

A Machine Learning Approach for Detection of Mental Health Conditions and Cyberbullying from Social Media

Edward Ajayi¹, Martha Kachweka¹, Mawuli Deku¹, Emily Aiken¹

¹Carnegie Mellon University Africa
Kigali, Rwanda
{eaajayi, mkachwek, mdeku, eaiken}@andrew.cmu.edu

Abstract

Mental health challenges and cyberbullying are increasingly prevalent in digital spaces, necessitating scalable and interpretable detection systems. This paper introduces a unified multiclass classification framework for detecting ten distinct mental health and cyberbullying categories from social media data. We curate datasets from Twitter and Reddit, implementing a rigorous 'split-then-balance' pipeline to train on balanced data while evaluating on a realistic, held-out imbalanced test set. We conduct a comprehensive evaluation comparing traditional lexical models, hybrid approaches, and several end-to-end fine-tuned transformers. Our results demonstrate that end-to-end fine-tuning is critical for performance, with the domain-adapted MentalBERT emerging as the top model, achieving an accuracy of 0.92 and a Macro F1 score of 0.76, surpassing both its generic counterpart and a zero-shot LLM baseline. Grounded in a comprehensive ethical analysis, we frame the system as a human-in-the-loop screening aid, not a diagnostic tool. To support this, we introduce a hybrid SHAP-LLM explainability framework and present a prototype dashboard ("Social Media Screener") designed to integrate model predictions and their explanations into a practical workflow for moderators. Our work provides a robust baseline, highlighting future needs for multi-label, clinically-validated datasets at the critical intersection of online safety and computational mental health.

NLP, mental health, cyberbullying

Code and Datasets are available —

https://github.com/codes2425/mental_health_detect

Introduction

Mental health disorders are a growing global concern, affecting one in eight people worldwide (World Health Organization 2025). Common mental health conditions like anxiety, depression, bipolar disorder, and stress-related illnesses contribute substantially to the global burden of disease. Despite the availability of effective prevention and treatment options, access to mental health care remains a major challenge, exacerbated by stigma, discrimination, and inadequate resources (World Health Organization 2025).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Social media platforms contribute to and mediate mental health conditions in an increasingly digital world. While social media sites have been shown to increase connection in some settings (Zsila and Reyes 2023), they have also amplified mental health risks by exposing users to cyberbullying (Naslund et al. 2020), emotionally charged content (Poddar et al. 2024), and negative sentiment (Wu et al. 2025). Social media also provides a valuable opportunity for scalable screening and content moderation, with the possibility of automated detection of signs of distress through natural language processing (NLP) techniques (Mobin et al. 2024). While existing research has made substantial progress in predicting mental health conditions from social media posts, most studies focus on binary classification (e.g., classifying posts as depression or not depression) (Kumar et al. 2022), overlooking the interconnected nature of mental health conditions (World Health Organization 2025). Moreover, existing research often focuses on narrow subsets of harmful content, failing to address the diverse range of mental health expressions and cyberbullying behaviors encountered online.

This study aims to fill this gap by developing and evaluating a unified multiclass classification framework for detecting ten distinct categories of mental health conditions and cyberbullying from public social media data. We aggregate datasets from Reddit and Twitter, comparing lexical (TF-IDF) and contextual (e.g., BERT) modeling approaches. We explicitly frame this framework not as a diagnostic tool, but as a human-in-the-loop screening aid intended for trained moderators. To support this practical integration, we introduce a hybrid SHAP-LLM explainability system and present a prototype dashboard ("Social Media Screener") to visualize how our model's outputs can be safely and transparently operationalized.

We propose a dual-purpose application: (1) as a content flagging tool for acute-risk classes like 'Suicide' that require urgent human review, and (2) as a component in longitudinal analysis tools to help practitioners identify linguistic patterns over time for nuanced conditions like 'Bipolar Disorder'. This human-in-the-loop framework serves as a direct parallel to clinical support systems, where similar AI tools can assist therapists in reviewing high volumes of patient-generated data. Our evaluation, conducted on a held-out, imbalanced test set, demonstrates the effectiveness of end-to-end fine-tuning, with the domain-adapted MentalBERT (Ji

et al. 2022) showing clear superiority. By integrating these technical findings with a concrete and ethically-grounded application, this study provides a comprehensive methodological baseline. Our analysis also highlights the critical limitations of existing public, weakly-labeled, single-label datasets, demonstrating the urgent need for future work in developing multi-label benchmarks sourced from consented, clinically-assessed cohorts for computational mental health.

Related Work

Mental Health Detection from Social Media

A large and growing body of research leverages natural language processing (NLP) to identify mental health conditions from social media platforms such as Twitter and Reddit (Mobin et al. 2024; Poddar et al. 2024; Nweke, Khan, and Pei 2024; Abdullah and Negied 2024; de Arriba-Pérez and García-Méndez 2024). Early studies often focused on binary classification tasks distinguishing between depressed and non-depressed users (Kumar et al. 2022). More recent papers have incorporated deep learning to capture nuanced emotional patterns (Jiang 2021). Transformer-based architectures, particularly BERT (Devlin et al. 2019a), have become the de facto standard for text classification due to their capacity to capture contextual dependencies and semantics at a fine-grained level. (Ji et al. 2022) introduced MentalBERT and MentalRoBERTa, domain-adapted versions of BERT fine-tuned on Reddit mental health forums, showing notable improvements in classifying user mental health status. However, most of these studies rely on label sets with limited numbers of classes, failing to capture the complex and interrelated nature of mental health conditions in the real world.

Cyberbullying Detection

Cyberbullying detection on social media has also received considerable attention from researchers in the recent years (Mathew et al. 2020; Antypas and Camacho-Collados 2023). Most relevantly to our work, (Wang, Fu, and Lu 2020) developed SOSNet, a domain-specific neural network architecture for distinguishing types of cyberbullying such as those based on gender, ethnicity, and religion. We build on the dataset curated by (Wang, Fu, and Lu 2020) in this paper.

Data and methods

Data

This study aggregates a number of datasets from Twitter and Reddit to analyze and classify various mental health conditions, including suicidal ideation, personality disorders, stress, anxiety, and bipolar disorder. We also work with data on cyberbullying data from Twitter. All datasets were sourced from Kaggle.

Mental health data The mental health data are sourced from various social media platforms, primarily Twitter and specific subreddits dedicated to mental health conversations (Sarkar 2024; Ghoshal 2025). The dataset is categorized into four conditions: anxiety, stress, bipolar disorder, and personality disorder, totaling 53,043 posts. The anxiety data was

gathered from Facebook and Twitter and subsequently manually annotated by four undergraduate English-speaking students. The data for bipolar disorder and personality disorder were collected from their respective subreddits. Similarly, the stress data was sourced from relevant subreddits; however, the specific annotation guidelines for this subset were not detailed by the original authors.

Cyberbullying data The cyberbullying dataset was adopted from (Wang, Fu, and Lu 2020), who developed a model to detect cyberbullying based on age, gender, ethnicity, and religion on Twitter. The dataset labels each post as one of five types of cyberbullying (gender, religion, ethnicity, age, other) or as not containing cyberbullying content. The fine-grained labels were initially created through a manual annotation process on a subset of the data. To expand the dataset, the authors (Wang, Fu, and Lu 2020) then employed a semi-supervised method, Dynamic Query Expansion (DQE), to increase the number of samples for each class. The final curated dataset contains a total of 47,692 posts.

