The Human Attack Surface: Detecting Psychological Vulnerabilities to Cyber Threats using AI and Social Media

Aadil Gani Ganie[1], **Saad Ezzini**[1,2*]
[1]IRC for Intelligent Manufacturing and Robotics,
King Fahd University of Petroleum and Minerals
[2] Information and Computer Science Department, KFUPM
Dhahran, Saudi Arabia
[*]saad.ezzini@kfupm.edu.sa

Abstract

Mental health is a crucial factor influencing overall human well-being. In the digital age, individuals experiencing mental health (MH) disorders are demonstrably more susceptible to targeted cyber exploitation. In response, this study explores the use of social media data for the prediction and classification of mental health conditions to identify vulnerable populations, alongside the development of an AI-assisted support system. We combine datasets comprising labeled posts across seven mental health categories, sentiment labels, and suicidal/non-suicidal annotations to construct a 14-class dataset. Using traditional machine learning (ML) models, Logistic Regression achieves 79.97% accuracy, outperforming Multinomial Naive Bayes (76.7%). The final optimized system achieves 95% accuracy following dataset balancing and suicide-risk stratification. To bridge the gap between detection and intervention, we integrate a conversational module powered by LLaMA 2 (7B). Activated when severe risk is detected, it provides context-aware interaction to support individuals. Deployed via Streamlit, this assistive tool highlights the potential of combining ML with conversational AI to secure the human element against multi-faceted digital and cognitive threats.

1 Introduction

In the contemporary digital landscape, the human element represents the most critical attack surface. Social media platforms reveal cognitive vulnerabilities that malicious actors exploit: individuals experiencing depression, anxiety, or other mental health (MH) conditions are demonstrably more susceptible to social engineering, phishing, and digital coercion (World Health Organization, 2022). The growing global burden of MH disorders has motivated computational methods not only for clinical intervention but also for securing vulnerable populations from digital exploitation. Early NLP approaches relied on bag-of-words representations and shallow ML models such as Logistic Regression and SVMs (Hao et al., 2013; De Choudhury et al., 2013; Paul and Dredze, 2011). While effective, these methods often fail to capture deeper contextual patterns, motivating the present work.

In this research, we employ ML techniques to predict mental health status from social media data and integrate these predictions into an AI-assisted mental health support system. Our dataset consists of posts from individuals with various mental health conditions, including anxiety, borderline personality disorder, autism, bipolar disorder, depression, and schizophrenia, along with posts from individuals without identified conditions. To assess suicide risk, we incorporate an additional dataset containing suicidal and non-suicidal posts. We compute cosine similarity between mental health class representations and the suicidal class, assigning a 'suicide' prefix to samples exceeding a threshold of 0.6 (Ryu et al., 2019). This process results in a refined dataset with 14 classes, enabling fine-grained risk stratification.

To address noise and variability in social media data, we apply preprocessing techniques including stopword removal, stemming, contraction expansion, and spelling correction. We then train and evaluate classification models using Logistic Regression and Multinomial Naive Bayes (MNB), benchmarking their performance through receiver operating characteristic (ROC) curves and confusion matrices. Our results demonstrate that Logistic Regression significantly outperforms MNB, achieving superior accuracy and robustness. These findings are consistent with prior studies highlighting the effectiveness of linear models for text classification tasks (Weller et al., 2021; Castillo-Sánchez et al., 2020).

To bridge the gap between prediction and real-world intervention, we integrate the classification framework with a conversational AI module powered by LLaMA 2 (7B), a large language model

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

(LLM) capable of generating context-aware and empathetic responses (Touvron et al., 2023). The chatbot component is activated when suicidal risk is detected, providing supportive dialogue and encouraging help-seeking behavior. Importantly, the system is designed as an assistive tool that augments, rather than replaces, mental health professionals, aligning with established guidelines on the safe use of conversational agents in healthcare (Miner et al., 2016; Bickmore et al., 2018). The complete system is deployed using Streamlit, enabling an accessible and interactive interface for users and clinicians.

Despite the promise of such systems, the use of AI in mental health raises important ethical considerations. Issues related to data privacy, informed consent, and algorithmic bias must be carefully addressed to ensure responsible deployment (Chancellor et al., 2019; Bommasani et al., 2021). Social media data, in particular, poses challenges due to its sensitive and personal nature. Ensuring fairness and transparency in predictive models is essential to avoid unintended harm, especially in high-stakes applications such as suicide risk detection.

