Journal of Computer Science

Original Research Paper

Suicide Ideation and Risk Detection from Social Media Using GPT Models

[1]Sara Lasri, [1]El Habib Nfaoui and [2]Karima Mrizik

[1]*Department of Computer Science, Faculty of Sciences Dhar El Mahraz, LISAC Laboratory,*
Sidi Mohamed Ben Abdellah University Fez, Morocco
[2]*Department of Psychology, Faculty of Letters and Human Sciences Dhar El Mahraz, Sociological, Psychological Laboratory,*
Sidi Mohamed Ben Abdellah University Fez, Morocco

Article history
Received: 18-05-2024
Revised: 21-06-2024
Accepted: 24-06-2024

Corresponding Author:
Sara Lasri
Department of Computer Science, Faculty of Sciences Dhar El Mahraz, LISAC Laboratory, Sidi Mohamed Ben Abdellah University Fez, Morocco
Email: sara.lasri@usmba.ac.ma

Abstract: As a reason for the sensitiveness of suicide ideation and its considerable impact on people's lives, the demand to treat and prevent the suicide ideation issue has become an obligation. Suicide ideation is a result of a combination of psychological pain and hopelessness. According to the World Health Organization, the task of reducing the global suicide mortality rate is a target, that should be attained. The entire population uses social media platforms to express their feelings, emotions, sentiments, and opinions. Social media platforms are among the most popular sources of datasets related to mental health issues. The process of detecting suicide ideation from social media platforms is based on recent methods of artificial intelligence such as machine learning and deep learning. In this study, we propose fine-tuning large language models to evaluate and find the level of suicide risk in posts published on Reddit. We fine-tuned four GPT-3 models using the UMD Reddit suicidality dataset, which is related to the subreddit of suicidal ideation. Our experimental results illustrate the efficiency of the LLMs in addressing our task. The model attains a high F1-score of 92.3%, an accuracy of 94.8%, and a training loss of 0.050.

Keywords: Suicide Ideation, Level of Suicide Risk, Large Language Models, Generative Pre-trained Transformer, Fine-Tuning, Reddit

Introduction

The word suicide tags to all instances of mortality resulting directly or indirectly from self-inflicted harm. Emile Durkheim's suicide study found several things to understanding suicide. According to this study, suicide is heavily influenced by social integration (Fournier, 2020). He distinguishes between four categories of suicide: [Anomic suicide, altruistic suicide, fatalistic suicide, and egoistic suicide]. The Centers for Disease Control and Prevention (CDC) defines [suicide as a type of mortality brought on by self-directed, harmful behavior to kill one's own life]. Following to, WHO suicide is defined as a serious global public health problem and a complex and multidimensional problem that is often linked to various mental health disorders. The Diagnostic and Statistical Manual of Mental Disorders (DSM 5) denotes the fact that there is a high correlation between mental health illnesses such as depressive disorders, anxiety disorders, bipolar-related disorders, personality disorders, and suicide risk. While suicide is not a condition, if a person experiences mental health problems, they will attempt or die by suicide. Following the World Health Organization, the suicide mortality data shows the need to prevent it and the task of reducing the global suicide rate by one-third by 2030 a target that should be realized. However, about 800,000 individuals worldwide commit suicide each year and it is also the second-largest cause of death for both males and females between the ages of 15 and 29. For each instance, numerous people attempt to take their own lives, and all ages, sexes, and regions of the world are affected. One in every 100 deaths globally is by suicide. Moreover, suicide happens everywhere in the world, with low and middle-income nations accounting for 77% of all suicides. To prevent suicide, we focus on data from social networking platforms, which are a source of opinions, feelings, emotions, and sentiments for users. In our previous studies (Lasri *et al*., 2022), we were interested in discovering the absence or presence of suicidal mental health issues on social media platforms. We used deep learning algorithms to realize this task. Recently, our goal has been not only to categorize the lack or existence of suicidal ideation but also to determine the level of suicide risk. By determining the level of suicide risk, we can easily predict and prevent suicidal ideation. In order to help the mental health workers, provide exact treatment and suitable monitoring to help the patients have and gain psychological well-being. The process of detecting the level of suicide risk is a big challenge that consists of the combination of clinical assessment and the recent technology of data analysis. For this reason, we focus on the recent techniques of artificial intelligence models. We employ extensive Language Models (LLMs) that signify the future of the medical field, harnessing deep learning for Natural Language Processing (NLP) and Natural Language Generation (NLG). These transformer-based models are trained on extensive datasets from the internet, allowing them to comprehend and produce human language. Designed to recognize, translate, predict, and create text and other content, LLMs are poised to revolutionize various aspects of healthcare. Hence, by applying LLMs in our context, we train the models on UMD Reddit to better understand the posts and effectively identify the level of suicide risk.

