CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Identificação de ideação suicida em textos usando aprendizado semi-supervisionado

Identifying suicidal ideation in texts using semi-supervised learning

Identificación de ideas suicidas en textos mediante aprendizaje semisupervisado

João Pedro Cavalcanti Azevedo[1], Adonias Caetano de Oliveira[2] , Ariel Soares Teles[3]

1 Mestrando em Ciência da Computação, Programa de Pós-graduação em Ciência da Computação, Universidade Federal do Maranhão, São Luís (MA), Brasil.
2 Doutorando em Biotecnologia, Programa de Pós-graduação em Biotecnologia, Universidade Federal do Delta do Parnaíba, Paranaíba (PI), Brasil.
3 Doutor em Engenharia Elétrica, Instituto Federal do Maranhão, Araioses (MA), Brasil.

Autor correspondente: João Pedro Cavalcanti Azevedo
E-mail: joaopedro.azevedo@lsdi.ufma.br
Link: https://doi.org/10.5281/zenodo.10070747

Resumo

Objetivo: Aprimorar o modelo *Boamente* usando métodos de aprendizado semi-supervisionado para a identificação de ideação suicida em textos não clínicos escritos em português brasileiro, a fim de melhorar o seu desempenho. Método: Foi realizada a coleta de novos dados e a aplicação de diferentes métodos de aprendizado semi-supervisionado com ênfase em análise de emoções para aprimorar o modelo existente. Resultados: Os resultados demostraram uma evolução entre 2,39% e 4,30% na métrica de acurácia em relação ao modelo original, com o método *self-learning* alcançando o melhor desempenho. Conclusão: A aplicação de métodos de aprendizado semi-supervisionado propiciou a melhoria no desempenho do modelo *Boamente* para a identificação de ideação suicida. Esse estudo então contribui para o desenvolvimento de uma ferramenta mais eficaz para os profissionais de saúde mental na prevenção ao suicídio, auxiliado-os em tomadas de decisão mais assertivas no monitoramento de seus pacientes.

Descritores: Saúde Mental; Ideação Suicida; Análise de Emoções;

Abstract

Objective: To improve the *Boamente* model using semi-supervised learning methods for the identification of suicidal ideation in non-clinical texts written in Brazilian Portuguese, in order to improve its performance. Method: New data was collected and different semi-supervised learning methods with an emphasis on emotion analysis were applied to improve the existing model. Results: The results showed an improvement of between 2.39% and 4.30% in the accuracy metric compared to the original model, with the self-learning method achieving the best performance. Conclusion: The application of semi-supervised learning methods improved the performance of the *Boamente* model for identifying suicidal ideation. This study therefore contributes to the development of a more effective tool for mental health professionals in suicide prevention, helping them to make more assertive decisions when monitoring their patients.

Keywords: Mental Health; Suicidal Ideation; Emotion Analysis;

Resumen

Objetivo: Mejorar el desempeño del modelo *Boamente* utilizando métodos de aprendizaje semi-supervisado para la identificación de ideación suicida en textos no clínicos escritos en portugués brasileño. Método: Se recolectaron nuevos datos y se aplicaron diferentes métodos de aprendizaje semisupervisado con énfasis en el análisis de emociones para mejorar el modelo existente. Resultados: Los resultados mostraron una mejora de entre el 2,39% y el 4,30% en la métrica de precisión en comparación con el modelo original, siendo el método de aprendizaje autosupervisado el que obtuvo el mejor rendimiento. Conclusión: La aplicación de métodos de aprendizaje semisupervisado mejoró el rendimiento del modelo *Boamente* para identificar la ideación suicida. Este estudio contribuye, por tanto, al desarrollo de una herramienta más eficaz para los profesionales de la salud mental en la prevención del suicidio, ayudándoles a tomar decisiones más asertivas en el seguimiento de sus pacientes.