Suicide and depression detection data Our third and final dataset focuses on detecting depression and suicidal ideation, with data collected from the SuicideWatch and depression subreddits (Komati 2021). The dataset also includes posts from the teenagers subreddit for non-suicidal and non-depression content. The dataset includes 232,074 posts in total, each labeled as suicidal or non-suicidal content.

Data cleaning All text entries were standardized to lowercase strings. Unwanted elements such as URLs, user mentions, and non-alphanumeric characters were removed using regular expressions, and extraneous whitespace was stripped to ensure data uniformity.

Dataset Curation, Splitting, and Balancing To ensure a robust evaluation and prevent data leakage, we implemented a strict "split-then-balance" pipeline. First, all ten datasets were merged into a master dataset of 274,150 posts. We then performed post-level deduplication to prevent identical posts from appearing in both training and test sets. After deduplication, the dataset was split into a training pool (80%) and a held-out test pool (20%) using a stratified split to preserve the original, imbalanced class distribution in both pools. The final test set was sampled from the held-out test pool and retained its original, highly imbalanced distribution to reflect real-world data.

The final training set was constructed exclusively from the 80% training pool using a multi-step balancing process. This process, summarized in Table 1, involved downsampling (DS) high-resource classes (e.g., Non-Suicide, Suicide) and applying deduplication (DD) and EDA-based oversampling (a technique that generates synthetic text by applying one of four random operations: synonym replacement, random insertion, random swap, or random deletion) (Wei and Zou 2019) to low-resource classes (e.g., Personality Disorder, Stress). Table 1 shows the class distribution after each step and the final test set.

Class	Before DS	After DS	After EDA & DD	Final Test Set
Age CB	7,992	2,400	2,400	175
Anxiety	3,841	2,400	2,400	80
Bipolar	2,777	2,001	2,400	55
Ethnicity CB	7,955	2,400	2,400	172
Gender CB	7,916	2,400	2,400	167
Non-Suicide	115,983	2,400	2,400	2,549
Personality Disorder	1,077	714	2,387	20
Religion CB	7,997	2,400	2,400	175
Stress	2,585	1,832	2,396	50
Suicide	116,027	2,400	2,400	2,557

Table 1: Class distribution across training preprocessing steps and final test set. DS = Downsampling, DD = Deduplication, CB = Cyberbullying.

Data Label Verification

To evaluate the reliability of the dataset’s weak labels, we performed a manual annotation study on a randomly selected subset of 300 posts (1% of the dataset), balanced across all ten classes. Two authors independently annotated the posts while blinded to the original labels. Inter-annotator agreement achieved a Cohen’s Kappa of $\kappa = 0.76$, indicating substantial consistency between annotators. We further compared each annotator’s labels with the original dataset labels, yielding Cohen’s Kappa scores of $\kappa = 0.71$ and $\kappa = 0.94$, respectively. These results demonstrate that the weak labels are highly aligned with human judgment, supporting the reliability of the dataset for our experiments.

Featurization

We experiment with two methods to extract numerical features from the preprocessed text:

- **TF-IDF Vectorization:** The Term Frequency-Inverse Document Frequency (TF-IDF) approach (Nweke, Khan, and Pei 2024; Robertson 2004) was implemented using the `TfidfVectorizer` from scikit-learn, configured to extract the top **5000** features. This transformation generated a sparse matrix of TF-IDF scores, emphasizing words that are important relative to their document frequency across datasets. We clarify that stopwords were removed only for exploratory lexical analysis, for all model training (both TF-IDF and BERT embeddings-based), stopwords were kept to preserve full semantic context and ensure a fair comparison.
- **BERT Embeddings:** For deeper semantic representation, BERT embeddings (Zhang and Liu 2024) were obtained using the `bert-base-uncased` model from the Hugging Face Transformers library. Each text sample was tokenized with a maximum length of 128 tokens.

These feature extraction strategies allowed the study to experiment with both lexical (TF-IDF) and contextual (BERT) representations.

Machine learning approaches

We test several distinct families of machine learning models for our 10-class classification task. All models were trained

on the balanced training set (Table 1) and evaluated on the held-out, imbalanced test set. Specific hyperparameters (Appendix) were selected based on standard practices.

Lexical Baseline Models To establish a non-contextual baseline, we trained two classical models on TF-IDF features. As described in Section 3.2, we used the top 5000 features (unigrams and bigrams), with stopwords preserved.

- **TF-IDF + Logistic Regression:** We used scikit-learn’s `LogisticRegressionCV` with the `saga` solver, a multinomial setting, and `cv=5` (5-fold cross-validation) on the training set to select the best regularization strength.
- **TF-IDF + Support Vector Machine (SVM):** We used scikit-learn’s `GridSearchCV` with `cv=5` to find the optimal regularization parameter for a linear SVC class, maximizing for macro F1 score.

Static Embedding Baseline Models To test the performance of static contextual embeddings, we fed pre-computed BERT embeddings into classical and neural classifiers. For these models, sentence-level embeddings were extracted using the CLS token from the `bert-base-uncased` model.

- **BERT Embeddings + Logistic Regression:** The static CLS embeddings were fed into the same `LogisticRegressionCV` model used in the lexical baseline to ensure a fair comparison.
- **BERT Embeddings + RNN:** A simple Recurrent Neural Network (RNN) with one hidden layer of 128 units was built using PyTorch. The model was trained for five epochs using the Adam (Kingma and Ba 2017) optimizer (learning rate = 0.001) on the static embeddings.
- **BERT Embeddings + DNN:** A feed-forward Deep Neural Network (DNN) was constructed with two hidden layers (256 and 128 units) with ReLU activations and dropout. This model was also trained for five epochs with the Adam optimizer (learning rate = 0.001) on the static embeddings.

End-to-End Fine-Tuned Transformer Models This group represents our primary end-to-end models, where the entire transformer architecture is updated during training. For these, we used a 90/10 split on our main training set for training and validation, respectively. All models were fine-tuned using the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of $2e-5$ and a batch size of 16. We compared four different architectures:

- **Finetuned BERT-base:** The `bert-base-uncased` model (Devlin et al. 2019b), fine-tuned end-to-end. This serves as our primary contextual model.
- **Finetuned MentalBERT:** A domain-adapted model (Ji et al. 2022) pre-trained on text from Reddit mental health forums.
- **Finetuned MentalRoBERTa:** A RoBERTa-base model also pre-trained on the same domain-specific mental health corpus as MentalBERT (Ji et al. 2022).

- **Finetuned ModernBERT:** We benchmark ModernBERT (Warner et al. 2024) to test whether recent architectural improvements offer performance gains over standard BERT on our task, even without domain-specific pre-training.

Zero-Shot LLM Baseline As a modern, non-finetuned baseline, we evaluated a powerful Large Language Model (GPT-OSS 120B)(Agarwal et al. 2025) in a zero-shot setting(Kojima et al. 2022). We prompted the model to classify samples from our test set into one of the ten categories. This approach required a post-processing step to map the model’s textual outputs to our valid class labels; approximately 31.6% of responses did not map to a valid label and were excluded from the LLM’s performance calculation.

Evaluation Metrics

All performance metrics are reported on the held-out, imbalanced test set (Table 1). This ensures our evaluation realistically assesses model performance on the original, real-world class distribution.

We report the following metrics of model performance:

- **Accuracy:** Overall proportion of correct predictions.
- **Macro F1:** The unweighted average of all F1 scores across classes. The macro F1 score treats all classes equally, which is critical for highlighting performance on our rare, low-resource classes.
- **Weighted F1:** The average of the per-class F1 scores, weighted by the size (support) of the class in the dataset.

