Ethical AI in Suicide Detection: A Multi-Model Analysis of Fairness & Bias

Kelvin Mock (300453668), Yifan Qin (300128779), Jing Hu (300252077), Shengchen Liu (300446349), and Jenifer Yu (300399089)
CSI5195 Ethics in AI Course Project, University of Ottawa

(Dated: February 21, 2025)

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. While AI-driven suicide detection models can provide timely interventions, they raise ethical concerns such as fairness, bias, and explainability. This project aims to develop an AI-powered suicide detection system utilizing state-of-the-art NLP models while ensuring fairness and mitigating bias. The goal is to improve predictive accuracy while maintaining ethical AI principles, reducing disparities in mental health predictions.

I. PROBLEM STATEMENT & MOTIVATION

Suicide detection using AI has gained attention as social media provides valuable real-time data for identifying individuals at risk [1]. for warning signs, ethical concerns arise regarding privacy, consent, and data security [2]. Users may not be aware their data is being used, and there is a risk of misuse, making ethical oversight essential.

To mitigate these 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 [3], GPT-based models [4], and DeepSeek [5]. 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.

Hence, there are a few research questions:

1. How do existing LLMs (e.g., BERT, GPT-4o, DeepSeek) exhibit bias in suicide detection, and how can it be mitigated?
2. What methods can be used to evaluate fairness in AI-based suicide risk prediction across different demographic groups?
3. How can explainability techniques such as SHAP analysis [6] 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?

II. DATASET / METHODOLOGY

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

- **CLPsych** [7]: Containing social media posts labeled for suicide risk.
- **Reddit SuicideWatch Posts**
- **Psychiatric Patient Records**
- **Social Media Sentiments Analysis** [8]: served as a corpus marked the sentiment of each piece of text.
- **Twitter Suicidal Data** [9]: A dataset with abundant suicidal texts best-suited for supervised fine-tuning (SFT) [10].
- **Depression Tweets** [11]: A JSON-structured dataset containing depression-related texts found on Twitter.

These datasets contain text-based user-generated content, allowing NLP models to assess risk levels. We will compare multiple NLP models, including:

1. **Baseline Model**: Decision Tree or Random Forest
2. **Deep Learning Models**
   - LSTM (Long Short-Term Memory)
   - BERT-based models [3](e.g., DistilBERT)
3. **LLM-based Models**
   - OpenAI GPT-based models [4]
   - DeepSeek [5]

III. EXPERIMENTS

- **Hyperparameter tuning**: learning rate, batch size, dropout rates.
- **Cross-validation**: to prevent overfitting.
- **Fine-tuning LLMs**: using domain-specific suicide-related text corpora.
- **SHAP (SHapley Additive exPlanations) Analysis** [6]: ensuring explainability for model predictions.

[1] S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, and C. L. Hanson, Validating machine learning algorithms for twitter data against suicide rates: A feasibility study, Biomedical Informatics Insights **10**, 10.1177/1178222618792860 (2018).

[2] B. H. Sciences and O. A. C. (BHSOAC), Emerging best practices in suicide prevention (2018), accessed: 2025-02-20.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Open-sourcing bert: State-of-the-art pre-training for natural language processing (2018), accessed: 2025-02-20.

[4] OpenAI, GPT-4o: OpenAI's Most Advanced Model (2024), accessed: 2025-02-20.

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

[6] S. Lundberg, Shap: Explainable ai and interpretability (2024), accessed: 2025-02-20.

[7] CLPsych, Mental health nlp datasets (2024), accessed: 2025-02-20.

[8] K. Parmar, Social media sentiments analysis dataset (2024), accessed: 2025-02-20.

[9] H. M. Ali, Twitter suicidal data (2024), accessed: 2025-02-20.

[10] Mantis NLP, Supervised fine-tuning: Customizing llms (2023), accessed: 2025-02-20.

[11] S. Rajesh, Depression tweets dataset (2024), accessed: 2025-02-20.