Descriptores: Salud Mental; Ideación Suicida; Análisis de Emociones;

Introdução

Na era em que a Inteligência Artificial (IA) e as mídias sociais proliferam, é crucial tirar vantagens dessas tecnologias para lidar com a crescente prevalência de problemas relacionados à saúde mental, incluindo a incidência significativa de suicídios. Anualmente, 703.000 vidas são perdidas para o suicídio no mundo todo, e um número ainda maior de indivíduos tenta esse ato desesperado[1]. Cada caso de suicídio representa uma tragédia impactante que reverbera em famílias, comunidades e até em escala nacional. Em 2019, o suicídio foi classificado como a quarta principal causa de morte entre jovens de 15 a 29 anos em todo o mundo[1].

A pesquisa em soluções digitais voltadas para a saúde mental, especialmente no que diz respeito à ideação suicida, torna-se imperativa diante dos desafios crescentes associados à saúde mental e à prevenção ao suicídio[2]. Com o aumento das taxas de transtornos mentais e casos de ideação suicida, as abordagens convencionais de intervenção podem ser insuficientes para lidar com a escala e a complexidade desses problemas. Neste cenário, as tecnologias digitais oferecem uma oportunidade inovadora para desenvolver intervenções acessíveis de acompanhamento e monitoramento remoto personalizadas e eficazes.

Formas de prevenção ao suicídio envolvem uma variedade de estratégias destinadas a identificar, monitorar e intervir em situações de risco. O monitoramento é uma meio fundamental e os métodos de monitoramento visam acompanhar indicadores de ideação suicida. Com avanços nessa área, o conceito de *just-in-time intervention*[5] (intervenção no momento certo) ganhou destaque. Esse método envolve o uso de algoritmos e ferramentas de análise para monitorar continuamente sinais comportamentais e emocionais, permitindo intervenções precisas e personalizadas, quando necessário. O objetivo é oferecer suporte no exato momento em que uma pessoa está precisando, maximizando a eficácia da intervenção.

O *Boamente*[6] é uma ferramenta desenvolvida para abordar o desafio de monitorar a ideação suicida em pessoas em risco. A ferramenta consiste em um aplicativo de teclado virtual para dispositivos móveis, que coleta passivamente dados textuais dos usuários. Esses dados são enviados para uma plataforma web, onde são processados usando técnicas de Processamento de Linguagem Natural (PLN) e um modelo de aprendizado profundo. Ao analisar os dados textuais dos usuários, o

Boamente é baseado na fenotipagem digital para identificar padrões e indicadores linguísticos associados à ideação suicida. Como definido pelos autores Torous et al.[7], fenotipagem digital é a "quantificação momento a momento do fenótipo humano em nível individual *in-situ* usando dados de smartphones e outros dispositivos digitais pessoais". A abordagem do *Boamente* permite uma detecção precoce de comportamentos de risco e a possibilidade para a realização de *just-in-time interventions* feitas por profissionais, contribuindo para a prevenção do suicídio.

Recentemente, houve avanços significativos no processamento de texto impulsionados pela IA[3]. No entanto, muitas tarefas de PLN demandam grandes conjuntos de dados rotulados para treinar modelos com desempenho satisfatório. A obtenção desses conjuntos de dados rotulados e de alta qualidade pode ser cara e trabalhosa, limitando a escalabilidade e a aplicabilidade dos modelos em cenários reais. Para enfrentar esse desafio, métodos semi-supervisionados surgiram como uma alternativa promissora, permitindo que modelos se beneficiem de dados não rotulados em conjunto com uma quantidade limitada de dados rotulados. O aprendizado semi-supervisionado representa uma abordagem intermediária entre o aprendizado supervisionado (que requer muitos dados rotulados) e o aprendizado não supervisionado (que depende exclusivamente de dados não rotulados)[4].

Nesse contexto, concentramos nossa atenção nos métodos de aprendizado semi-supervisionado aplicados ao processamento de texto. O objetivo deste trabalho foi aprimorar o modelo de aprendizado profundo do Boamente, a fim de obter um melhor desempenho e, como consequência, torná-lo mais confiável e eficiente para auxiliar na prevenção do suicídio.