We additionally assess the performance of each ML model in each class, to identify which classes are more and less challenging to predict. We report the following per-class metrics of model performance:

- **Precision & Recall:** Precision is the proportion of correct positive predictions; recall is the proportion of actual positives identified as positive.
- **Per-Class F1:** F1 score reported individually per class. The F1 score is the harmonic mean of precision and recall.
- **AUPRC:** For the high-risk *Suicide* class, we additionally report the Area Under the Precision–Recall Curve (McDermott et al. 2024), which is more informative under class imbalance.
- **Calibration Plot:** We also provide a calibration analysis for the ‘Suicide’ class to assess whether the models’ predicted confidence scores are reliable.

Exploratory data analysis

We conducted both raw word frequency analysis and TF-IDF (Term Frequency–Inverse Document Frequency) analysis to identify representative words within each dataset. While the raw frequency approach consistently surfaced high-frequency function words such as “I”, “the”, “to”, and “my”, TF-IDF assigns higher importance to words that are characteristic of a specific category but rare in others. This approach enabled us to uncover more semantically meaningful and category-specific terms such as bipolar (bipolar

class), rape (gender cyberbullying class), and bullied (age cyberbullying class). Figure 1 presents the top ten most frequent words associated with each class label in the dataset.

To examine the lexical similarity between classes, we performed a correlation analysis of the TF-IDF features. Inspired by past natural language processing work taking similar approaches to compare classes (Mobin et al. 2024; Liu, Lin, and Sun 2020), we computed the mean TF-IDF vector by averaging across all its documents. The Pearson correlation coefficient was then computed between the mean TF-IDF vectors of each dataset pair to assess the degree of lexical similarity. The resulting heatmap is presented in Figure 2. High positive correlations indicate shared vocabulary patterns, while low correlations suggest distinct linguistic structures. Strong correlations are observed among mental health-related classes such as stress, bipolar, and personality Disorder. In contrast, cyberbullying classes exhibit much lower correlations with mental health datasets, indicating distinct linguistic patterns.

Results

This section presents the empirical results of our experiments. We first provide a comparative analysis of the overall performance of all models, followed by a detailed per-class breakdown to identify specific strengths and weaknesses. Finally, we conduct a focused analysis on the critical task of suicide detection, evaluating model reliability through AUPRC and calibration plots.

Overall Model Performance

Our results demonstrate a clear performance advantage for end-to-end fine-tuned transformer models over both traditional machine learning (ML) methods and hybrid approaches that use static BERT embeddings. As shown in Table 2, **Finetuned MentalBERT** emerged as the top-performing model, achieving an accuracy of 0.92 and a Macro F1 score of 0.76. The other fine-tuned variants, including the generic BERT-base, RoBERTa, and ModernBERT, also delivered strong performance, with Macro F1 scores ranging from 0.70 to 0.71. The competitive performance of the generic fine-tuned BERT highlights the effectiveness of our data curation and balancing pipeline in adapting a general-purpose model to this specific task.

In contrast, the traditional and hybrid ML models exhibited lower performance (Table 3). The best-performing ML model, TF-IDF_LogReg, achieved a Macro F1 score of 0.67. Models using static BERT embeddings as features for classical classifiers (BERT_LogReg, BERT_RNN, BERT_DNN) performed the poorest, with Macro F1 scores between 0.53 and 0.58. This significant gap underscores the necessity of fine-tuning the entire transformer architecture to capture the complex contextual nuances present in the data.

Per-Class Performance Analysis

A granular, per-class analysis reveals significant performance disparities across the ten categories, reflecting the

Dataset	1	2	3	4	5	6	7	8	9	10
Anxiety	im	anxiety	like	just	ive	restless	feel	dont	know	really
Stress	im	stress	just	like	feel	dont	ive	know	time	really
Bipolar	im	just	like	feel	ive	dont	bipolar	know	really	want
Personality Disorder	im	like	just	people	dont	feel	avpd	know	ive	want
Suicide	im	just	dont	want	like	feel	life	know	ive	people
Non-Suicide	im	just	like	dont	want	know	people	day	filler	got
Gender Cyberbullying	jokes	rape	gay	joke	im	sexist	just	bitch	like	dont
Religion Cyberbullying	muslims	muslim	idiot	christian	islam	islamic	idiots	like	terrorism	dont
Ethnicity Cyberbullying	fuck	dumb	nigger	ass	obama	nigers	black	white	bitch	ur
Age Cyberbullying	school	high	bullied	bully	girls	girl	bullies	like	just	im

Figure 1: Table showing top 10 TF-IDF words for each class label

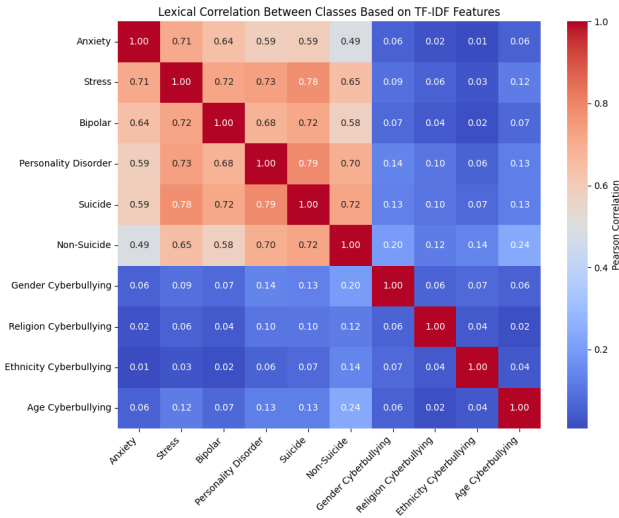


Figure 2: Plot of correlation of TF-IDF embeddings across different class labels

challenges posed by our realistic, imbalanced test set (Table 4). The fine-tuned models, particularly MentalBERT, excelled on high-signal categories with explicit lexical cues. For instance, MentalBERT achieved outstanding F1-scores on cyberbullying classes like *Age CB* (0.95) and *Religion CB* (0.96), as well as the critical mental health class *Suicide* (0.96).

Conversely, performance was substantially lower for nuanced mental health conditions that were sparsely represented in the imbalanced test set and are characterized by greater lexical ambiguity. The F1-scores for *Personality Disorder* (0.32), *Stress* (0.46), and *Bipolar* (0.70) were markedly lower for MentalBERT. This outcome is not an indication of model failure but rather a realistic reflection of the inherent difficulty of detecting these conditions from isolated posts without longitudinal context. The severe class

Table 2: Overall Model Performance Comparison for Fine-tuned BERT variants.

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
BERT	0.87	0.70	0.88	0.64	0.88
MentalBERT	0.92	0.76	0.93	0.70	0.89
RoBERTa	0.88	0.70	0.90	0.64	0.90
ModernBERT	0.89	0.71	0.91	0.64	0.88

Table 3: Overall ML Model Performance Comparison.

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
TF-IDF_LogReg	0.82	0.67	0.84	0.61	0.82
TF-IDF_SVM	0.80	0.64	0.83	0.58	0.81
BERT_LogReg	0.75	0.55	0.79	0.49	0.76
BERT_RNN	0.68	0.53	0.74	0.48	0.74
BERT_DNN	0.77	0.58	0.80	0.51	0.76

imbalance in the test set, which mirrors real-world data distribution, correctly penalizes models for misclassifying these rare but important cases, leading to lower but more credible Macro F1 scores. The traditional ML models followed a similar trend but with uniformly lower scores across all classes (Table 5).

Analysis on Suicide Detection: AUPRC and Calibration

For the high-stakes task of suicide detection, we evaluated model reliability. All fine-tuned models demonstrated excellent discriminative power, achieving near-perfect AUPRC scores (≥ 0.992), with MentalBERT and ModernBERT reaching 0.994 (Figure 3). This confirms their ability to identify suicidal content with high precision.