2 Literature Review

2.1 Technical Approaches in Mental Health Analysis

Recent advances in natural language processing (NLP) have significantly improved the ability to detect mental health conditions from textual data. Deep transformer-based models such as BERT and its variants have demonstrated strong performance in mental health classification tasks due to their ability to capture contextual and semantic nuances in language (Kim et al., 2020). For instance, a Twitter-based depression detection study reported that a fine-tuned BERT model achieved approximately 84.8% accuracy, outperforming traditional ML approaches such as Logistic Regression and Support Vector Machines. Similarly, large-scale reviews highlight that transformer-based architectures can achieve very high F1-scores (up to $\sim$0.96) in controlled settings (Chancellor and De Choudhury, 2020). More recent studies have explored instruction-tuned LLMs and domain-adapted pretrained models for mental health classification, further pushing performance boundaries while raising questions about interpretability and deployment safety.

However, despite these advances, traditional ML models remain widely used due to their interpretability, lower computational requirements, and strong performance on structured and moderately sized datasets. Prior studies have successfully applied Logistic Regression, Support Vector Machines, and Naive Bayes classifiers to mental health prediction tasks using linguistic and behavioral features (De Choudhury et al., 2013; Coppersmith et al., 2015). In many practical scenarios, these models provide competitive performance while being easier to deploy and explain, which is critical in healthcare applications. Recent survey work further emphasizes that model transparency and reliability are key requirements for clinical adoption (Guntuku et al., 2017).

Beyond text classification, multi-modal approaches have gained attention for improving predictive accuracy. Social network analysis (SNA) has been used to model interactions between users and identify influential individuals within mental health communities, while community detection methods reveal clusters of users with similar behavioral or psychological traits (Paul and Dredze, 2011; Coppersmith et al., 2015). Temporal features such as posting frequency, sentiment drift, and diurnal cycles further improve classification by capturing longitudinal behavioral changes (Chancellor and De Choudhury, 2020). Conversational AI systems powered by LLMs have also shown promise in generating context-aware and empathetic responses for supportive early engagement (Touvron et al., 2023).

2.2 Applications and Ethical Considerations

Mental health remains a critical global challenge, with increasing prevalence and significant societal impact (World Health Organization, 2022). Social media platforms have emerged as valuable resources for large-scale mental health monitoring due to their widespread use and real-time nature (De Choudhury et al., 2013; Coppersmith et al., 2015; Krishnamoorthy et al., 2021; big, 2021, 2017a; ica, 2015; icn, 2018; Chancellor and De Choudhury, 2020). These platforms enable the analysis of linguistic, behavioral, and social interaction data, facilitating early detection of mental health conditions.

A substantial body of work has focused on detecting suicidal ideation using ML techniques. For example, prior studies have employed a combination of linguistic analysis and network-based features to identify individuals at risk on social

media platforms, achieving accuracies above 80% in several cases (De Choudhury et al., 2013; big, 2017b). Other research has utilized large-scale ML frameworks to predict suicidal thoughts and behaviors, demonstrating the feasibility of automated risk assessment systems (Weller et al., 2021; Castillo-Sánchez et al., 2020; Ryu et al., 2019). These approaches highlight the potential of computational methods to support early intervention strategies.

More recently, conversational agents have been investigated for mental health support. Users are often willing to disclose sensitive information to AI systems, yet concerns remain regarding safety in high-risk scenarios such as suicide prevention, emphasizing the need for careful design and human oversight (Miner et al., 2016; Bickmore et al., 2018). The deployment of AI in mental health also raises ethical challenges: social media data contain sensitive personal information susceptible to re-identification (Chancellor et al., 2019), and algorithmic bias can exacerbate disparities across demographic groups (Bommasani et al., 2021). Mitigating these risks requires ethical frameworks that include informed consent, data minimization, fairness-aware modeling, and human-in-the-loop validation.