Métodos

Nesse estudo, foram realizados as seguintes etapas presentes na Figura 1 e descritas nas próximas seções: coleta e preparação de dados, treinamento dos modelos e avaliação de desempenho. Usamos a linguagem de programação *Python*, e as bibliotecas *Pandas*, *NumPy*, *PyTorch* e *Matplotlib*, com o *Google Colaboratory* como plataforma de desenvolvimento. No ambiente do *Google Colaboratory*, as configurações de hardware da máquina virtual foram: processador *Intel(R) Xeon(R)*, 65GB de memória *RAM*, e uma *GPU NVIDIA L4* com 22,5GB de memória *RAM*.

Figura 1 - Etapas da Metodologia.

Coleta e Preparação de Dados

O conjunto inicial de dados é o original do *Boamente* (ver Material Suplementar no link). Esse conjunto de dados passou por rotulação manual especializada, em que psicólogos analisaram as sentenças e as classificaram contendo ideação suicida ou não. Inicialmente, um total de 5.699 tweets foi coletado. Após a coleta de dados, três psicólogos foram convidados a realizar a anotação dos dados. Para evitar vieses no processo de anotação, foram selecionados psicólogos com diferentes abordagens da área de psicologia: terapia cognitivo-comportamental, teoria psicanalítica, e a teoria humanística. Os profissionais tiveram que classificar cada sentença como positiva (contém ideação suicida) ou negativa (não contém ideação suicida). Sentenças que tiveram discordância entre os profissionais foram excluídas, gerando assim um conjunto de dados rotulados de 3.788 amostras.

Os dados não rotulados foram coletados da rede social *Reddit*. Eles foram extraídos mais especificamente de três *subreddits* com temas relacionados à ideação suicida: *Ansiedade e Depressão*, *Transtornos Mentais* e *Desabafos*. Foi utilizada a biblioteca *Praw* para acessar a *API* do *Reddit* e efetuar a coleta dos dados. A coleta gerou um total de 11.240 novas amostras, representando quase quatro vezes o tamanho do conjunto de dados original. Esses dados passaram por uma limpeza, em que links, emojis e carácteres especiais foram removidos.

No conjunto de dados original, a coleta foi feita usando palavras-chave específicas[6]. Já na coleta que fizemos no *Reddit*, os dados foram coletados sem filtro, ou seja, sem qualquer restrição pela presença de palavras-chave ou termos específicos. A Figura 2 exibe a nuvem de palavras do conjunto de dados original do *Boamente.* A grande presença das palavras de cunho negativo (e.g., "matar") é devido à coleta desse conjunto de dados ter sido feita considerando palavras-chave[6], tais como "me matar", "quero morrer, "estar morto", dentre outras. A Figura 3 mostra a

nuvem de palavras da coleta do *Reddit*, a qual não realizou filtros por palavras-chave no conteúdo das postagens. Ao analisar as duas figuras, é possível observar que as diferenças de termos é significativa entre os dois conjuntos de dados.

Figura 2 - Nuvem de palavras do conjunto de dados do *Boamente*.

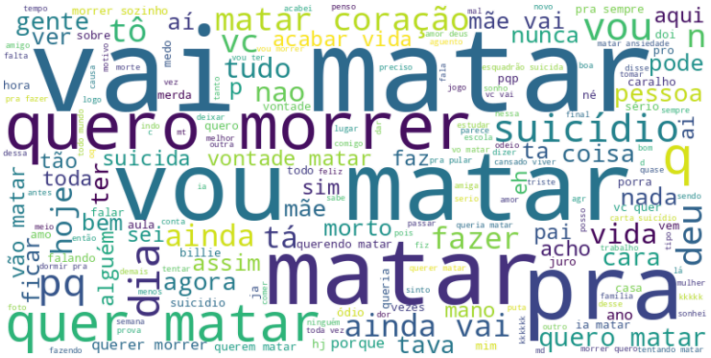


Figura 3 - Nuvem de palavras do conjunto de dados não rotulado.