Beyond discrimination, the calibration analysis (Figure 4) identifies MentalBERT as the most trustworthy model. Its confidence scores are well-calibrated, in contrast to the under-confident standard BERT and over-confident ModernBERT. This superior reliability makes MentalBERT the most suitable model for this critical screening task.

Table 4: Per-Class F1, Precision, and Recall – Fine-tuned Models.

Model	Age CB	Anx.	Bip.	Eth. CB	Gen. CB	Non-Su.	Pers. Dis.	Rel. CB	Str.	Suic.
F1-Score										
BERT	0.92	0.71	0.62	0.87	0.64	0.85	0.19	0.93	0.30	0.96
MentalBERT	0.95	0.65	0.70	0.91	0.74	0.93	0.32	0.96	0.46	0.96
RoBERTa	0.92	0.60	0.40	0.89	0.68	0.88	0.17	0.94	0.56	0.96
ModemBERT	0.91	0.66	0.61	0.87	0.71	0.90	0.24	0.94	0.33	0.96
Precision										
BERT	0.88	0.61	0.50	0.79	0.48	0.98	0.11	0.89	0.18	0.96
MentalBERT	0.96	0.52	0.62	0.86	0.61	0.97	0.21	0.95	0.32	0.99
RoBERTa	0.88	0.45	0.26	0.83	0.52	0.98	0.10	0.91	0.46	0.98
ModemBERT	0.86	0.52	0.48	0.78	0.57	0.97	0.14	0.91	0.22	0.99
Recall										
BERT	0.97	0.86	0.82	0.97	0.95	0.75	0.75	0.98	0.82	0.96
MentalBERT	0.95	0.89	0.80	0.97	0.93	0.89	0.70	0.97	0.82	0.94
RoBERTa	0.97	0.91	0.87	0.95	0.97	0.79	0.90	0.98	0.74	0.94
ModemBERT	0.97	0.90	0.84	0.98	0.95	0.84	0.65	0.98	0.74	0.93

Note: CB—Cyberbullying, Anx.—Anxiety, Bip.—Bipolar, Eth.—Ethnicity, Gen.—Gender, Non-Su.—Non-Suicidal, Pers. Dis.—Personality Disorder, Rel.—Religion, Str.—Stress, Suic.—Suicide.

Table 5: Per-Class F1, Precision, and Recall – ML Models.

Model	Age CB	Anx.	Bip.	Eth. CB	Gen. CB	Non-Su.	Pers. Dis.	Rel. CB	Str.	Suic.
F1-Score										
TF-IDF.LogReg	0.89	0.57	0.44	0.86	0.67	0.85	0.43	0.93	0.19	0.87
TF-IDF.SVM	0.92	0.56	0.39	0.86	0.62	0.83	0.28	0.93	0.17	0.85
BERT.LogReg	0.73	0.43	0.26	0.72	0.52	0.78	0.14	0.89	0.17	0.85
BERT.RNN	0.70	0.45	0.30	0.68	0.41	0.73	0.20	0.88	0.09	0.81
BERT.DNN	0.73	0.52	0.30	0.71	0.62	0.79	0.25	0.81	0.16	0.86
Precision										
TF-IDF.LogReg	0.84	0.43	0.32	0.80	0.54	0.89	0.31	0.89	0.11	0.93
TF-IDF.SVM	0.88	0.41	0.27	0.78	0.48	0.88	0.18	0.91	0.10	0.93
BERT.LogReg	0.62	0.30	0.17	0.60	0.37	0.91	0.08	0.83	0.10	0.93
BERT.RNN	0.59	0.31	0.21	0.55	0.27	0.93	0.13	0.83	0.05	0.95
BERT.DNN	0.61	0.40	0.20	0.58	0.49	0.91	0.17	0.69	0.09	0.93
Recall										
TF-IDF.LogReg	0.95	0.88	0.71	0.94	0.87	0.80	0.70	0.97	0.58	0.82
TF-IDF.SVM	0.96	0.89	0.71	0.97	0.89	0.79	0.65	0.96	0.52	0.79
BERT.LogReg	0.89	0.76	0.65	0.89	0.87	0.68	0.50	0.96	0.60	0.79
BERT.RNN	0.87	0.81	0.56	0.90	0.92	0.60	0.40	0.95	0.72	0.71
BERT.DNN	0.91	0.74	0.58	0.91	0.83	0.71	0.50	0.97	0.64	0.81

Note: CB—Cyberbullying, Anx.—Anxiety, Bip.—Bipolar, Eth.—Ethnicity, Gen.—Gender, Non-Su.—Non-Suicidal, Pers. Dis.—Personality Disorder, Rel.—Religion, Str.—Stress, Suic.—Suicide.

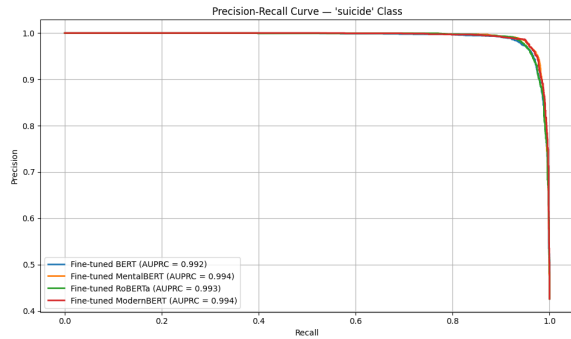


Figure 3: Precision-Recall Curve for the 'Suicide' class, showing near-perfect AUPRC scores for all fine-tuned models.

Impact of Data Balancing Strategy

To evaluate the effectiveness of our dual data balancing approach, we compared model performance before and after its application. The results confirm a crucial improvement in overall robustness. For our top-performing model, MentalBERT, the Macro F1 score, the most significant metric for imbalanced classes, increased from **0.73 to 0.76**, while accuracy rose from 0.90 to 0.92. This demonstrates the value

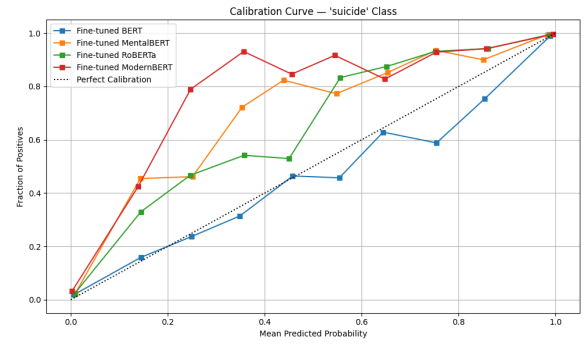


Figure 4: Calibration Curve for the 'Suicide' class. MentalBERT demonstrates the best calibration, with its predicted probabilities closely matching the observed frequencies.

of our pipeline in creating a more equitable and effective classifier.

The strategy's primary benefits are most evident at the per-class level, validating its targeted impact. The most substantial gain occurred in the **Bipolar** class, where the F1-score rose dramatically from **0.49 to 0.70**. Notable improvements were also observed for other underrepresented mental health classes like **Anxiety** (0.59 to 0.65) and **Personality Disorder** (0.28 to 0.32). Interestingly, the strategy did not improve all rare classes equally; the F1-score for **Stress** remained unchanged, suggesting its lexical ambiguity presents a challenge that oversampling alone cannot solve. Overall, these results justify our methodological choice, confirming that the balancing strategy provides a targeted performance lift for key underrepresented classes.