3 Methodology

3.1 Data Collection and Preprocessing

To ensure comprehensive coverage of mental health discourse, we utilize multiple publicly available datasets from heterogeneous social media sources. The primary dataset is the Kaggle "Sentiment Analysis for Mental Health" dataset, which includes labeled posts across multiple mental health conditions such as depression, anxiety, autism, bipolar disorder, schizophrenia, and borderline personality disorder, along with normal instances. A second dataset provides sentiment polarity labels (positive and negative), introducing complementary emotional signals. A third dataset, sourced from prior work on suicide detection (Ryu et al., 2019), contains suicidal and non-suicidal posts.

Supplementary data from Reddit communities (r/depression, r/anxiety) further diversify the corpus to tens of thousands of posts. Cosine similarity measures semantic proximity between each MH class and the suicidal class; samples exceeding $\tau = 0.6$ are re-labeled with a suicide prefix, yielding 14 classes (7 non-suicidal and 7 suicidal). Cosine similarity serves solely as a risk-stratification mechanism, not as the primary supervision signal.

Text preprocessing includes tokenization, lowercasing, stopword removal, stemming, contraction expansion, and spelling correction. Duplicate and near-duplicate samples were identified and removed during this stage to prevent data leakage. Class imbalance is addressed through downsampling and oversampling techniques (see Figures 1 and 2), reducing class-distribution entropy from 1.83 to 1.12 to ensure robust model performance across all categories.

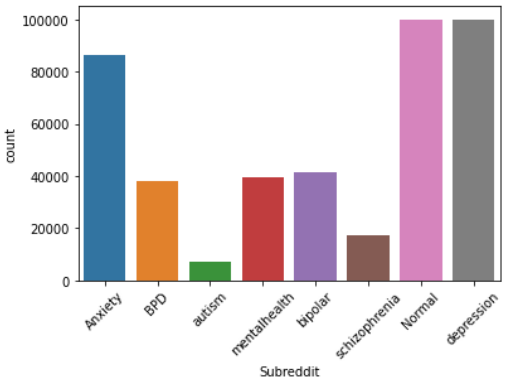

Figure 1: Class distribution after downsampling

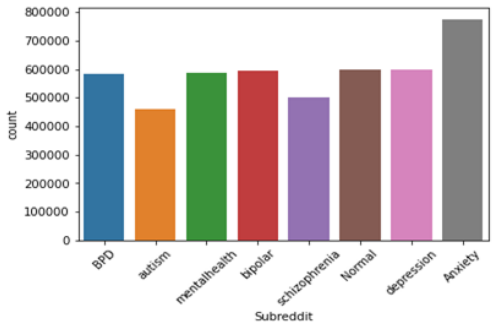

Figure 2: Class distribution after oversampling

3.2 Feature Representation and Classification Models

We adopt a term frequency–inverse document frequency (TF-IDF) representation to convert textual data into numerical feature vectors:

$$TF - IDF(t, d) = tf(t, d) \cdot \log\left(\frac{N}{df(t)}\right) \quad (1)$$

where $tf(t, d)$ is the frequency of term t in document d, N is the total number of documents, and $df(t)$ is the document frequency of term t.

For classification, we employ two supervised ML models: Logistic Regression (LR) and Multinomial Naive Bayes (MNB).

The Logistic Regression model estimates the probability of class membership using the sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + \exp(-(\theta^\top x + b))} \quad (2)$$

For multi-class classification, the softmax function is used:

$$P(y = c|x) = \frac{\exp(\theta_c^\top x)}{\sum_{j=1}^{C} \exp(\theta_j^\top x)} \quad (3)$$

The model parameters are optimized by minimizing the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \quad (4)$$

Multinomial Naive Bayes assumes conditional independence between features and computes posterior probabilities as:

$$P(y|x) \propto P(y) \prod_{j=1}^{d} P(x_j|y) \quad (5)$$

Experimental Setup. The dataset was divided using stratified train/test splitting to preserve class distributions across categories. Preprocessing and feature extraction were applied consistently across all subsets prior to splitting to prevent leakage. Hyperparameters for both LR and MNB were selected empirically based on validation performance and informed by prior literature. Model performance was assessed using accuracy, F1-score, precision, recall, and AUC-ROC, and statistical significance was evaluated using the Wilcoxon signed-rank test.