Treinamento dos Modelos com Métodos Semi-supervisionados

O modelo pré-treinado utilizado neste estudo para todos os métodos semi-supervisionados foi o *BERTimbau Large*[13], o qual obteve o melhor desempenho no estudo inicial do *Boamente*[6]. O modelo *BERTimbau*[13] foi treinado em uma vasta quantidade de texto em português pelo corpus *brWaC*[14], 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*, os quais são 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 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 aprendizado semi-supervisionado: *self-learning*, *co-training*, *label propagation*, *weighting* e *boosting*. O self-learning[8] é um método em que um modelo é inicialmente treinado com um conjunto de dados rotulados. Após esse treinamento inicial, o modelo é usado para fazer predições sobre dados não rotulados, gerando pseudo-rótulos para esses dados com base em suas próprias predições. Esses pseudo-rótulos, as predições do modelo que têm alta confiança, são então incorporados ao conjunto de dados rotulados. Em seguida, o modelo é re-treinado com o conjunto expandido de dados rotulados, que agora inclui os exemplos inicialmente não rotulados com seus pseudo-rótulos. O processo é iterativo, normalmente envolvendo múltiplas rodadas de retreinamento e geração de pseudo-rótulos. A cada iteração, o modelo utiliza os pseudo-rótulos dos dados não rotulados para aprender mais e aprimorar sua capacidade de generalização. A ideia é que, com cada iteração adicional, o modelo se torne mais competente em suas predições, aumentando a confiança dos pseudo-rótulos. Em certos casos, uma única iteração pode ser suficiente para melhorar significativamente a capacidade do modelo de generalizar a partir dos dados não rotulados disponíveis.

O método *co-training*[9] é uma abordagem de aprendizado semi-supervisionado que se baseia na interação entre dois modelos de aprendizado de máquina, cada um treinado em uma perspectiva diferente dos dados. Inicialmente, os modelos são treinados com um conjunto limitado de dados rotulados e, em seguida, são utilizados para rotular um conjunto maior de dados não rotulados. Os exemplos rotulados com alta confiança, em limiares estabelecidos, são então trocados entre os modelos, permitindo que cada um aprenda com os exemplos rotulados pelo outro. Embora tipicamente envolva múltiplas iterações de rotulagem, troca e re-treinamento, é crucial salientar que o co-training pode ser executado com uma única iteração, dependendo das demandas e recursos disponíveis.

O método *boosting*[10] é uma técnica de aprendizado de máquina que combina vários modelos de aprendizado fraco para formar um modelo forte. Em cada iteração, o algoritmo de *boosting* ajusta o peso dos exemplos de treinamento com base no

desempenho do modelo atual, dando mais peso aos exemplos mal classificados. Isso permite que o modelo se concentre nos exemplos mais difíceis, melhorando gradualmente sua capacidade de generalização. O *boosting* é uma abordagem iterativa, por implicar múltiplas rodadas de ajuste e combinação de modelos fracos ao longo do processo. Em cada iteração subsequente, um novo modelo fraco é treinado, direcionando-se para corrigir os erros dos modelos anteriores e ajustando os pesos dos exemplos de treinamento para esse fim. Este processo contínuo de treinamento e ajuste persiste até que um critério de parada seja alcançado. Dessa forma, ao longo iterações pré-estabelecidas, é gerado um classificador final que reduz tanto o viés quanto a variância.

