2192-6360. Citado na página 38.

RABANI, S. T.; KHAN, Q. R.; KHANDAY, A. M. U. D. Detection of suicidal ideation on twitter using machine learning & ensemble approaches. *Baghdad Science Journal*, v. 17, n. 4, p. 1328, Dec. 2020. Citado na página 20.

RAFFEL, C. *et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. Citado na página 34.

RAO, V. M. *et al.* Multimodal generative ai for medical image interpretation. *Nature*, v. 639, n. 8056, p. 888–896, Mar 2025. ISSN 1476-4687. Citado na página 35.

RENAUD, J. *et al.* Suicidal ideation and behavior in youth in low- and middle-income countries: A brief review of risk factors and implications for prevention. *Frontiers in Psychiatry*, v. 13, 2022. ISSN 1664-0640. Citado 2 vezes nas páginas 15 e 17.

RIBEIRO, J. D. *et al.* Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. *British Journal of Psychiatry*, v. 212, n. 5, p. 279–286, 2018. Citado na página 16.

RIERA-SERRA, P. *et al.* Clinical predictors of suicidal ideation, suicide attempts and suicide death in depressive disorder: a systematic review and meta-analysis. *European Archives of Psychiatry and Clinical Neuroscience*, v. 274, n. 7, p. 1543–1563, Oct 2024. ISSN 1433-8491. Citado na página 24.

SADRI, F. Ambient intelligence: A survey. Association for Computing Machinery, New York, NY, USA, v. 43, n. 4, out. 2011. ISSN 0360-0300. Citado na página 26.

SAI, S. *et al.* Generative ai for transformative healthcare: A comprehensive study of emerging models, applications, case studies, and limitations. *IEEE Access*, v. 12, p. 31078–31106, 2024. Citado na página 38.

SAWHNEY, R. *et al.* Towards emotion- and time-aware classification of tweets to assist human moderation for suicide prevention. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 15, n. 1, p. 609–620, May 2021. Citado 2 vezes nas páginas 20 e 21.

SCHERER, S.; PESTIAN, J.; MORENCY, L.-P. Investigating the speech characteristics of suicidal adolescents. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2013. p. 709–713. Citado na página 18.

SHAO, M. *et al.* Survey of different large language model architectures: Trends, benchmarks, and challenges. *IEEE Access*, v. 12, p. 188664–188706, 2024. Citado na página 35.

SHI, W. *et al.* Transductive semi-supervised deep learning using min-max features. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. Citado na página 29.

SHOIB, S. *et al.* Facebook and suicidal behaviour: User experiences of suicide notes, live-streaming, grieving and preventive strategies—a scoping review. *International Journal of Environmental Research and Public Health*, v. 19, n. 20, 2022. ISSN 1660-4601. Citado na página 25.

SIKANDER, D. *et al.* Predicting risk of suicide using resting state heart rate. In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. [S.l.: s.n.], 2016. p. 1–4. Citado na página 18.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado na página 42.

SQUIRES, M. *et al. Enhancing Suicide Risk Detection on Social Media through Semi-Supervised Deep Label Smoothing*. 2024. Citado na página 36.

STERN, W. *et al.* Natural language explanations for suicide risk classification using large language models. In: *Proceedings of Machine Learning for Cognitive and Mental Health Workshop (ML4CMH 2024)*. [S.l.: s.n.], 2024. Citado na página 39.

TANK, C. *et al. Su-RoBERTa: A Semi-supervised Approach to Predicting Suicide Risk through Social Media using Base Language Models*. 2024. Citado na página 36.

TELES, A. S. *et al. Dataset of suicidal ideation texts in Brazilian Portuguese - Boamente System*. [S.l.]: Zenodo, 2023. 698 p. Citado na página 49.

TELES, A. S. *et al.* Enriching mental health mobile assessment and intervention with situation awareness. *Sensors*, v. 17, n. 1, 2017. ISSN 1424-8220. Citado 2 vezes nas páginas 20 e 21.

TOROUS, J. *et al.* New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Ment Health*, Canada, v. 3, n. 2, p. e16, maio 2016. Citado 2 vezes nas páginas 16 e 26.

TURECKI, G.; BRENT, D. A. Suicide and suicidal behaviour. *Lancet (London, England)*, v. 387, n. 10024, p. 1227–1239, 2016. Citado 2 vezes nas páginas 19 e 24.

VAIDYAM, A.; HALAMKA, J.; TOROUS, J. Actionable digital phenotyping: a framework for the delivery of just-in-time and longitudinal interventions in clinical healthcare. *Mhealth*, China, v. 5, p. 25, ago. 2019. Citado 2 vezes nas páginas 20 e 21.

VASWANI, A. *et al. Attention Is All You Need*. 2023. Citado 2 vezes nas páginas 33 e 34.

WAGNER, J. *et al.* The brwac corpus: A new open resource for brazilian portuguese. In: . [S.l.: s.n.], 2018. Citado na página 43.

WASTLER, H. M. *et al.* An empirical investigation of the distinction between passive and active ideation: Understanding the latent structure of suicidal thought content. *Suicide and Life-Threatening Behavior*, v. 53, n. 2, p. 219–226, 2023. Citado na página 16.

WEISS, S. J. *et al.* Potential paths to suicidal ideation and suicide attempts among high-risk women. *Journal of Psychiatric Research*, Elsevier Science, Netherlands, v. 155, p. 493–500, 2022. Citado 3 vezes nas páginas 15, 16 e 19.

WONGKOBLAP, A.; VADILLO, M. A.; CURCIN, V. Researching mental health disorders in the era of social media: Systematic review. *J Med Internet Res*, v. 19, n. 6, p. e228, Jun 2017. ISSN 1438-8871. Citado na página 25.

World Health Organization. *World Mental Health Report: Transforming Mental Health for All*. [S.l.]: World Health Organization, 2022. Citado na página 25.

YANG, J. *et al.* Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, Association for Computing Machinery, New York, NY, USA, feb 2024. ISSN 1556-4681. Just Accepted. Citado na página 34.

YANG, R. *et al.* Large language models in health care: Development, applications, and challenges. *Health Care Science*, v. 2, n. 4, p. 255–263, 2023. Citado na página 33.

ZHANG, T. *et al.* Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, v. 5, n. 1, p. 46, Apr 2022. ISSN 2398-6352. Citado na página 19.

ZHAO, H. *et al.* Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, Association for Computing Machinery, New York, NY, USA, v. 15, n. 2, feb 2024. ISSN 2157-6904. Citado na página 19.

ZHAO, W. X. *et al. A Survey of Large Language Models*. 2023. Citado na página 19.

ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. 1. ed. [S.l.]: Springer Cham, 2022. ISSN 1939-4608. Citado na página 29.

ZOU, W.; TANG, L.; BIE, B. The stigmatization of suicide: A study of stories told by college students in china. *Death studies*, v. 46, n. 9, p. 2035—2045, 2022. ISSN 0748-1187. Citado na página 17.