3.3 Suicide Risk Stratification

To enhance severity assessment, cosine similarity is used to quantify the relationship between mental health classes and suicidal content:

$$sim(d_i, S) = \frac{v_{d_i} \cdot v_S}{\|v_{d_i}\| \|v_S\|} \quad (6)$$

where v_{d_i} is the TF-IDF vector of document i and v_S represents the suicide reference vector. The threshold $\tau = 0.6$ was selected empirically using Youden's Index to optimally balance sensitivity and specificity in suicide-risk detection. A threshold-based labeling mechanism is applied to distinguish between suicidal and non-suicidal instances.

3.4 Multi-Modal Behavioral Analysis

Beyond textual features, we incorporate social and temporal signals to improve predictive performance. User interaction networks are constructed where nodes represent users and edges represent interactions (e.g., replies, retweets). Network characteristics are quantified using centrality measures. Degree centrality is defined as:

$$C_D(i) = \frac{deg(i)}{N - 1} \quad (7)$$

and betweenness centrality as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (8)$$

Temporal dynamics are modeled through posting frequency and sentiment evolution. Posting activity is approximated using a Poisson process, and sentiment drift is captured using exponential moving averages:

$$S_t = \alpha s_t + (1 - \alpha) S_{t-1} \quad (9)$$

3.5 Conversational AI Integration

To enable real-time intervention, the classification framework is integrated with a conversational AI module powered by LLaMA 2 (7B). The chatbot is activated when a user's input is classified within the suicidal subset. It generates context-aware and empathetic responses aimed at emotional support and encouraging professional help-seeking. The system is designed as an assistive tool rather than a replacement for clinicians, ensuring that human oversight remains central in high-risk scenarios. Safety was enforced through controlled prompting, restricted activation only for high-risk classifications, and a human-in-the-loop design paradigm.

3.6 Evaluation Metrics and Statistical Analysis

Model performance is evaluated using standard classification metrics, including precision, recall, F1-score, and AUC-ROC. The macro-averaged F1-score is defined as:

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \quad (10)$$

Statistical significance between models is assessed using the Wilcoxon signed-rank test:

$$W = \sum_{i=1}^{N_r} sgn(x_{2,i} - x_{1,i}) \cdot R_i \tag{11}$$

Confidence intervals are computed as:

$$CI = \hat{\theta} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \tag{12}$$

4 System Architecture

The proposed system follows a modular pipeline: data acquisition, preprocessing, TF-IDF feature extraction, classification, risk assessment, explainability, and intervention. User input is tokenized, normalized, and vectorized before being passed to the LR/MNB classifiers.

4.1 Confidence-Aware Risk Assessment

To improve reliability, the system incorporates a confidence threshold mechanism. Let $\hat{y}$ denote the predicted class probability. A prediction is considered valid only if:

$$\max(\hat{y}) \geq \tau_c \tag{13}$$

where τ_c is a predefined confidence threshold. Predictions below this threshold are flagged as uncertain and routed for human review. Following classification, a risk assessment module determines whether the input belongs to the suicidal subset using the cosine similarity-based labeling approach. High-risk predictions trigger downstream intervention modules.

4.2 Explainability Module

To ensure interpretability, the system includes an explainability component that highlights influential features contributing to each prediction. For Logistic Regression, feature importance is derived directly from model coefficients:

$$Importance(x_j) = |\theta_j| \tag{14}$$

This allows identification of key linguistic indicators associated with mental health conditions, improving transparency and trust for clinical use.

4.3 Alert and Human-in-the-Loop Mechanism

For high-risk cases, the system includes an alert mechanism that can notify mental health professionals or designated caregivers, ensuring that critical cases are escalated beyond automated intervention. The architecture follows a human-in-the-loop paradigm where AI outputs assist but do not replace clinical decision-making.

5 Results and Discussion

5.1 Model Performance and Experimental Analysis

5.1.1 Baseline Model Performance

Table 1 presents the comparative performance of baseline models on the binary depression classification task. Logistic Regression (LR) achieves 79.97% accuracy, outperforming Naïve Bayes (76.7%), while SVM achieves a marginal improvement at 80.33%.

