Project Proposal:
Detecting Suicide Intention on Social Media using NLP

Kelvin Mock (300453668), Sabrina Cai (300399089), Jenifer Yu (300399089)

CSI5386 Natural Language Processing Course Project, University of Ottawa

kmock073@uOttawa.ca, wyu094@uOttawa.ca, hcai062@uOttawa.ca

Abstract – Suicide is a significant public health concern, with millions of lives affected worldwide. The increasing use of AI in mental health applications presents both opportunities and challenges. It is believed that AI-driven suicide detection models can provide timely interventions before crisis occurs. As a result, this project aims to develop a multi-model suicide detection system utilizing state-of-the-art NLP models while maintaining suitable ethical considerations such as fairness, bias, and explainability that helps reduce disparities in mental health predictions. The goal is to optimize its predictive accuracy such that it identifies suicidal contents across social media platforms appropriately.

1 Introduction

1.1 Problem Statement & Motivation

Introduction Suicide is a significant global health concern, with millions of individuals expressing distress online before attempting self-harm. While social media platforms are supposed to provide a space where individuals share their thoughts and emotions, people who suffer from suicidal thoughts might find those platforms a valuable source to identify their suicidal intentions. However, this is never a healthy atmosphere since it encourages negative sentiments being spread over the internet and thus, yields a vicious cycle where people tend to believe expressing negative thoughts on the internet is an appropriate way. In turn, the detection of suicidal ideation on social media is a crucial task [13]. Therefore, this project intends to establish a multi-model solution in the detection of suicidal intentions through contents posted on social media platforms.

Ethical Challenges On the contrary, critics might pinpoint ethical concerns regarding privacy, consent, and data security [21]. Suicide detection using AI has gained attention as social media provides valuable real-time data for identifying individuals at risk [3]. Users may not be aware their data are being used, while there is a risk of misuse, making ethical oversight essential.

An Ethical Solution To mitigate ethical concerns, this project exclusively employs publicly-accessible datasets while ensuring fairness and bias mitigation in LLM-based suicide risk prediction. It will develop a suicide detection model based on pre-trained models like BERT [7], GPT-based models [14], and DeepSeek [5]. Hence, the project will focus on fairness, bias mitigation, and explainability to ensure responsible AI practices. By enhancing predictive accuracy and maintaining ethical standards, this approach aims to improve mental health monitoring while protecting user rights.

Linguistic Challenges Technically speaking, detecting suicidal intentions through automated means poses another challenge due to the nuances of human language, sarcasm, slangs and varying contextual expressions, which are beyond the definition in an official dictionary. Therefore, as part of the modeling, this project aims to train an NLP-based model which is capable of accurately identifying suicidal intent in social media posts, helping in early intervention efforts.

1.2 Objectives

The primary objective of this research is to build a robust NLP model to detect suicide intention in from social media text, which might also occasionally involve emojis. The system should efficiently analyze the sentiment of text, determine the likelihood of suicidal intent, and provide actionable insights that can assist mental health organizations in early intervention.

1.3 Research Gaps

Existing models for suicide detection might face limitations in the following criteria:

- Understanding contextual nuances and emojis within suicidal expressions;
- Handling imbalanced datasets where suicidal content is rare;
- Lack of real-time processing for early intervention; and,
- Ethical concerns surrounding data privacy and accuracy.

Addressing the Gap This project will address these challenges with advanced NLP techniques that we learnt throughout the course, namely transformer-based models, as well as maintaining suitable ethical principles.

1.4 Research Questions

Hence, there are a few research questions:

1.4.1 Technical Aspects

1. How effective are transformer-based models (e.g., BERT [6], RoBERTa [10], DistilBERT [20] [9]) in detecting suicidal intent in social media posts compared to traditional machine learning models?
2. What linguistic patterns and sentiment markers are most indicative of suicidal intent in social media text?
3. Can hybrid models combining deep learning with rule-based filtering improve suicide detection performance?
4. How do different feature extraction techniques (e.g., sentiment analysis, emotion detection, word embeddings) impact the model's accuracy in identifying suicide intention?