Related Work

An Overview of Methods for Detecting Suicidal Ideation

The literature describes various techniques for automatically detecting suicide ideation across different social networking platforms, (Renjith *et al*., 2022; Ben Hassine *et al*., 2022; Mendes *et al*., 2023), including machine learning and deep learning methods (Parsapoor *et al*., 2023). Because of the importance of suicide disorder mental health and the massive amounts of data published in this context, several studies are based on the large language model. They employ the pre-trained model to examine the large volumes of data from social media that are related to suicide ideation and employed in specific tasks. Concerning machine learning methods, several studies use them to detect suicidal ideation this research Ji *et al*. (2021) reviews the use of machine learning for detecting suicidal ideation across various datasets, including questionnaires, electronic health records, suicide notes, and online social texts. The majority of these studies implement different machine learning algorithms to detect the most effective ones, such as this study (Rabani *et al*., 2020), sought to assess the feasibility of identifying and distinguishing suicidal tweets from non-suicidal ones by analyzing data from social media platforms. To achieve this, the researchers employed various supervised machine learning approaches, such as decision trees, Naïve Bayes, logistic regression, multinomial Naïve Bayes, random forest, support vector machine, voting, AdaBoost, and stacking. They achieve an accuracy of 98.5% by using a random forest. Additionally, this study (Aldhyani *et al*., 2022) utilizes a diverse range of features, encompassing syntactic, statistical, linguistic, word embedding, and topic features. These features are assessed using various classifiers from both traditional and deep learning algorithms. The classifiers employed include support vector machines, random forests, Multilayer Feed-Forward Neural Networks (MLFFNN), Gradient Boosted Decision Trees (GBDT), Long Short-Term Memory (LSTM), and XGBoost. Among these methods, XGBoost achieved the highest accuracy when all feature groups were utilized as inputs based on machine learning algorithms, particularly random forest, logistic regression, Multilayer Perceptron (MLP), XGBoost, and Convolutional Neural Network (CNN), to identify individuals with suicidal ideation who may be at risk of attempting suicide. To realize this task, they used data from the Korean National Health and Nutrition Examination Survey. It analyzes extensive data spanning around four years (from 2017-2020) gathered by the Korea Youth Policy Institute (KYPI). In this study (Rabani *et al*., 2020), researchers explore the effectiveness of deep learning architectures (Sawhney *et al*., 2018) like LSTMs, RNNs, and C-LSTMs in constructing precise and resilient models for detecting suicidal ideation. They evaluate these models against standard baselines in text classification tasks and use search queries aligned with a generated lexicon to identify and classify tweets from the Twitter REST API into either suicidal or non-suicidal categories. Quantitative comparisons among several models demonstrated the efficacy of a CLSTM-based model in identifying suicidal ideation in tweets, achieving an accuracy of 81.2%. Besides, this study uses (Selvi *et al*., 2023) deep learning systems to automatically and proactively detect suicidal thoughts and sudden changes in user behavior by analyzing users' social media posts.to detect suicidal ideation using word encoding techniques like TF-IDF, Word2Vec, and deep learning for classification using the publicly available Reddit dataset. They utilize a Bilateral Long-Term Memory (BiLSTM) model to categorize social media messages as suicidal, which has an accuracy of 94.2%. In a related context (Du *et al*., 2018) the study concentrates on identifying suicide-related psychiatric stressors from Twitter using deep learning methods and a transfer learning approach that leverages annotations from clinical text datasets. They utilized a Convolutional Neural Network (CNN) algorithm to construct a binary classifier for this purpose. Additionally, the identification of psychiatric stressors in suicide-related tweets was approached as a Named Entity Recognition (NER) task and addressed using Recurrent Neural Network (RNN) based methods. With a 78% accuracy rate in identifying tweets related to suicide, CNN is performing better than everyone else. Multiple studies have utilized both machine learning and deep learning methods to distinguish suicidal ideation, comparing them to determine the most operational algorithms for classifying signs of suicide ideation. Such as this study Ji *et al*. (2021) which is based on both deep learning and machine learning algorithms to categorize Reddit posts as either suicidal or non-suicidal, their outcomes show that the CNN-BiLSTM model outperformed the XGBoost model, achieving a 95% accuracy rate in identifying suicidal thoughts, compared to XGBoost's 91.5% accuracy rate. In another paper (Haque *et al*., 2022), presents a comparative study between machine learning and deep learning models for distinguishing suicidal thoughts on the social media platform Twitter. CountVectorizer feature extraction was utilized in ML models such as MNB, LR, SGD, RF, and SVC. Additionally, word embeddings were employed in DL models like CLSTM, BiLSTM, BiGRU, and LSTM. The performance of the ML and DL models was then compared using evaluation matrices following training. According to their experimental results, the RF model demonstrated a high accuracy of 93%, making it the top performer among the machine learning methods in terms of classification score. Recently, numerous studies have highlighted Large Language Models (LLMs) as pivotal in recognizing suicide indicators through social media. These models are increasingly regarded as the future of psychological domains, offering effective tools for detecting suicidal ideation. Cheng *et al*. (2023); Hua *et al*. (2024); Demszky (2023). Hence, what is the impact of Large Language Models (LLMs) on suicide ideation? Moreover, what is the benefit that we can gain when we implement Large Language Models (LLMs) in mental health in general?