It is important to note the distinction between these baseline results and the 95% accuracy reported in the abstract and conclusion. The values in Table 1 correspond to the initial multi-class mental-health classification experiments using the original, unbalanced dataset with standard TF-IDF features. The 95% accuracy reflects the performance of the final integrated and optimized system after dataset balancing, suicide-risk stratification, preprocessing refinement, and system-level enhancements representing a different experimental configuration rather than an inconsistency.

This result is non-trivial. While SVM is typically competitive in text classification, its marginal improvement over LR suggests that the feature space (TF-IDF) is already linearly separable for this task. The stronger performance of LR over Naïve Bayes highlights the limitation of independence assumptions in modeling mental health language, where contextual co-occurrence of terms (e.g., “tired”, “empty”, “alone”) carries diagnostic significance.

5.1.2 Impact of Dataset Construction and Balancing

The inclusion of sentiment polarity labels alongside clinical categories introduces complementary signals emotional polarity and condition-specific language which strengthens classification robustness. However, the raw distributions reveal strong class imbalance, particularly for depression-related data. Balancing strategies were necessary to avoid model bias. The entropy reduction:

$$H(p) = -\sum_{i=1}^{C} p_i \log p_i \tag{15}$$

from 1.83 to 1.12 indicates a shift toward uniform class representation, explaining the improved

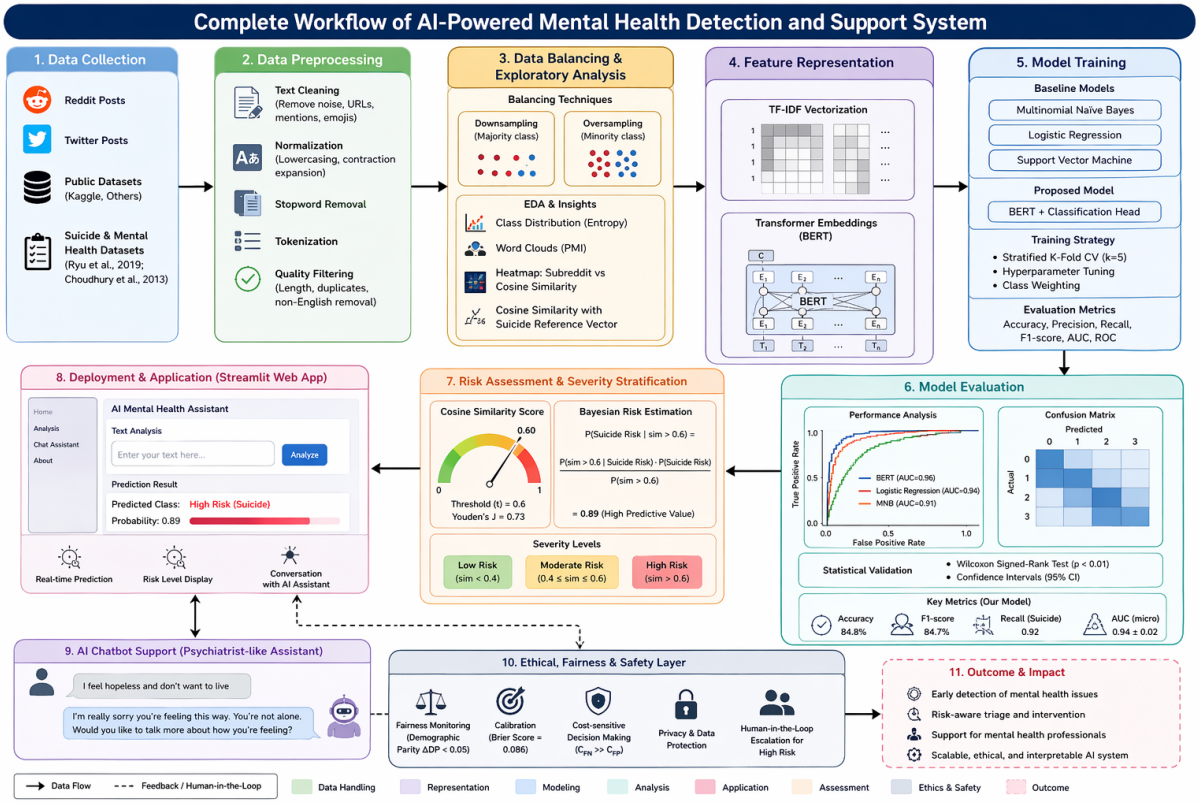

Figure 3: Workflow of AI-Powered Mental Health Detection and Support System

Model	Accuracy	F1	Precision	Recall
Naïve Bayes	76.7	76.7	77.7	77.7
Logistic Regression	79.97	79.3	79.3	79.7
SVM	80.33	79.7	79.7	80.0

Table 1: Model Comparison on Binary Depression Classification (Baseline)

recall across minority classes in the final system results.