1.4.2 Ethical Aspects

In combination with the work from the course project of *CSI5195 - Ethics in AI* - this project will also dive into a few ethical questions:

1. How do existing LLMs (e.g., BERT [6], GPT-4o [14], DeepSeek [5]) exhibit bias in suicide detection, and how can it be mitigated?
2. What methods can be used in the evaluation of fairness in an AI-based suicide risk prediction across different demographic groups?
3. How can explainability (or known as "interpretability") techniques such as the SHAP analysis [11] help to make AI-driven suicide detection models more explainable (i.e., transparent and trustworthy)?
4. How do feature attribution methods compare in explaining AI predictions?

2 Methodology

2.1 Datasets

We will consider publicly available mental health and suicide risk datasets:

- **CLPsych** [4]: Containing social media posts labeled for suicide risk.
- **Reddit SuicideWatch Posts**
- **Psychiatric Patient Records**
- **Social Media Sentiments Analysis** [15]: served as a corpus marked the sentiment of each piece of text.
- **Twitter Suicidal Data** [1]: A dataset with abundant suicidal texts best-suited for supervised fine-tuning (SFT) [12].
- **Depression Tweets** [16]: A JSON-structured dataset containing depression-related texts found on Twitter.

2.2 Models

The above mentioned datasets contain text-based user-generated content, allowing NLP models to assess risk levels. We will make a comparison across multiple NLP models, including:

1. **Baseline Model**: Based on the traditional machine learning approach, 2 models will be implemented as the baseline in comparison with other models.
   - Decision Tree or Random Forest classifier
   - SVM classifier
   - Logistic Regression - predicting in a binary classification problem probabilistically
2. **Deep Learning Models**
   - LSTM (Long Short-Term Memory)
   - Transformer-based models (e.g., BERT [7], RoBERTa [10], DistilBERT [20] [9])
3. **LLM-based Models**
   - OpenAI GPT-based models [14]
   - DeepSeek [5]
   - Google Gemini (as a backup plan)
4. **Hybrid model**: combining deep learning with rule-based filtering.

2.3 Data Collection

This project intends to collect necessary data (for modeling) from a combination of **web scraping** which adheres to ethical considerations as well as **publicly-available datasets** as mentioned above.

2.4 Preprocessing

The following procedures will be followed in order to get the data in the format which is compatible in our analysis:

1. Removing noise,
2. Handling spelling variations, and
3. Normalizing text.

2.5 Annotation and Labelling

We admit the significance of manual annotation especially those done by domain experts. Obviously, we will need to rely on some existing labelled datasets.

2.6 Labeling Categories

In our analysis, text from social media platforms is normally expected to be in either category:

- High risk
- Moderate risk
- Low risk
- Neutral (i.e., not suicidal)

2.7 Feature Extraction & Embedding Techniques

Sentence embeddings have been explored in various studies [17]. This project intends to apply a variety of word embedding techniques, namely:

- BERT [6]
- SBERT [18] [19]
- Word2Vec

Contextual analysis will be done to achieve the following goals:

- Sentiment analysis
- Emotion detection
- Linguistic cues

3 Evaluation

Last but not least, throughout our experiment, we intend to evaluate how desirable and reliable our resulting model is. We also acknowledge the success of similar models in the domain of suicide detection [8] [22]. As a matter of fact, we intend to use metrics that are commonly used in other works in our evaluation, which align our results to standards. These are metrics in consideration:

- Overall Accuracy
- Precision
- Recall
- F1-score
- Area Under the Curve (AUC) from a Receiver Operating Characteristic (ROC) curve - for imbalanced datasets specifically

Ethical Explainability By visualizing our results as a conclusion of the project, we will also have to ensure our model's trustability. Maintaining the model's explainability is an important consideration in ethics. In this context, we intend to implement a simple algorithm which evaluates SHapley Additive exPlanations (**SHAP**) [11] values.

4 Expected Outcome and Impact

As a deliverable of the project, one could be a model capable of identifying suicide-related posts with high accuracy.

4.1 Broader Impacts

As an extension to the project, we do want to make a wider impact to the society and the academic research field.

Mental Health Contributions We sincerely hope our contributions being valuable to the field in mental health and ethics in AI, so that mental health professionals will be able to alleviate the problem of suicidal intentions being spread over the internet on time, before crisis occurs.