Error Analysis

An analysis of the confusion matrix for our top model, MentalBERT (Figure 5), reveals key learning behaviors. The model demonstrates high performance on well-represented classes with clear lexical signals, such as **Suicide** and the cyberbullying categories, as indicated by the strong diagonal.

More importantly, the off-diagonal error patterns provide strong evidence that the model is learning true semantic relationships, not superficial domain cues. Misclassifications are concentrated between semantically coherent categories, such as the confusion between **Anxiety** and **Stress**. In contrast, cross-domain errors between the mental health (primarily Reddit-based) and cyberbullying (primarily Twitter-based) categories are minimal: only 2.3% (123 out of 5,311) of mental health posts were misclassified as a cyberbullying class, while a mere 1.5% (10 out of 689) of cyberbullying posts were misclassified as mental health. This pattern refutes the hypothesis of the model learning spurious platform heuristics and confirms it is addressing the intended nuanced classification task.

Ethical Considerations

The development of AI for mental health screening requires a rigorous ethical framework to ensure responsible innovation. Our work is therefore grounded in a framework de-

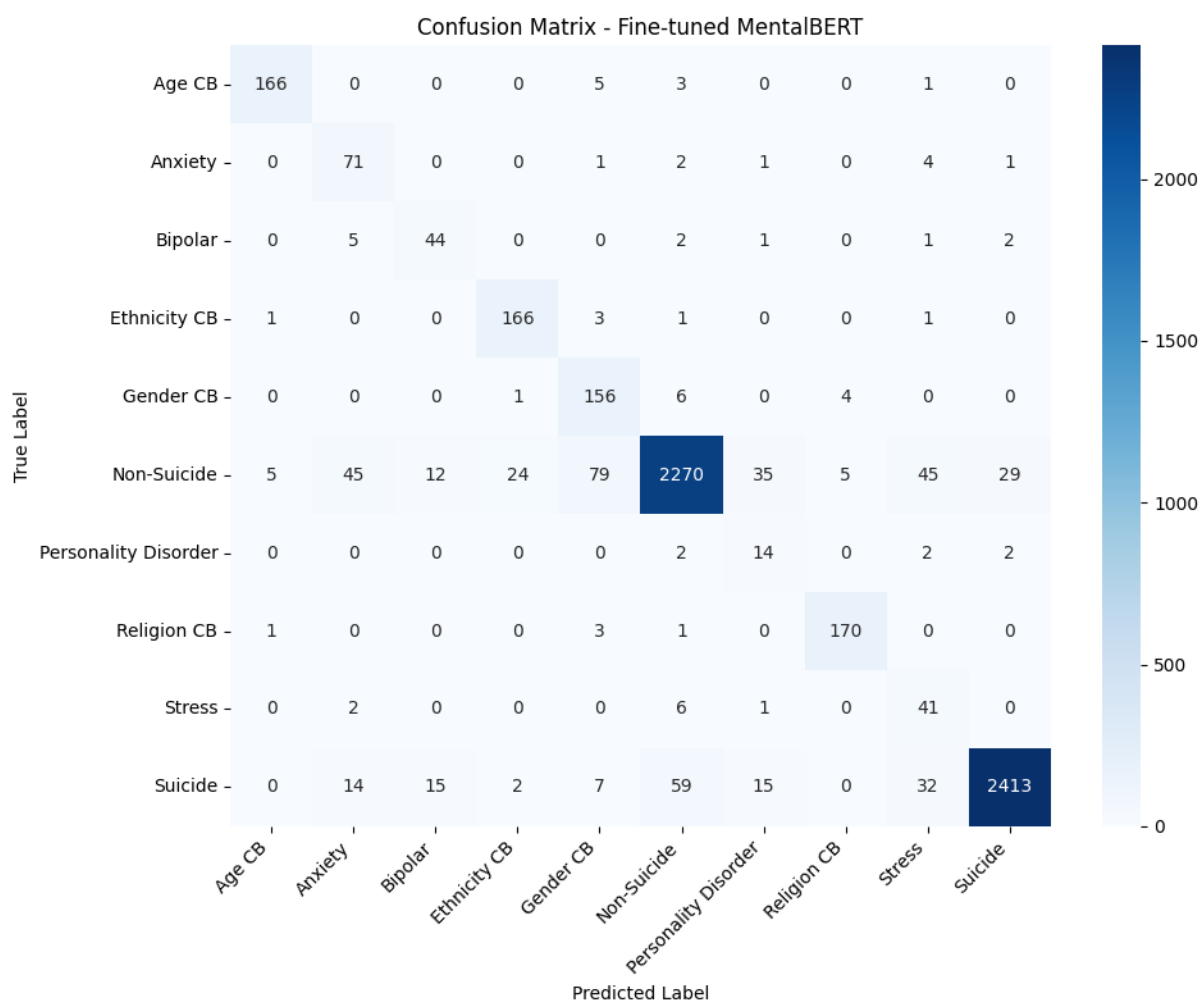


Figure 5: Confusion matrix for Fine-tuned MentalBERT. The model’s primary errors occur between semantically related classes (e.g., Anxiety and Stress), confirming that it learns content over platform-specific artifacts.

signed to proactively address the critical challenges of model application, data provenance, and fairness.

Intended Application and Limitations

A primary ethical risk is the misinterpretation of model outputs as clinical diagnoses. We explicitly state that our model is not a diagnostic tool. Instead, we propose its use as a screening assistant for trained human moderators within a dual-purpose framework:

- **For Immediate Risk Flagging:** For high-signal classes like *Suicide*, the model can effectively flag content for urgent human review, helping to prioritize potentially life-saving interventions.
- **Nuanced Pattern Analysis:** For conditions like *Bipolar Disorder* that require longitudinal assessment, the model supports a broader decision-support tool by highlighting linguistic shifts over time. For example, tracking changes in language associated with mood episodes can offer an early signal for clinical intervention, rather than labeling

individual posts in isolation.

In all cases, the system is designed to inform, not replace, professional human judgment.

Data Privacy and Provenance

Our work uses anonymized, publicly available data from Kaggle. All datasets were curated in accordance with platform terms and community research standards. The data were further processed to remove any personally identifiable information, ensuring compliance with ethical use guidelines. Our label verification study (Section) served as an additional quality assurance step to validate labeling consistency and dataset integrity.

Accountability and Misuse

This framework is designed for a strictly supportive purpose, operationalized through a human-in-the-loop system where trained professionals verify all automated flags. The intended application of the resulting tool is to assist content

moderation and connect individuals with resources. Consequently, any use for surveillance, censorship, or punitive action would constitute a misuse and falls outside the tool’s ethical scope.

Model Explainability and Inference Architecture

To ensure our framework is transparent and trustworthy for real-world applications, we developed a hybrid explainability system that decouples real-time detection from deep semantic analysis. This approach combines the quantitative precision of SHAP (SHapley Additive exPlanations)(Lundberg and Lee 2017) with the narrative clarity of a Large Language Model (GPT-OSS)(Agarwal et al. 2025), accessed via the **Groq API**.

First, SHAP identifies the precise mathematical contribution of each word to the model’s prediction. This quantitative evidence is then synthesized by the LLM to generate a coherent, human-readable explanation. Crucially, this high-resource component operates on an **on-demand, post-by-post basis**, triggered only when detailed auditor review is required. This design ensures that the system remains computationally efficient, reserving the API-based overhead for targeted inquiries while maintaining low-latency performance for the primary screening pipeline.

To demonstrate how this system is integrated into our proposed human-in-the-loop application, Figure 6 shows a prototype of the “Social Media Screener” dashboard.