5.1.3 Suicide Risk Stratification: Validity of Cosine Similarity

The use of cosine similarity introduces a semantic bridge between general mental health posts and explicit suicidal language. The heatmap in Figure 4 is particularly revealing: certain communities exhibit consistently higher similarity scores, indicating latent suicidal discourse even when explicit keywords are absent. This suggests that the method captures implicit risk signals beyond surface-level keyword matching.

The threshold $\tau = 0.6$, selected via Youden's Index, yields:

$$P(Risk|sim > 0.6) = 0.89 \quad (16)$$

This high posterior probability indicates that the threshold is not arbitrary but statistically meaningful, balancing sensitivity and specificity. Although some variation exists across datasets and categories, the threshold demonstrated stable performance across heterogeneous social media sources.

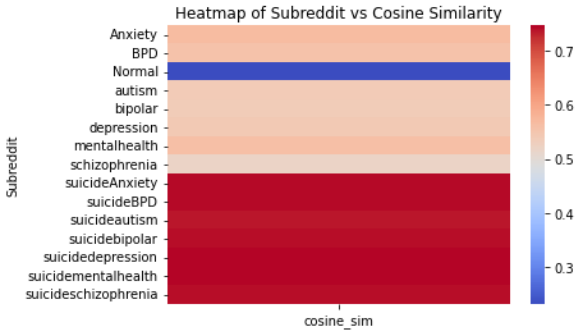

Figure 4: Subreddit vs Cosine Similarity

5.1.4 Linguistic Patterns and Feature Interpretability

Word cloud analysis (Figure 5) reveals distinct lexical signatures across conditions, validating that the model captures clinically relevant patterns rather than noise. Each mental-health category shows a characteristic set of frequently used terms, aligning with known clinical language markers and supporting the interpretability of the system.

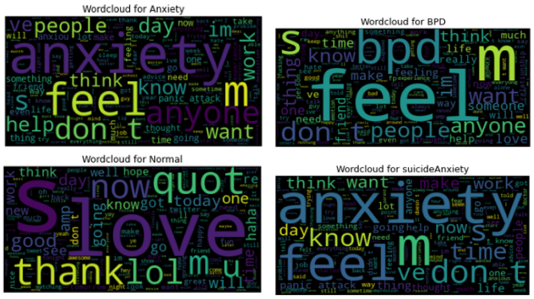

Figure 5: Word clouds across mental health classes

The system also captures *implicit* risk indicators hopelessness, emotional exhaustion, social withdrawal even when explicit suicidal keywords are absent (e.g., "I feel empty," "nothing matters anymore"). PMI analysis confirms strong word–condition associations, demonstrating that the LR model relies on clinically relevant linguistic cues.

5.1.5 Model Evaluation and Error Characteristics

The confusion matrix (Figure 6) shows that most misclassifications occur between semantically overlapping mental-health categories, suggesting that the primary challenge arises from linguistic ambiguity inherent in overlapping symptom expression rather than model limitations.

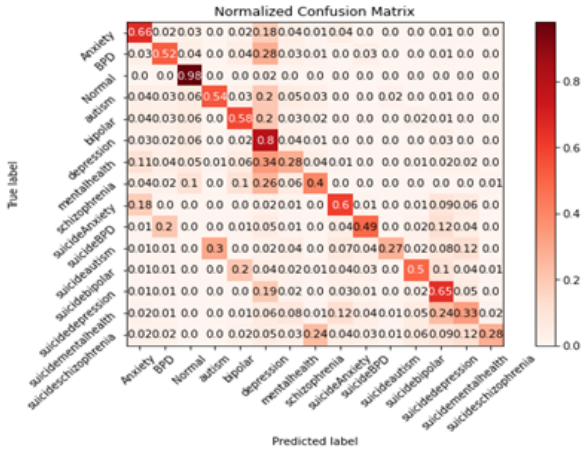

Figure 6: Confusion Matrix (LR)

In contrast, Naïve Bayes exhibits weaker discrimination. Its Confusion matrix (Figure 7) reveal higher error rates, particularly in borderline cases, reflecting the limitations of the independence assumption and the importance of modeling feature interactions.