Domain Knowledge Further collaboration opportunities with mental health professionals and crisis support organizations will be desirable. Especially in the context of machine learning modeling, the more data annotated by domain experts are, the higher accuracy of our model it could achieve.

Visualization Potentially, when the model is deployed in use, a web-based application with a nicely looking user interface will beenfit real-time monitoring and alerts. It is however far beyond the scope of this course project.

5 Ethical Considerations

Besides explainability, to ensure ethical issues properly addressed, we intend to achieve the following goals:

- Data privacy and anonymization,
- Bias mitigation in model predictions, and
- Responsible AI deployment to avoid false positives and harmful consequences.

6 Timeline & Resources

Computational Requirements Throughout the project, we considered the high computational power, especially when we intend to train deep learning transformer-based neural networks. In case if we do not have a GPU which is powerful enough to train our classifiers within reasonable time locally, we will rely on an external low-cost cloud-based service such as Amazon Web Services (AWS) SageMaker [2], Kaggle, GitHub Codespaces, or Google Collab.

Week	Task
Week 1-2	Data collection and preprocessing
Week 3-4	Model selection and training
Week 5-6	Hyperparameter tuning and evaluation
Week 7	Ethical analysis and bias mitigation
Week 8	Finalizing results and documentation

Table 1: Project Timeline

Modeling Tools In order to gain an access to labelled datasets, handy NLP libraries and pre-trained models, we rely on HuggingFace and Kaggle mainly, with the help of the Natural Language Toolkit (NLTK).

Technologies Transformers like BERT have revolutionized NLP tasks [6]. To implement the layers and to train a neural network, we will make use of **TensorFlow** in Python.

7 Conclusion

This project aims to leverage advanced NLP techniques to identify suicidal intent from social media posts. By focusing on deep learning models, recent generative LLMs, and ethical considerations, we hope to contribute to mental health research and provide a tool that can assist in real-world suicide prevention efforts.

References

[1] Hosam Mhmd Ali. Twitter suicidal data, 2024. Accessed: 2025-02-20.

[2] Amazon Web Services. Amazon sagemaker, 2025. Accessed: 21-Feb-2025.

[3] Scott R. Braithwaite, Christophe Giraud-Carrier, Joshua West, Michael D. Barnes, and Cindy L. Hanson. Validating machine learning algorithms for twitter data against suicide rates: A feasibility study. *Biomedical Informatics Insights*, 10, 2018.

[4] CLPsych. Mental health nlp datasets, 2024. Accessed: 2025-02-20.

[5] DeepSeek-AI. DeepSeek-VL: Multimodal Large Language Model, 2024. Accessed: 2025-02-20.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 2018.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Open-sourcing bert: State-of-the-art pre-training for natural language processing, 2018. Accessed: 2025-02-20.

[8] Zein Hasan. Suicidal detection sentiment analysis, 2024. Accessed: 2025-02-21.

[9] Hugging Face. DistilBERT Model Documentation, 2025. Accessed: 21-Feb-2025.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[11] Scott Lundberg. Shap: Explainable ai and interpretability, 2024. Accessed: 2025-02-20.

[12] Mantis NLP. Supervised fine-tuning: Customizing llms, 2023. Accessed: 2025-02-20.

[13] E. Mohammadi et al. Detecting suicidal ideation in social media. *Journal of Artificial Intelligence Research*, 2020.

[14] OpenAI. GPT-4o: OpenAI's Most Advanced Model, 2024. Accessed: 2025-02-20.

[15] Kashish Parmar. Social media sentiments analysis dataset, 2024. Accessed: 2025-02-20.

[16] Senapati Rajesh. Depression tweets dataset, 2024. Accessed: 2025-02-20.

[17] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2019.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[19] Nils Reimers and Iryna Gurevych. Sentence-bert documentation, 2025. Accessed: 21-Feb-2025.

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[21] Behavioral Health Sciences and Oversight Accountability Commission (BHSOAC). Emerging best practices in suicide prevention, 2018. Accessed: 2025-02-20.

[22] Chau Pham Vy Nguyen. Leveraging large language models for suicide detection on social media with limited labels. *arXiv preprint arXiv:2410.04501*, 2024.