Large Language Model (LLMS) and Mental Health Application

Large Language Models (LLMs) are among the most recent artificial intelligence techniques that have had a significant effect on the medical domain or healthcare in general (Noopur Rewatkar and Rewatkar, 2024). That means we can benefit from LLMs to enhance the process of diagnosing, preventing, treating, and managing the disorder's health and the patient's medical papers about the mental health aspect (Clusmann *et al*., 2023). All WHO member states are dedicated to enhancing the mental health of individuals, Moreover, considering it an essential human right, Mental health is essential for personal wellness, community resilience, and socio-economic development. There are various applications of Large Language Models (LLMs) in the mental health domain, such as the creation of mental health catboats, therapy support and guidance, automated psychological-text based counseling, language-based interventions for Cognitive Behavioral Therapy (CBT), Electronic Health Records (EHR), sentiment analysis for social media monitoring, personalized treatment plans, digital mental health screeners. The employing of a large language model in the creation of catboats or virtual assistants, clinical documentation, and note taking, which are question-answering systems based on large language models that patients may use to communicate with people and acquire knowledge, advice and assistance around their mental health and medical needs. These catboats, offer sympathetic replies, analyze patients' well-being and provide solutions to their problems. Corresponding to (Lai, 2023), they propose a Psy-LLM framework, which is an AI-based assistive tool that utilizes Large Language Models (LLMs) for question-answering in psychological consultation contexts. This framework intends to reduce the request for mental health professionals by offering a practical tool for efficient screening and rapid response to individuals in urgent need of mental support, thereby addressing and alleviating pressing needs within the healthcare sector. In addition, large language models are the most effective technique in the mental health support field for enhancing therapy support and guidance, therapeutic interaction, or the process of analyzing and interpreting mental health issues. Among the papers that talked about this topic, we found this study (Stade *et al*., 2023). LLMs show promising capabilities in performing tasks essential for psychotherapy, such as conducting assessments, providing psychoeducation, and demonstrating interventions. According to this study, large language models can support, augment, or even potentially replace human-led psychotherapy in some cases. This advancement could improve the quality, availability, reliability, and flexibility of therapeutic interventions and clinical science research. Consequently, LLMs could lead to the future of clinical science by aiding in psychodiagnostics evaluations and identifying harm risks, especially in recognizing suicidal or homicidal thoughts, child or elder abuse, and intimate partner violence. Using the large language model in automated text-based counseling to present psychological counseling that involves the provision of professional guidance and support to individuals dealing with mental health challenges. According to this area, this study (Fu *et al*., 2023) presents the efficiency of psychological interventions on social media platforms. Hence, recent advances in large language models can detect the alarming number of users expressing negative emotions on social media and help train psychological counselors to provide effective mental interventions. This study introduces a novel model built on large language models designed to assist nonprofessionals in delivering psychological interventions during online user interactions. The model, called the LLM counselors support system, aims to enhance counselors' communications with individuals experiencing depression. In this iterative process, when a person with depression initiates a conversation, the counselor uses the system to formulate a response. In addition, large language models are the most effective technique in the mental health support field for enhancing therapy support and guidance, therapeutic interaction, or the process of analyzing and interpreting mental health issues. The impact of adopting language models in Electronic Health Records (EHR) systems is providing real-time decision support. Therefore, language models assist Electronic Health Records (EHR) in identifying and aggregating relevant data from the patient's medical history, such as demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports, to recommend personalized treatment and promptly review the protocol for the specific