Label propagation[11] é uma técnica de aprendizado semi-supervisionado que se baseia na estrutura de um grafo para inferir rótulos para dados não rotulados. Começa-se construindo um grafo de vizinhos mais próximos a partir das representações dos dados, em que os nós representam os dados e as arestas indicam a proximidade entre eles. Os rótulos conhecidos são propagados através do grafo para os dados não rotulados, considerando a similaridade entre os exemplos, de modo que os rótulos dos vizinhos próximos são utilizados para estimar os rótulos dos exemplos não rotulados. Esses rótulos estimados são então incorporados ao conjunto de dados para o treinamento do modelo, com a incerteza associada aos rótulos sendo considerada durante o processo. A incerteza associada aos rótulos refere-se à confiança ou confiabilidade das predições feitas pelo modelo em relação aos exemplos não rotulados.

O método *weigthing*[12] introduz uma abordagem para melhorar a classificação semi-supervisionada baseada em grafos. Ele se destaca ao atribuir pesos às amostras com base em índices difíceis de agrupar derivados de múltiplos agrupamentos, visando melhorar à acurácia da classificação. Ao considerar amostras rotuladas próximas à fronteira de decisão entre diferentes classes como mais relevantes do que amostras rotuladas distantes dessa fronteira, o método otimiza o processo de propagação de rótulos no grafo, resultando em um desempenho superior.

Inicialmente, para o treinamento, os dados rotulados e não rotulados foram carregados e pré-processados. O modelo *BERTimbau Large*[13] e seu *tokenizador* foram inicializados, e então configurando o ambiente para uso da *GPU*. O otimizador

AdamW e a função de perda *CrossEntropyLoss* foram definidos, com o número de três épocas. Outros hiperparâmetros usados foram o *learning rate* com o valor de 2e-5, e o número do *batch size*, que foi 8. *Learning rate* controla o tamanho das atualizações dos pesos do modelo durante o treinamento. 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.

Além disso, realizamos uma propagação final de rótulos para os dados não rotulados. Dados com uma confiança maior ou igual a um limiar de 0,7, 0,8 e 0,9 foram usados no *self-learning*. No *co-training*, apenas o limitar de 0,9 foi utilizado. Então, apenas os dados dentro desses limiares foram rotulados (i.e., receberam pseudo-rótulos). O número final de dados rotulados e descartados foi contabilizado, e a distribuição das classes nos dados rotulados foi exibida.

Avaliação e Métricas

A avaliação dos métodos *self-learning* e co-training foi realizada utilizando a técnica de validação cruzada estratificada *Stratified K-Fold Cross-Validation* com o valor de *K=5*, em que as novas amostras são adicionadas conforme o valor de confiança preestabelecido. Nos demais três métodos, foi utilizada a técnica *Hold-out*, com proporção (80/20).

As classes *CustomDataset* ou *TextDataset* foram criadas para manipular textos e rótulos, definindo tokens para os textos com o *tokenizador BERTimbau*. Após cada época, o modelo prediz rótulos para os dados não rotulados, salvando os rótulos preditos e as probabilidades de confiança na última época. Para a avaliação, as métricas de desempenho calculadas incluem acurácia, precisão, *recall*, *F1-score* e *Area under the ROC Curve* (AUC). A média das métricas de todas as *folds* é calculada para fornecer uma estimativa geral do desempenho dos modelos em que foi utilizado esse método.