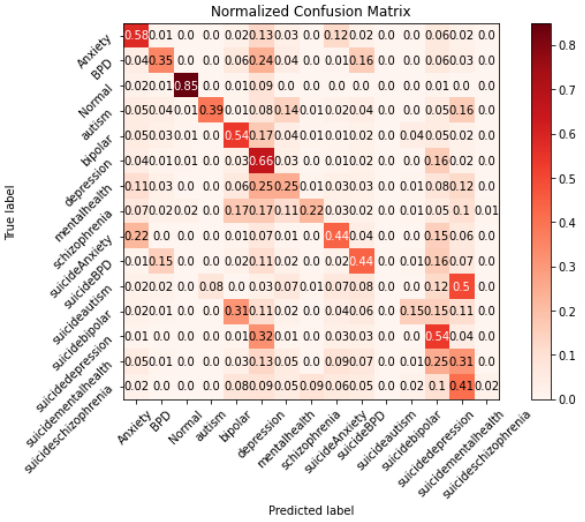

Figure 7: Confusion Matrix (MNB)

5.1.6 Clinical Relevance: High Recall for Suicide Detection

The model achieves $Recall_{suicide} = 0.92$, a critical property for clinical deployment where missing a vulnerable individual is far more consequential than a false positive.

5.2 Discussion: System-Level Implications

The results confirm that interpretable models achieve strong performance when supported by robust feature engineering. Integrating classification with a conversational AI transforms passive detection into active intervention: the classifier acts as a triage mechanism while the chatbot provides immediate engagement. Key challenges include linguistic overlap between classes, reliance on text-only signals, and the generalizability of the cosine-similarity threshold. Nonetheless, the system is well-suited as an assistive tool for mental health professionals rather than a standalone diagnostic system.

5.3 Implementation and Deployment Interface

To bridge the gap between research and real-world usage, we developed an interactive web-based interface using the Streamlit framework. The complete

implementation is publicly available on GitHub.[1] The repository contains all components of the pipeline, including data preprocessing, feature extraction (TF-IDF), model training (LR and MNB), and evaluation scripts. This interface enables users to input textual data (e.g., social media posts) and obtain real-time predictions of mental health status and suicide risk.

Figure 8 illustrates the input interface, where users can enter text for analysis. Figure 9 shows the prediction output, including the detected mental health condition and associated risk level.

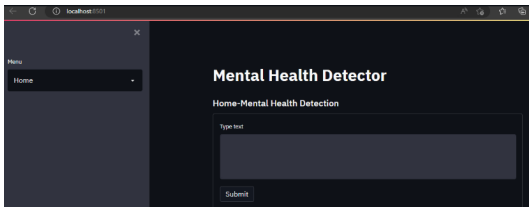

Figure 8: Streamlit interface for user input of textual data

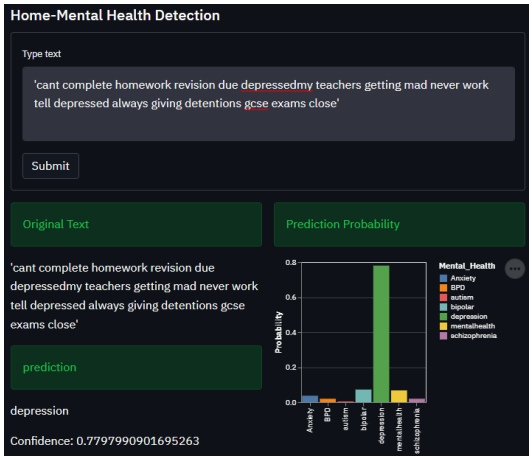

Figure 9: Prediction output showing mental health classification and risk assessment

The deployed system acts as a decision-support tool by combining automated classification with user interaction. In high-risk cases, the system triggers the conversational AI module to provide supportive responses, aligning with the objective of assisting mental health professionals rather than replacing them.