patient. The research described in (Kariotis *et al*., 2022) seeks to examine the impact of Electronic Health Records (EHRs) on information practices in mental health environments. The review revealed that Electronic Health Records (EHRs) in mental health settings impact clinicians' information practices, with implications for care provision. The cornerstone of mental health services is the therapeutic relationship, which necessitates a distinct workflow that many current EHR systems do not adequately support. Among the important applications of LLMs in mental health is sentiment analysis for social media monitoring. In this case, LLMs can analyze social media content to understand user sentiments related to mental health topics to detect mental health issues on social media platforms. Among the papers that talked about this topic, we found (Ji *et al*., 2023a), a study that aimed to evaluate the effectiveness of using pre-trained language models to identify binary and multi-class mental disorders, including stress, anxiety, and depression. They utilized a range of pre-trained language models, including BERT, RoBERTa, BioBERT, ClinicalBERT, MentalBERT, and MentalRoBERTa. Additionally, they utilized the Reddit corpus, focusing on subreddits related to the mental health domain. For depression detection, they use two datasets, CLPsych15 and Depression_Reddit (DR). Concerning stress, they adopt two datasets, Dreaddit and SAD. They also use the T-SID, UMD, and SWMH datasets for suicidal ideation detection and CAMS for causal analysis of mental health issues i.e., bias or abuse, jobs and careers, medication, relationships, and alienation. In addition, in this study (Pourkeyvan *et al*., 2024) researchers employed a pretrained BERT model from the hugging face library to diagnose mental health disorders in social networks. They fine-tuned four different pre-trained BERT models distilbert-base-uncased, fine-tuned-sst-2 English, bert-base-uncased, distilroberta-base and mental-bert-base-uncased using two distinct datasets consisting of users' tweets and bios. The study illustrates these models' potential in accurately predicting depression symptoms, achieving high accuracy and F1-scores that exceed previous research. This represents a significant advancement in the automated detection of depression symptoms. The findings confirm that BERT models from the Hugging Face library can serve as valuable tools for mental health screening and monitoring, enhancing the understanding of mental health conditions and the development of effective interventions. In another study (Rathje *et al.*, 2023), GPT and all large-language models may be the future of psychological text analysis and may help facilitate more cross-linguistic research with the understudied. They tested the ability of GPT to accurately detect psychological constructs in text across 15 datasets. For each psychological construct, they first examined GPT's performance in English as well as a second, unrelated language, using six publicly available datasets with categorical labels. Finally, to examine whether GPT performs equally well with less commonly spoken or studied languages, we tested GPT's ability to detect sentiment in eight African languages, such as Swahili, Amharic, Yoruba, and Kinyarwanda. According to this study, GPT appears to be effective at multilingual sentiment analysis, with performance comparable to top-performing machine learning models from previous studies. This suggests GPT-4 might have a cross-linguistic bias towards overestimating sentiment in a given text compared to humans. Furthermore, the researchers of this study (Ji *et al*., 2023b) researchers trained the pre-trained MentalBERT and MentalRoBERTa models using mental health data sourced from Reddit and Twitter to detect and classify several mental disorders such as stress, anxiety, suicide, and depression. The study highlights that domain-specific pre-trained models for mental health, like MentalBERT and MentalRoBERTa, generally outperform those trained on general corpora. The findings suggest that ongoing pretraining with mental health-related datasets can enhance classification accuracy significantly. This study (Qi *et al*., 2023) is also part of the research centered on utilizing Large Language Models (LLMs) to identify mental health issues associated with suicide. The study assessed the efficacy of LLMs in two mental health tasks on Chinese social media: Classifying suicide risk and multi-label classification of cognitive distortions. It compared LLM performance using three approaches: Zero-shot, few-shot, and fine-tuning. The findings suggest that GPT-4 demonstrates superior performance across different scenarios, while GPT-3.5 exhibits significant improvement in suicide risk classification after fine-tuning.