Resultados

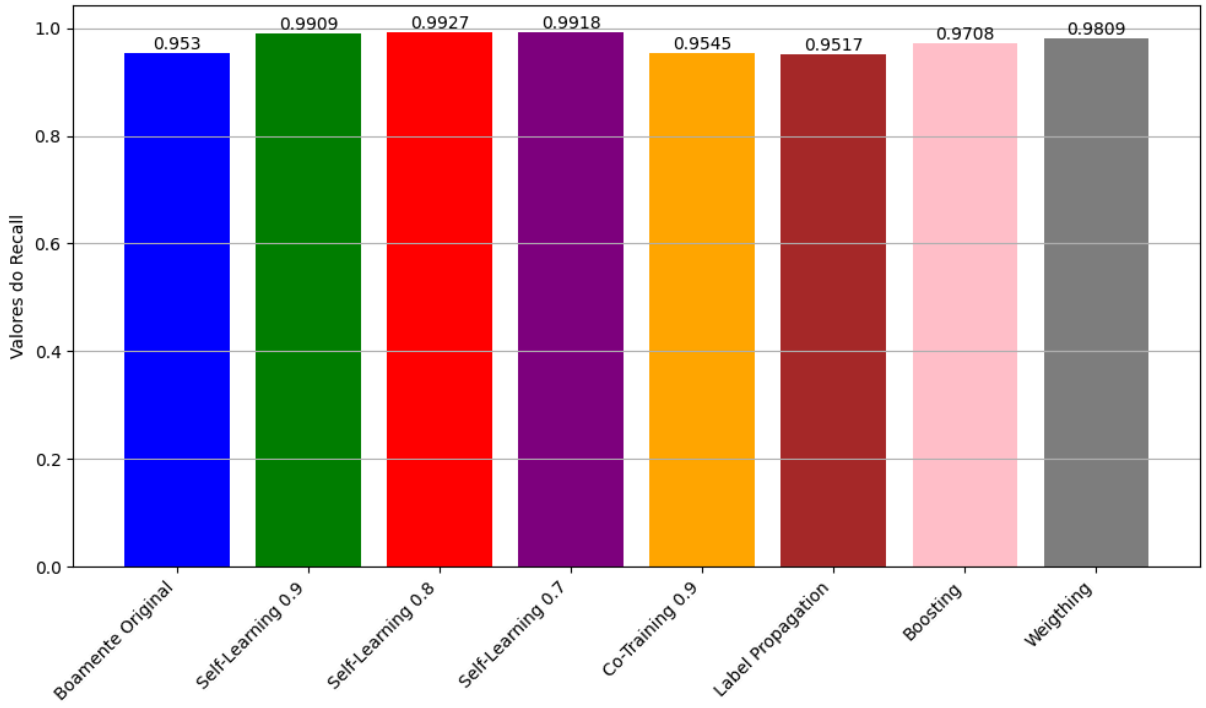
A Tabela 1 apresenta os resultados parciais obtidos dos métodos usados neste trabalho. Os melhores resultados de cada métrica estão destacados em negrito.

Tabela 1 - Desempenho dos modelos.

Método	Acurácia	Precisão	*Recall*	*F1-Score*	*AUC*
Boamente Original	0,9555	0,9612	0,9530	0,9544	0,9545
Self-Learning (Confiança de 0,9)	0,9960	**0,9954**	0,9909	0,9931	0,9945
Self-Learning (Confiança de 0,8)	**0,9966**	**0,9954**	**0,9927**	**0,9941**	**0,9954**
Self-Learning (Confiança de 0,7)	0,9945	0,9892	0,9918	0,9905	0,9937
Co-Training (Confiança de 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

A Figura 4 mostra o gráfico de comparativo da métrica *recall* entre os resultados obtidos e o valor do modelo original do *Boamente*. Esta métrica é de grande importância neste estudo, uma vez que os falsos-positivos são especialmente preocupantes, pois indicarem que o modelo está erroneamente classificando um texto que não contém ideação suicida como verdadeiro positivo. Os resultados evidenciam que, em sua maioria, os métodos avaliados superam o desempenho do modelo original. Os resultados demonstram que o método *self-learning* obteve o melhor desempenho, seguido pelo método *weigthing*, *boosting*, *co-training* e *label propagation*.

Figura 4 - Comparação dos resultados obtidos na métrica *recall*.



Discussão

Principais Achados

Os resultados demonstram que o método *self-learning* com confiança de 0,8 obteve o melhor desempenho, seguido pelo método *weigthing*, *boosting*, *co-training* e *label propagation*. Um ponto importante a se destacar é que a maioria dos métodos foi superior ao modelo original do *Boamente*. Portanto, como demonstrado nos resultados, os métodos de aprendizado semi-supervisionado possibilitaram uma melhorar significativa no modelo do *Boamente*. Dessa forma, acreditamos que os profissionais de saúde mental possam ter uma confiança maior na ferramenta proposta para a identificação de ideação suicida, além de permitir que os pacientes sejam monitorados de forma mais precisa.