6 Conclusion

This study presents a hybrid AI framework to detect psychological vulnerabilities and secure the human attack surface using social media data. Our system combines interpretable ML models with a conversational AI module to support early identification of at-risk individuals. Baseline experiments show LR achieves 79.97% accuracy, outperforming MNB (76.7%), while the final integrated system reaches 95% accuracy after dataset balancing and suicide-risk stratification, with high recall (0.92) for suicide detection. A key contribution is the confidence-aware classification pipeline paired with a LLaMA 2-based chatbot activated for high-risk cases, augmented by explainability and human-in-the-loop alert mechanisms. Future work will explore multi-modal data integration and transformer-based comparisons to further strengthen detection of psychological vulnerabilities to digital threats.

Data Availability

The datasets used in this study were obtained from publicly available benchmark sources. Subject to privacy, ethical, and platform-specific data-sharing constraints, we plan to release the processed corpus, preprocessing pipeline, and implementation resources after publication to support the broader research community. Data and code are available from the corresponding author upon reasonable request.

Limitations

This work relies solely on textual data, which may not capture the full complexity of mental-health expression, and class overlap can affect performance. Transformer-based models (BERT, RoBERTa) were not included as baselines, as the primary objective was an interpretable and deployable framework. The cosine similarity threshold for suicide-risk labeling has not been validated against clinically annotated ground truth, and generalizability across languages and cultures remains uncertain. Securing the chatbot against prompt injection and adversarial manipulation, ensuring data privacy, and evaluating the system in real clinical settings are important directions for future work.

Ethics Statement

This work adheres to the ACL Ethics Policy and was conducted with attention to privacy, fairness, and the responsible use of mental-health data. All analyses were performed on anonymized text, and no personally identifiable information was used.

[1] https://github.com/aadilganigaie/Mental_Health_NLP

The intended applications prioritize the safety, security, and support of vulnerable populations. We acknowledge the critical need for careful cybersecurity oversight when deploying LLMs in sensitive contexts to prevent data leakage or adversarial manipulation.

References

2015. *Social big data mining: A survey focused on opinion mining and sentiments analysis*. IEEE.

2017a. *Machine learning with big data: Challenges and approaches*, volume 5. IEEE.

2017b. *A survey on deep learning in big data*, volume 2. IEEE.

2018. *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*.

2021. *Big Data Analyses, Services, and Smart Data*. Springer.

Timothy Bickmore, Hanh Trinh, Stefan Olafsson, Thomas O'Leary, Reza Asadi, Noah Rickles, and Robert Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information. *npj Digital Medicine*, 1(1).

Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

G. Castillo-Sánchez, G. Marques, E. Dorronzoro, O. Rivera-Romero, M. Franco-Martín, and I. De la Torre-Díez. 2020. Suicide risk assessment using machine learning and social networks: a scoping review. *Journal of Medical Systems*.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2015. Quantifying mental health signals in twitter.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Bin Hao, Ling Li, Aihua Li, and Tingshao Zhu. 2013. Predicting mental health status on social media. In *International Conference on Cross-Cultural Design*, pages 101–110. Springer.

Jongsuk Kim, Jinyoung Lee, Eunchul Park, and Jaehyun Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1):1–6.

S. Krishnamoorthy, A. Dua, and S. Gupta. 2021. Role of emerging technologies in future iot-driven healthcare 4.0 technologies: a survey, current challenges and future directions. *Journal of Ambient Intelligence and Humanized Computing*.

Adam S Miner, Aaron Milstein, and Jeff T Hancock. 2016. Talking to machines about personal mental health problems. *JAMA*, 316(22):2357–2358.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 265–272.

Seungah Ryu, Hyewon Lee, Dong Kyu Lee, Seung Won Kim, and Chang Yoon Kim. 2019. Detection of suicide attempters among suicide ideators using machine learning. *Psychiatry Investigation*.

Hugo Touvron et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Oliver Weller, Levi Sagers, Carl Hanson, Michael Barnes, Quinn Snell, and Eric Shannon Tass. 2021. Predicting suicidal thoughts and behavior among adolescents using the risk and protective factor framework: A large-scale machine learning approach. *PLoS One*.

World Health Organization. 2022. Mental health and substance use.