Materials and Methods

This study deals with the task of suicide ideation identification from social media posts, which is a serious public mental health problem. In particular, we aim to establish the suicide risk related to the posts published on the Reddit social media platform. It can be formulated as a multi-class classification problem. We propose to fine-tune GPT-3 as a general-purpose LLM for training a task-specific fine-tuned model able to detect the risk of suicide

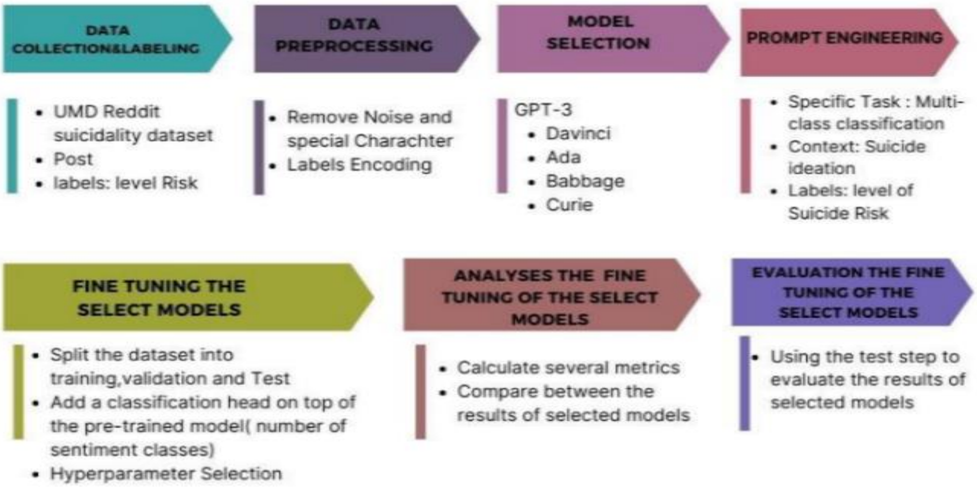
DATA COLLECTION&LABELING
UMD Reddit suicidality dataset
Post
labels: level Risk
DATA PREPROCESSING
Remove Noise and special Charachter
Labels Encoding
MODEL SELECTION
GPT-3
Davinci
Ada
Babbage
Curie
PROMPT ENGINEERING
Specific Task : Multi-class classification
Context: Suicide ideation
Labels: level of Suicide Risk
FINE TUNING THE SELECT MODELS
Split the dataset into training,validation and Test
Add a classification head on top of the pre-trained model(number of sentiment classes)
Hyperparameter Selection
ANALYSES THE FINE TUNING OF THE SELECT MODELS
Calculate several metrics
Compare between the results of selected models
EVALUATION THE FINE TUNING OF THE SELECT MODELS
Using the test step to evaluate the results of selected models

Fig. 1: The process of data pre-processing and models fine-tuning

conveyed in Reddit posts. The level of risk ranges from low, moderate, severe, or none. Figure 1 shows the overall process of data processing and model fine-tuning for suicide ideation detection from Reddit.