Comparação com Trabalhos Prévios

No âmbito dos estudos dedicados à detecção de ideação suicida, diversas revisões da literatura, como as conduzidas por Lasri *et al.*[15], Heckler *et al.*[16] e Ji *et al.*[17], focam na prevenção do suicido a partir da identificação de sinais de ideação suicida. Dentre os diversos trabalhos que visam identificar ideação suicida,

consideramos como trabalhos relacionados aqueles que detectaram ideação suicida a partir de textos.

Trabalhos voltados para avaliar o impacto da ideação suicida na predição de comportamentos suicidas iminentes, como o de McMullen *et al.*[18], trazem a tona a necessidade de diagnóstico rápido, eficiente e preciso para um melhor resultado. Os pesquisadores realizaram um estudo com participantes de unidades hospitalares e aplicaram o *Suicide Crisis Inventory* (SCI) para avaliar a intensidade da Síndrome de Crise Suicida. Eles treinaram algoritmos de aprendizado de máquina, como *Random Forest* e *XGBoost*, para analisar os dados coletados e comparar a eficácia preditiva da SCI com e sem a inclusão da ideação suicida.

O trabalho de Birjali *et al.*[19] traz como contribuição a construção de um vocabulário associado ao suicídio para lidar com a falta de recursos terminológicos nessa área. Já no trabalho de Chatterjee *et al.*[20], os pesquisadores coletaram dados de postagens em redes sociais como *Reddit* e *Twitter* e criaram um conjunto de dados bem rotulado contendo pensamentos suicidas. Para identificar esses sinais, eles desenvolveram seis grupos de atributos que abrangem aspectos linguísticos, comportamentais e emocionais presentes nas postagens dos usuários. O modelo desenvolvido combina esses grupos de atributos em uma abordagem multimodal, utilizando técnicas de aprendizado de máquina para identificar sinais de risco de suicídio.

O *Boamente* é também um trabalho relacionado e base para este estudo envolvendo métodos de aprendizado semi-supervisionado. Ao analisar os trabalhos prévios, a contribuição científica desta pesquisa é a investigação pioneira de diferentes métodos de aprendizado semi-supervisionado para melhorar o desempenho de um modelo de aprendizado profundo para a identificação de ideação suicida a partir de textos.

Limitações

Esse estudo possui limitações que precisam ser reconhecidas. A primeira limitação é o tamanho do conjunto de dados não rotulado, que poderia ter sido maior. A coleta original do *Boamente* foi feita no *Twitter*, enquanto os dados atuais foram obtidos no *Reddit*, uma rede social menos popular no Brasil. A *API* atual do *Twitter* só

permite a coleta de dados mediante uma conta paga, o que inviabilizou o uso dessa plataforma no estudo. Ademais, outros métodos de aprendizado semi-supervisionado poderiam ser explorados, os quais teriam chances de ter resultados superiores aos apresentados aqui. Por fim, a experimentação realizada não considerou um ajuste fino exaustivo de todas as combinações de parâmetros dos métodos semi-supervisionados.

Conclusão

O presente estudo abordou o aprimoramento do modelo *Boamente* por meio de métodos semi-supervisionadas para a detecção de ideação suicida em textos não clínicos. Os resultados apresentados demonstram uma clara evolução ao modelo original que é usado como baseline para o nosso trabalho. O fato de abordamos exclusivamente textos em português-brasileiro para classificação de textos não clínicos de ideação suicida e técnicas de aprendizagem semi-supervisionado, tornam o nosso trabalho único. O método *self-learning* com o limiar de confiança de 0,8 obteve um percentual de melhora ao modelo original de 4,30%. Trabalhos futuros envolvem o uso de modelos generativos para a identificação de ideação suicida. O objetivo em usar os modelos generativos é para permitir não somente a identificação, mas também a geração de explicações da(s) razão(ões) para um texto apresentar ou não sinais de ideação suicida.