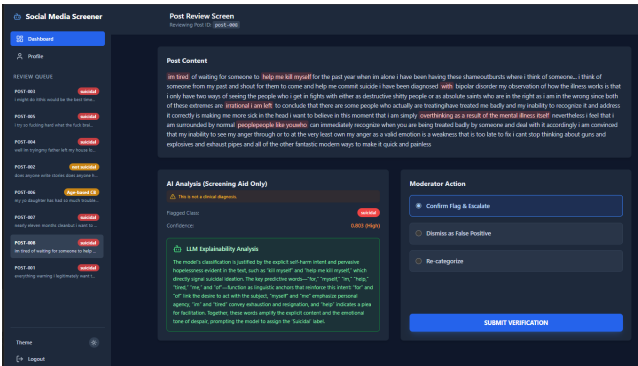


Figure 6: The proposed “Social Media Screener” prototype. This interface integrates the hybrid explainability system directly for a human moderator. The SHAP-derived token importance is shown via red highlights in the ‘Post Content’ section, while the ‘LLM Explainability Analysis’ is presented in a clear text box below the model’s flag.

This prototype visualizes how a human moderator interacts with the system. Instead of a raw plot, the SHAP analysis is rendered as simple **red highlights** on the most impactful words (e.g., “help me kill,” “myself,” “tired of”). This quantitative data is paired with the LLM Explainability Analysis in the green box, which provides a narrative summary. This interface also includes the critical ethical safeguard (“This is not a clinical diagnosis.”) and the “Moderator Action” panel, which requires a human to verify every flag, completing the human-in-the-loop framework. Additional examples can be found in Appendix .

Analysis of Performance Disparities and Domain Cues

Our per-class results show performance variations across categories, which our analysis attributes to the inherent characteristics of the data rather than model artifacts. As confirmed by our error analysis (Section), the model does not rely on superficial platform cues (e.g., Reddit vs. Twitter) but instead learns from the content’s semantic properties. Performance differences are primarily driven by two factors:

- **Lexical Specificity:** Cyberbullying classes, typically sourced from shorter Twitter posts, often contain explicit, high-signal keywords. This results in higher classification accuracy due to the clear and consistent linguistic markers.
- **Narrative Complexity:** Mental health classes, often from longer, narrative-style Reddit posts, express distress in more nuanced and varied ways. The critical signal may be diffused within a larger volume of text, creating a greater challenge for the model. This aligns with our observation that the most common errors occur between semantically related mental health classes like *Anxiety* and *Stress*.

These findings demonstrate the model’s ability to adapt to diverse text styles and confirm that performance disparities are a function of the task’s complexity, not a reliance on spurious correlations.

Limitations and Future Work

Our work provides a robust framework for multi-class mental health screening, but we acknowledge limitations that pave the way for future research.

Limitations

- **Data Provenance and Generalizability:** While we verified label quality using a sample of the datasets, our use of aggregated Kaggle datasets limits broad claims of generalizability. The model’s performance on datasets from different platforms, time periods, or demographic groups remains untested.
- **Single-Label Task Framing:** Our multi-class framework treats each post as having a single, primary label. This simplifies the clinical reality where mental health conditions and cyberbullying can co-occur (e.g., a post expressing suicidal ideation within a narrative about Bipolar Disorder). As shown in our error analysis (Example 1, Appendix), the model sometimes correctly identifies a high-risk secondary theme (Suicide) while misclassifying the primary label (Bipolar Disorder). A multi-label framework is a critical next step to capture this comorbidity and provide a more clinically nuanced output.
- **Scope of Evaluation:** The current study is limited to English-language text, excluding the rich signals available in multilingual and multimodal (e.g., images, videos) content.

Future Work Based on these limitations, this work highlights several key directions for future research in this domain:

- **Curation of a Multi-Label Benchmark:** The most critical next step for the field is the creation of a new, ethically sourced, and expertly annotated benchmark dataset. Such a dataset should feature multi-label annotations to enable the study of co-occurring conditions, reflecting a more clinically realistic scenario.
- **Validating Utility with Human-in-the-Loop Studies:** The promising results of our prototype pave the way for formal Human-Computer Interaction (HCI) studies. Future work should engage trained moderators in a user study to quantitatively validate the real-world impact of our hybrid explainability system on decision-making accuracy, efficiency, and trust in AI-assisted workflows.
- **Advanced Model Architectures:** The availability of a multi-label benchmark would enable the exploration of multi-task learning architectures. These models could feature a shared encoder with separate classification heads to explicitly model the interplay between different mental health and cyberbullying phenomena.
- **Robustness and Multimodal Evaluation:** Future research should prioritize rigorous cross-domain and temporal generalization tests to ensure models are robust in real-world environments. Furthermore, expanding these frameworks to incorporate multilingual and multimodal analysis is essential for building more comprehensive and inclusive screening tools.

References

- Abdullah, M.; and Negied, N. 2024. Detection and Prediction of Future Mental Disorder From Social Media Data Using Machine Learning, Ensemble Learning, and Large Language Models. *IEEE Access*, 12: 120553–120569.
- Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R. K.; Bai, Y.; Baker, B.; Bao, H.; et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Antypas, D.; and Camacho-Collados, J. 2023. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, 231–242. Toronto, Canada.
- de Arriba-Pérez, F.; and García-Méndez, S. 2024. Detecting anxiety and depression in dialogues: a multi-label and explainable approach. *ArXiv:2412.17651*.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019b. 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, volume 1 (long and short papers)*, 4171–4186.
- Ghoshal, N. 2025. Reddit Mental Health Data. <https://www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data>. Accessed: May 14, 2024.
- Ji, Z.; Reddy, T.; Nagda, A.; and Singh, H. 2022. MentalBERT: Publicly available pretrained transformer-based model for mental healthcare social media text mining on Reddit. In *Proceedings of LREC*.
- Jiang, Y. 2021. Problematic Social Media Usage and Anxiety Among University Students During the COVID-19 Pandemic: The Mediating Role of Psychological Capital and the Moderating Role of Academic Burnout. *Frontiers in Psychology*, 12: 612007.
- Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.
- Komati, N. 2021. Suicide and Depression Detection. <https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch>. Accessed: May 14, 2025.
- Kumar, P.; Samanta, P.; Dutta, S.; Chatterjee, M.; and Sarkar, D. 2022. Feature Based Depression Detection from Twitter Data Using Machine Learning Techniques. *Journal of Scientific Research*, 66(02): 220–228.
- Liu, Z.; Lin, Y.; and Sun, M. 2020. Sentence Representation. In Liu, Z.; Lin, Y.; and Sun, M., eds., *Representation Learning for Natural Language Processing*, 59–89. Singapore: Springer Nature. ISBN 9789811555732.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. *arXiv:1711.05101*.
- Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874*.
- Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *CoRR*, abs/2012.10289.
- McDermott, M.; Zhang, H.; Hansen, L.; Angelotti, G.; and Gallifant, J. 2024. A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, 37: 44102–44163.
- Mobin, M. I.; Akhter, A. F. M. S.; Mridha, M. F.; Mahmud, S. M. H.; and Aung, Z. 2024. Social Media as a Mirror: Reflecting Mental Health Through Computational Linguistics. *IEEE Access*, 12: 130143–130164.
- Naslund, J.; Bondre, A.; Torous, J.; and Aschbrenner, K. 2020. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. *Journal of Technology in Behavioral Science*, 5.
- Nweke, A.; Khan, M.; and Pei, Y. 2024. Explainable Multi-Label Classification Framework for Behavioral Health Based on Domain Concepts. *IEEE Transactions on Artificial Intelligence*, 15: 89–105.
- Poddar, S.; Mukherjee, R.; Samad, A.; Ganguly, N.; and Ghosh, S. 2024. MuLX-QA: Classifying Multi-Labels and Extracting Rationale Spans in Social Media Posts. *ACM Transactions on Information Systems*, 42(3).
- Robertson, S. 2004. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation - J DOC*, 60: 503–520.