GPT-3 (Generative Pre-Trained Transformer)

GPT-3, created and trained by OpenAI, is a sophisticated language model capable of comprehending and generating natural language. It has been succeeded by the GPT-3.5 and GPT-4 models. The original GPT-3 base models consist of Curie, Davinci, Ada, and Babbage. Comparing Davinci and Ada, the cost of using Davinci is often higher. Ada, the fastest and most cost-effective model in the GPT-3 series, excels at performing straightforward tasks. Its efficiency and affordability make it particularly successful in various applications, including sentiment analysis. Babbage, a less flexible variant of GPT-3, excels at straightforward tasks and is both very fast and cost-effective. Curie is highly capable, offering a balance of speed and cost that is faster and cheaper than Davinci. In this study, first, we used these base models for the task of suicide ideation detection. Second, we fine-tuned them using the Reddit dataset in order to select the most capable model on this NLP sub-task.

LLM Fine-Tuning

Fine-tuning is an effective method for enhancing pre-trained models. The primary goal of fine-tuning is to customize the pre-trained model for specific topics or tasks such as text classification, named entity recognition, question answering, sentiment analysis, and text summarization thereby improving its effectiveness and accuracy in task-related predictions. In this study, we fine-tuned the original GPT-3 base models Davinci, Curie, Ada, and Babbage for multiclass classification tasks using the UMD Reddit suicidality dataset. We have kept the default parameters provided by OpenAI fine-tunes API (batch_size, n_epochs, and learning_rate_multiplier).

Dataset: UMD Reddit Suicidality Dataset

Researchers at the University of Maryland created the University of Maryland Reddit suicidality dataset by gathering data from Reddit, specifically from subreddits focused on suicidality and suicide prevention. This dataset was introduced at the 2019 workshop on computational linguistics and clinical psychology and utilized information extracted from the 2015 full Reddit submission corpus. In this study (Zirikly *et al*., 2019), annotations were gathered from both experts and crowdsource workers for a randomly selected subset of users based on their postings in SuicideWatch on Reddit. They categorized users into four risk levels as follows (Shing *et al*., 2018):

a) [No risk ("none"): There is no indication that this person is at risk for suicide]
b) [Low risk: Some factors may suggest risk, but it is unlikely that this person is at significant risk of suicide]
c) [Moderate risk: Symptoms indicate a real possibility that this person may attempt suicide]
d) [Severe risk: This person is highly likely to attempt suicide in the near future]

The UMD Reddit suicidality dataset comprises expert ratings for 245 users and crowdsourced ratings for a larger set of 865 users. This dataset comprises various essential attributes for each post, including post ID (a unique identifier), user ID (a unique numeric identifier for the post author), timestamp (Unix epoch time indicating when the post was created), subreddit (the name of the subreddit where the post appeared), post title (the title of the post) and post body (the textual content of the post) (Shing *et al*., 2018).

Results and Discussion

In this section, we provide a comprehensive analysis of the fine-tuning process and the subsequent evaluation of the fine-tuned models on the UMD Reddit suicidality dataset. Our objective was to detect varying levels of suicide ideation, and to this end, we fine-tuned four models: Curie, Ada, Babbage, and Davinci.

Table 1 summarizes the performance metrics of these models on the Reddit test set. The Curie model emerged as the most effective, achieving an impressive accuracy of 94.8%. This high accuracy indicates that Curie is particularly adept at identifying the suicide risk level in Reddit posts. Moreover, the model's F1-score of 92.3% underscores its robustness in classification tasks, balancing precision and recall to effectively distinguish between different levels of suicide ideation.

Comparatively, the other models Ada, Babbage, and Davinci also performed well but did not match Curie’s level of accuracy and F1-score. This suggests that while all four models are capable of detecting suicide ideation, Curie has a superior ability to capture the nuances in language that indicate varying levels of risk.

The findings from this study highlight the importance of selecting and fine-tuning the right model for specific tasks such as mental health risk assessment. The Curie model, in particular, demonstrates a strong potential for application in real-world scenarios where accurate detection of suicide risk is crucial.

These results also suggest that further refinement of the fine-tuning process could enhance the performance of the other models, potentially closing the gap between them and Curie. Overall, the experimental results validate the effectiveness of the Curie model in the context of suicide ideation detection and offer insights into how different models can be leveraged for similar tasks in the future.