Agradecimentos

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (308059/2022-0).

Referências

1. Shin S, Kim K. Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: A case study in South Korea where suicide is the leading cause of death. Asian Journal of Psychiatry [Internet]. 2023 Oct 1;88:103725.

2. Choi M, Eun Hae Lee, Joshua Kirabo Sempungu, Yo Han Lee. Long-term trajectories of suicide ideation and its socioeconomic predictors: A longitudinal 8-year follow-up study. Social science & medicine. 2023 Jun 1;326:115926–6.

3. Facchinetti T, Benetti G, Giuffrida D, Nocera A. slr-kit: A semi-supervised machine learning framework for systematic literature reviews. Knowledge-Based Systems. 2022

Sep;251:109266.

4. Chen H, Han W, Soujanya Poria. SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training. arXiv (Cornell University). 2022 Jan 1.

5. Coppersmith DDL, Dempsey W, Kleiman EM, Bentley KH, Murphy SA, Nock MK. Just-in-Time Adaptive Interventions for Suicide Prevention: Promise, Challenges, and Future Directions. Psychiatry. 2022 Jul 18;1–17.

6. Diniz EJS, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, et al. Boamente: A Natural Language Processing-Based Digital Phenotyping Tool for Smart Monitoring of Suicidal Ideation. Healthcare. 2022 Apr 8;10(4):698.

7. Torous J, Kiang MV, Lorme J, Onnela JP. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. JMIR Mental Health. 2016 May 5;3(2):e16.

8. Amini MR, Feofanov V, Pauletto L, Hadjadj L, Devijver E, Maximov Y. Self-Training: A Survey [Internet]. arXiv.org. 2023.

9. Lang H, Agrawal MN, Kim Y, Sontag D. Co-training Improves Prompt-based Learning for Large Language Models [Internet]. proceedings.mlr.press. PMLR; 2022. p. 11985–2003.

10. Chen Y, Tan X, Zhao B, Chen Z, Song R, Liang J, et al. Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data [Internet]. openaccess.thecvf.com. 2023. p. 7548–57.

11. Iscen A, Tolias G, Avrithis Y, Chum O. Label Propagation for Deep Semi-Supervised Learning [Internet]. openaccess.thecvf.com. 2019. p. 5070–9.

12. Chen X, Yu G, Tan Q, Wang J. Weighted samples based semi-supervised classification. Applied soft computing. 2019 Jun 1;79:46–58.

13. Souza F, Nogueira R, Lotufo R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. Intelligent Systems. 2020;403–17.

14. Wagner Filho JA, Wilkens R, Idiart M, Villavicencio A. The brWaC Corpus: A New Open Resource for Brazilian Portuguese [Internet]. Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, et al., editors. ACLWeb. Miyazaki, Japan: European Language Resources Association (ELRA); 2018.

15. Lasri S, Nfaoui EH, El haoussi F. Suicide Ideation Detection on Social Networks: Short Literature Review. Procedia Computer Science. 2022;215:713–21.

16. Heckler WF, de Carvalho JV, Barbosa JLV. Machine learning for suicidal ideation identification: A systematic literature review. Computers in Human Behavior. 2022 Mar;128:107095.

17. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. IEEE Transactions on Computational Social Systems. 2021 Feb;8(1):214–26.

18. McMullen L, Parghi N, Rogers ML, Yao H, Bloch-Elkouby S, Galynker I. The role of suicide ideation in assessing near-term suicide risk: A machine learning approach. Psychiatry

Research. 2021. Oct;304:114118.

19. Birjali M, Beni-Hssane A, Erritali M. Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. Procedia Computer Science. 2017;113:65–72.

20. Chatterjee M, Kumar P, Samanta P, Sarkar D. Suicide ideation detection from online social media: A multi-modal feature based technique. International Journal of Information Management Data Insights. 2022 Nov;2(2):100103.