Sarkar, S. 2024. Sentiment Analysis for Mental Health. <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>. Kaggle dataset. Accessed: May 14, 2025.

Wang, J.; Fu, K.; and Lu, C.-T. 2020. SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. In *2020 IEEE International Conference on Big Data (Big Data)*, 1699–1708.

Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; Cooper, N.; Adams, G.; Howard, J.; and Poli, I. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv:1901.11196.

World Health Organization. 2025. Mental Disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>. Accessed: April 22, 2025.

Wu, M.; Chang, J.; Epstein, Z.; and Rand, D. 2025. Beyond Friends: Exploring the Effects of Unknown Users’ Social Media Posts on Individuals’ Perceptions and Behaviors.

Zhang, Y.; and Liu, J. 2024. Depression Detection on Social Media with Large Language Models. <https://arxiv.org/abs/2403.10750>. ArXiv preprint.

Zsila, A.; and Reyes, M. E. S. 2023. Pros & cons: impacts of social media on mental health. *BMC Psychology*, 11(1): 201.

Appendix

Model Hyperparameters

This section details the hyperparameters used for all feature extraction methods and machine learning models evaluated in this study. For models trained using cross-validation or grid search, the search space is specified.

Table 6: Hyperparameters for Feature Extraction and Trained Models

Component	Model / Method	Hyperparameters
Feature Extraction	TF-IDF Vectorizer	Max Features: 5000 Stop Words: None (preserved for training) N-gram Range: (1, 2)
	BERT Embeddings	Model: <i>bert-base-uncased</i> Max Sequence Length: 128 Solver: 'saga', Setting: 'multinomial' Max Iterations: 1000 Regularization Strengths (C): [0.1, 1, 10] Cross-Validation Folds: 5
Lexical Based Models	Logistic Regression (CV)	Kernel: 'linear'
	Support Vector Machine (Grid Search)	Regularization Strengths (C): [0.01, 0.1, 1, 10] Cross-Validation Folds: 5 Solver: 'saga', Setting: 'multinomial' Max Iterations: 1000, CV Folds: 5 Hidden Layers: [256, 128] Activation Function: ReLU, Dropout: 0.3 Optimizer: Adam, Learning Rate: 0.001 Batch Size: 64, Epochs: 5
Static BERT Embedding Based Models	Logistic Regression (CV)	Hidden Layer Dimension: 128 Optimizer: Adam, Learning Rate: 0.001 Batch Size: 64, Epochs: 5
	Feedforward NN (DNN)	Optimizer: AdamW Learning Rate: 2e-5 Batch Size: 16 Training/Validation Split: 90/10
End-to-End Fine-Tuned Models	Recurrent Neural Network (RNN)	
	Finetuned BERT-base Finetuned MentalBERT Finetuned MentalRoBERTa Finetuned ModernBERT	

Model Performances before oversampling

The performance of the model was tested without oversampling the data and the performance metrics can be seen below

Table 7: Overall ML Model Performance on Non-Oversampled Data.

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
TF-IDF_LogReg	0.82	0.67	0.84	0.61	0.82
TF-IDF_SVM	0.81	0.64	0.83	0.58	0.81
BERT_LogReg	0.75	0.55	0.79	0.49	0.76
BERT_RNN	0.78	0.58	0.81	0.52	0.76
BERT_DNN	0.72	0.55	0.77	0.51	0.74

Table 8: Per-Class Metrics for ML Models on Non-Oversampled Data.

Model	Age CB	Anx.	Bip.	Eth. CB	Gen. CB	Non-Su.	Pers. Dis.	Rel. CB	Str.	Suic.
F1-Score										
TF-IDF_LogReg	0.89	0.57	0.45	0.86	0.67	0.85	0.43	0.93	0.19	0.87
TF-IDF_SVM	0.92	0.56	0.39	0.86	0.63	0.84	0.29	0.93	0.17	0.85
BERT_LogReg	0.73	0.43	0.26	0.72	0.52	0.78	0.13	0.89	0.17	0.85
BERT_RNN	0.72	0.46	0.24	0.81	0.54	0.81	0.29	0.86	0.22	0.87
BERT_DNN	0.77	0.47	0.24	0.79	0.47	0.80	0.19	0.91	0.12	0.79
Precision										
TF-IDF_LogReg	0.84	0.43	0.33	0.80	0.54	0.89	0.31	0.89	0.11	0.93
TF-IDF_SVM	0.88	0.41	0.27	0.78	0.49	0.88	0.18	0.91	0.10	0.93
BERT_LogReg	0.62	0.30	0.16	0.60	0.37	0.92	0.08	0.84	0.10	0.93
BERT_RNN	0.59	0.34	0.15	0.77	0.39	0.91	0.20	0.78	0.14	0.92
BERT_DNN	0.68	0.34	0.15	0.73	0.31	0.88	0.12	0.90	0.06	0.96
Recall										
TF-IDF_LogReg	0.95	0.88	0.71	0.94	0.87	0.81	0.70	0.97	0.58	0.82
TF-IDF_SVM	0.96	0.89	0.71	0.97	0.89	0.79	0.65	0.96	0.52	0.79
BERT_LogReg	0.89	0.76	0.65	0.89	0.87	0.68	0.50	0.96	0.60	0.79
BERT_RNN	0.91	0.74	0.71	0.85	0.87	0.73	0.50	0.96	0.46	0.82
BERT_DNN	0.87	0.78	0.58	0.87	0.93	0.73	0.45	0.91	0.66	0.68

Note: CB—Cyberbullying, Anx.—Anxiety, Bip.—Bipolar, Eth.—Ethnicity, Gen.—Gender, Non-Su.—Non-Suicidal, Pers. Dis.—Personality Disorder, Rel.—Religion, Str.—Stress, Suic.—Suicide.

Table 9: Overall Fine-tuned Model Performance on Non-Oversampled Data.

Model	Accuracy	Macro F1	Weighted F1	Precision	Recall
Fine-tuned BERT	0.89	0.71	0.91	0.66	0.89
Fine-tuned MentalBERT	0.90	0.73	0.92	0.66	0.89
Fine-tuned RoBERTa	0.89	0.73	0.91	0.67	0.90
Fine-tuned ModernBERT	0.88	0.69	0.90	0.63	0.88

Table 10: Per-Class Metrics for Fine-tuned Models on Non-Oversampled Data.