In the fine-tuning process, the Curie model demonstrated exceptional performance, achieving a final training loss of 0.050%. This low training loss highlights the model's efficiency in learning and understanding the intricacies of posts related to varying levels of suicide risk. The model's ability to minimize loss to such a degree underscores its effectiveness in discerning the subtle linguistic cues that often indicate suicide ideation.

Table 1: Models performance metrics comparison using the Reddit test set

Model	Accuracy (%)	F1-score (%)
Fine-tuned Ada	0.945	0.921
Fine-tuned Curie	0.948	0.923
Fine-tuned Babbage	0.945	0.921
Fine-tuned Davinci	0.777	0.670

Moreover, the Curie model attained an impressive final training accuracy of 98.9% across three epochs. This high accuracy level suggests that the model is highly proficient at capturing complex patterns, features, and contextual nuances within the text sequences. Such proficiency is crucial for accurately classifying posts according to their suicide risk levels during the training phase. The model’s success in achieving these metrics not only illustrates its capability in handling the task but also positions it as a valuable tool for future applications in mental health risk assessment.

The implications of these findings are significant. The Curie model’s ability to effectively classify and detect suicide ideation can have a profound impact on real-world applications, particularly in aiding mental health professionals in monitoring and supporting individuals at risk. By accurately predicting the level of suicide risk, the model can contribute to timely interventions, potentially preventing suicides and improving outcomes for individuals experiencing mental health crises.

As advancements in AI continue, OpenAI’s recent release of GPT-3.5-turbo and GPT-4 presents new opportunities for further enhancement of this task. These more advanced models are likely to offer improved performance, with enhanced capabilities in understanding and predicting suicide ideation. Exploring these models for future iterations of this task could lead to even greater precision in identifying at-risk individuals and predicting their potential actions.

In conclusion, the fine-tuning process has shown that the Curie model is highly effective in classifying posts related to suicide risk. Its strong performance in both training loss and accuracy metrics highlights its potential for practical application in mental health monitoring and intervention. The planned exploration of GPT-3.5-turbo and GPT-4 models offers a promising avenue for future improvements, with the goal of further refining suicide ideation detection and ultimately contributing to better mental health outcomes.

Conclusion

In this study, we proposed to fine-tune general-purpose LLMs for preventing suicide by detecting suicidal ideation published on social networks such as Reddit. Our purpose is to help human beings have stable mental health and psychological well-being. We achieve significant results by finetuning GPT-3 models Curie, Ada, Babbage, and Davinci. Our results show that the Curie fine-tuned model achieved satisfactory performance in detecting levels of suicide risk. Consequently, we can support individuals experiencing suicidal ideation or attempts by developing prevention and psychological care programs similar to those offered by the American Psychological Association

(APA). These programs utilize psychological research to combat mental health stigma and advocate for initiatives that enhance education and community services, aiming to identify and assist individuals at risk of suicide and their families.

Acknowledgment

The authors would like to express their gratitude to Philip Resnik for providing the UMD Reddit suicidality dataset and to mark Dredze for providing the dataset used in the CLPsych 2015 shared task.

Author’s Contributions

Sara Lasri: Literature review, research design, data collection and preprocessing, model development, experimentations, analysis and interpretation the results, manuscript written.

El Habib Nfaoui: Guidance and supervision, conceptual input, review and feedback, strategic decision-maker, co-author responsibilities, intellectual leadership.

Karima Mrizik: Psychological expertise, annotation guidance.

Ethics

All information presented in this study is confidential and original. This study has not been published or submitted for review elsewhere. There are no ethical concerns related to this research.

References

Aldhyani, T. H. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., & Ahmed, Z. A. T. (2022). Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*, *19*(19), 12635. https://doi.org/10.3390/ijerph191912635

Ben Hassine, M. A., Abdellatif, S., & Ben Yahia, S. (2022). A novel imbalanced data classification approach for suicidal ideation detection on social media. *Computing*, *104*(4), 741–765. https://doi.org/10.1007/s00607-021-00984-0

Cheng, S., Chang, C., Chang, W., Wang, H., Liang, C., Kishimoto, T., Chang, J. P., Kuo, J. S., & Su, K. (2023). The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and Clinical Neurosciences*, *77*(11), 592–596. https://doi.org/10.1111/pcn.13588

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., & Löffler, C. M. L. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 141. https://doi.org/10.1038/s43856-02300370-1

Demszky, D. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5.

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making*, *18*(S2), 77–87. https://doi.org/10.1186/s12911-018-0632-8

Fournier, M. (2020). Émile Durkheim (1858-1917). In *Bibliothèque idéale de psychologie* (pp. 56–58). Éditions Sciences Humaines. https://doi.org/10.3917/sh.marmi.2020.02.0056

Fu, G., Zhao, Q., Li, J., Luo, D., Song, C., Zhai, W., Liu, S., Wang, F., Wang, Y., Cheng, L., & Zhang, J. (2023). Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals. *ArXiv*, arXiv:2308.15192.

Haque, R., Islam, N., Islam, M., & Ahsan, M. M. (2022). A Comparative Analysis of Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies*, *10*(3), 57. https://doi.org/10.3390/technologies10030057

Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y., Zhou, P., Moran, L., Ananiadou, S., & Beam, A. (2024). Large Language Models in Mental Health Care: a Scoping Review. *ArXiv*, arXiv:2401.02984.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021). Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Transactions on Computational Social Systems*, *8*(1), 214–226. https://doi.org/10.1109/tcss.2020.3021467

Ji, S., Zhang, T., Yang, K., Ananiadou, S., Cambria, E., & Tiedemann, J. (2023a). Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health. *ArXiv*, arXiv:2304.10447.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2023b). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *ArXiv*, arXiv:2110.15621.

Kariotis, T. C., Prictor, M., Chang, S., & Gray, K. (2022). Impact of Electronic Health Records on Information Practices in Mental Health Contexts: Scoping Review. *Journal of Medical Internet Research*, *24*(5), e30405. https://doi.org/10.2196/30405

Lai, T. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *ArXiv*, arXiv:2307.11991.

Lasri, S., Nfaoui, E. H., & El haoussi, F. (2022). Suicide Ideation Detection on Social Networks: Short Literature Review. *Procedia Computer Science*, *215*, 713–721. https://doi.org/10.1016/j.procs.2022.12.073

Mendes, L., Leonido, L., & Morgado, E. (2023). Correlation between Suicidal Ideation and Addiction to Various Social Media Platforms in a Sample of Young Adults: The Benefits of Physical Activity. *Societies*, *13*(4), 82. https://doi.org/10.3390/soc13040082

Noopur Rewatkar, S. P., & Rewatkar, N. (2024). *Large Language Models and Generative AI's Expanding Role in Healthcare*. https://doi.org/10.13140/RG.2.2.20109.72168

Parsapoor (Mah Parsa), M., Koudys, J. W., & Ruocco, A. C. (2023). Suicide risk detection using artificial intelligence: the promise of creating a benchmark dataset for research on the detection of suicide risk. *Frontiers in Psychiatry*, *14*, 1186569. https://doi.org/10.3389/fpsyt.2023.1186569

Pourkeyvan, A., Safa, R., & Sorourkhah, A. (2024). Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks. *IEEE Access*, *12*, 28025–28035.

Qi, H., Zhao, Q., Li, J., Song, C., Zhai, W., Dan, L., Liu, S., Yu, Y. J., Wang, F., Zou, H., Yang, B. X., & Fu, G. (2023). Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media. *ArXiv*.

Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2020). Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches. *Baghdad Science Journal*, *17*(4), 1328. https://doi.org/10.21123/bsj.2020.17.4.1328

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/sekf5

Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 9564–9575. https://doi.org/10.1016/j.jksuci.2021.11.010

Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. (2018). Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 167–175. https://doi.org/10.18653/v1/w18-6223

Selvi, S. S., Parthiban, R., & Sivakumar, G. (2023). Deep Learning Algorithms for Suicide Prediction Based on Bilateral Long-Term Memory Using Social Media Behaviour Dataset. *International Journal of Innovative Research in Technology*, *9*(11), 930–935.

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 25–36. https://doi.org/10.18653/v1/w18-0603

Stade, E. C., tirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., & Eichstaedt, J. C. (2023). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *PsyArXiv*. https://doi.org/10.31234/osf.io/cuzvr

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. *CLPsych (Computational Linguistics and Clinical Psychology) 2019*, 24–33.