Model	Age CB	Anx.	Bip.	Eth. CB	Gen. CB	Non-Su.	Pers. Dis.	Rel. CB	Str.	Suic.
F1-Score										
Fine-tuned BERT	0.94	0.58	0.52	0.91	0.84	0.90	0.16	0.96	0.38	0.95
Fine-tuned MentalBERT	0.94	0.59	0.49	0.90	0.78	0.91	0.28	0.97	0.46	0.96
Fine-tuned RoBERTa	0.97	0.59	0.52	0.82	0.79	0.89	0.17	0.94	0.62	0.96
Fine-tuned ModernBERT	0.95	0.53	0.55	0.86	0.81	0.89	0.22	0.91	0.27	0.96
Precision										
Fine-tuned BERT	0.91	0.42	0.38	0.87	0.77	0.97	0.09	0.95	0.26	0.98
Fine-tuned MentalBERT	0.92	0.44	0.34	0.85	0.65	0.98	0.17	0.95	0.32	0.98
Fine-tuned RoBERTa	0.98	0.44	0.37	0.71	0.69	0.98	0.09	0.91	0.54	0.97
Fine-tuned ModernBERT	0.94	0.37	0.41	0.77	0.73	0.97	0.13	0.84	0.16	0.99
Recall										
Fine-tuned BERT	0.96	0.93	0.82	0.95	0.93	0.84	0.90	0.97	0.72	0.92
Fine-tuned MentalBERT	0.95	0.88	0.84	0.94	0.97	0.85	0.70	0.99	0.72	0.95
Fine-tuned RoBERTa	0.95	0.93	0.87	0.97	0.92	0.81	0.90	0.98	0.72	0.95
Fine-tuned ModernBERT	0.97	0.94	0.82	0.97	0.90	0.81	0.65	0.98	0.71	0.92

Note: CB—Cyberbullying, Anx.—Anxiety, Bip.—Bipolar, Eth.—Ethnicity, Gen.—Gender, Non-Su.—Non-Suicidal, Pers. Dis.—Personality Disorder, Rel.—Religion, Str.—Stress, Suic.—Suicide.

Explainability Examples

Below are some examples of our hybrid explainability framework in action. These instances demonstrate how the system provides clear rationales for its predictions.

Prototype Interface Walkthrough

The “Social Media Screener” interface shown in each example is designed for a human-in-the-loop workflow. The key components are:

- **Post Content:** The original text of the post. Words identified by SHAP as having the highest impact on the model’s prediction are highlighted in red, providing immediate, quantitative evidence.
- **AI Analysis (Screening Aid Only):** Displays the model’s predicted label and confidence score. It includes a critical disclaimer that this is not a clinical diagnosis.
- **LLM Explainability Analysis:** A human-readable narrative, synthesized by an LLM from the SHAP values, explaining *why* the model made its decision in plain language.
- **Moderator Action:** A required action panel where the human user must confirm, dismiss, or re-categorize the flag, ensuring no decision is fully automated.

Example 1: Misclassification Highlighting Model Nuance

Text: “everything warning i legitimately want to kill myself just to spite my father i lived on my own for many years and about years ago i was guiltd into ...”

Actual Label: Bipolar Disorder

Predicted Label: Suicidal (Confidence: 1.000)

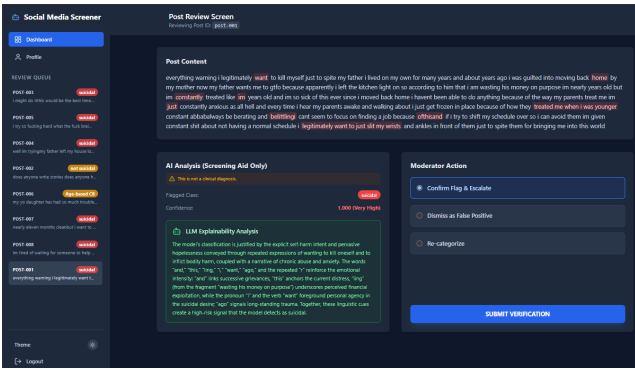


Figure 7: Hybrid explainability output for a text labeled ‘Bipolar Disorder’ but classified as ‘Suicidal’ due to strong self-harm intent. This highlights the model’s ability to detect co-occurring high-risk language.

Example 2: Age-based Cyberbullying

Text: “my yo daughter has had so much trouble at school being left out bullied and the final straw she was jumped by girls on high street of town we live in ...”

Actual Label: Age-based CB

Predicted Label: Age-based CB (Confidence: 1.000)

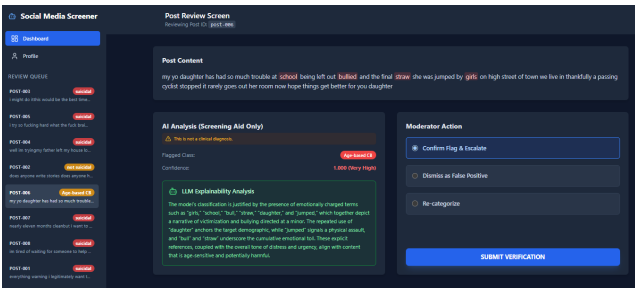


Figure 8: Hybrid explainability output for a text correctly classified as Age-based Cyberbullying. The LLM identifies terms like ‘bullied’ and ‘jumped’ as key indicators.

Example 3: Suicidal Ideation (Explicit)

Text: “nearly eleven months clean but i want to cut it short by bleeding myself dry in the bathroom fuck i really want to escape whatever the fuck i am whatev...”

Actual Label: Suicidal

Predicted Label: Suicidal (Confidence: 0.999)

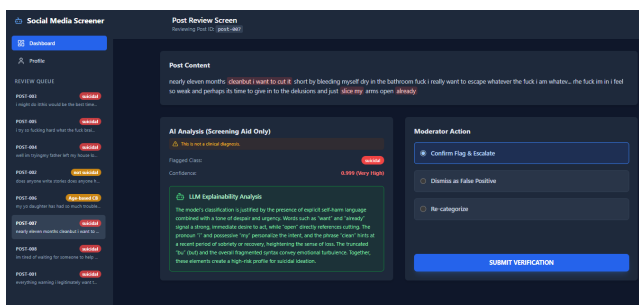


Figure 9: Hybrid explainability output for a text correctly classified as Suicidal, showing explicit self-harm language ('bleeding myself dry', 'slice my arms').

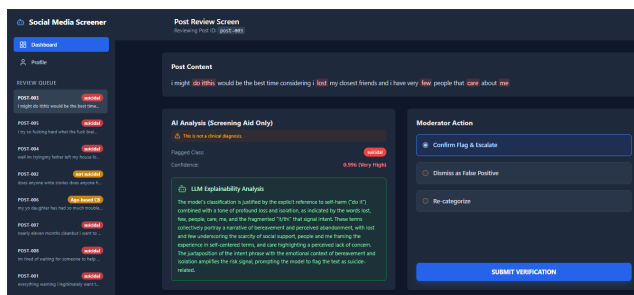


Figure 11: Hybrid explainability output for a text correctly classified as 'suicidal'. The model identifies nuanced indicators of isolation and hopelessness ('lost', 'few people', 'care about me') as contributing to the risk profile.

Example 4: Non-Suicidal Content

Text: “does anyone write stories does anyone here write stories with characters is it tough in my experience i feel like i dont know enough about people to write characters who arent just like me...”

Actual Label: not suicidal

Predicted Label: not suicidal (Confidence: 0.444)

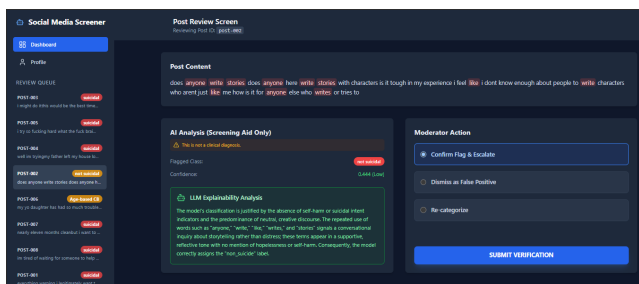


Figure 10: Hybrid explainability output for a text correctly classified as 'not suicidal'. The LLM notes the absence of self-harm indicators and the focus on creative discussion.

Example 5: Suicidal Ideation (Nuanced)

Text: “i might do this would be the best time considering i lost my closest friends and i have very few people that care about me”

Actual Label: suicidal

Predicted Label: suicidal (Confidence: 0.996)