Echoes of the Mind: Detecting Mental Health Signals in the Noise of Social Media

Kirati Tangjitrmaneesakda
Software Engineering Group B
Machine Learning and Smart Systems
University of Europe for Applied Science
Kirati.Tangjitrmaneesakda@ue-germany.de

Raja Hashim Ali
Department of Business
Univ. of Europe for Applied Sciences
Potsdam 14469, Germany
hashim.ali@ue-germany.de

***Abstract*—The rise of social media as a space for expressing mental health struggles presents both a challenge and an opportunity for suicide prevention efforts. This project proposes a machine learning and deep learning-based approach to detect suicidal ideation in Reddit posts using a publicly available dataset. The methodology explores five focused research questions addressing emotional variance, transformer model benchmarking, hybrid lexical-emotional modeling, linguistic marker identification, and real-time deployment feasibility. Findings show that emotion-variance features (Q1) enhance prediction, and hybrid models (Q3) combining TF-IDF with NRC emotions significantly boost performance, though still fall short of Transformers. RoBERTa (Q2) achieves near-perfect accuracy at high computational cost. SHAP analysis (Q4) highlights key linguistic cues and the limits of context-insensitive models. Lightweight hybrids (Q5) offer an excellent balance of performance and efficiency, ideal for real-time or budget-constrained use.**

***Index Terms*—suicide detection, Logistics Regression, Rainforest, Naive Bayes, RoBERTa, BERT, DistillBERT, Machine Learning, Deep Learning, Transformer**

I. Introduction

More and more individuals utilize social media to express and vent their distress, loneliness, and suicidal ideation without seeking for real assistance. Suicide is a leading cause of death all over the world, especially among young adults, and earliest warning signs tend to appear first in online communication. The traditional diagnosis of mental illness relies on face-to-face clinical interviews that are unaffordable, inaccessible, and stigmatized. This creates a critical demand for real-time, scalable technology that can passively monitor user-generated content for signs of a mental health crisis. However, implementing such technologies is a monumental challenge. Natural Language Processing (NLP) models will need to overcome the hurdle of interpreting emotionally nuanced, sarcastic, or coded language. In addition, there are profound ethical and technical considerations involved in the development of AI models that are equitable, accurate, and sensitive to privacy among vulnerable groups. This project addresses some of these challenges by constructing and experimenting with numerous models in an effort to detect suicidal ideation on Reddit. The principal objectives are the following:

1) Evaluate systematically the performance of traditional machine learning, hybrid, and transformer models.
2) Investigate characteristics, such as intra-post emotional variability, to increase detection accuracy.
3) Use model interpretability techniques (SHAP) to identify the most relevant linguistic patterns for suicide risk.
4) Compare the trade-offs in model accuracy, complexity, and inference speed for deployment.

II. Literature Review & Gap Analysis

A. Literature Review

Recent efforts have increasingly focused on deep learning architectures for their ability to capture complex linguistic patterns. For example, Bhuiyan, Islam, Kamarudin, and Ismail (2025) proposed a hybrid CNN-BiLSTM model that incorporated an attention mechanism to classify suicidal Reddit posts, achieving a high test accuracy of 94.29% and using SHAP for model explainability. [1] Similarly, transformer architectures have become prominent. Tavčioski, Robnik-Sikonja, and Pollak (2023) demonstrated the effectiveness of BERT, RoBERTa, and ensemble methods for depression detection on both Reddit and Twitter. Their work highlighted that ensembles outperform individual transformer models and that models can achieve successful cross-platform transfer learning. [2] Alongside end-to-end deep learning, researchers have explored hybrid models that integrate engineered features. Yeskuatov, Chua, and Foo (2024) showed the benefit of combining traditional TF-IDF vectors with the psycholinguistic characteristics of the NRC Emotion Lexicon and LIWC to classify suicidal ideation in online forums. This approach underscores the value of psychologically-grounded, interpretable features. [3] In a similar vein, Kaur, Kaur, and Kumar (2023) applied traditional machine learning algorithms, including SVM and Random Forest, to sentiment analysis features for depression detection on Twitter, achieving 88% accuracy. These studies show the continued relevance of feature-rich classical models. [4] The field is also defined by critical reviews and the development of new datasets that advance the research. The review by Iyortsuun et al. (2023) surveyed a wide range of machine learning and deep learning models for mental health diagnosis, highlighting persistent technical challenges such as data quality, model interpretability, and generalizability. [5] More specifically focused on Reddit, the systematic review by

Yeskuatov, Chua, and Foo (2022) compared 26 studies on suicidal ideation detection, calling out the need for standardized evaluation and improved preprocessing techniques. [6] Further contributing to explainable AI, Garg et al. (2023) introduced a novel Reddit dataset annotated for specific psychological risk factors—thwarted belongingness and perceived burdensomeness—providing a resource for building more clinically transparent models. [7] Finally, work by Inamdar et al. (2023) explored stress detection on Reddit using a variety of NLP features, including ELMo, BERT, and Bag-of-Words, achieving a 0.76 F1-score on a dataset of 2,800 posts. [8]

B. Gap Analysis

- Small and Imbalanced Datasets Limit Generalizability: Several works used relatively small datasets—Inamdar et al. (2,800 posts), Yeskuatov et al. (785 posts)—with notable class imbalance (e.g., 205 suicidal vs. 580 non-suicidal posts). These scale and balance issues can lead to overfitting and inflated performance metrics.
- Inconsistent Evaluation Metrics and Benchmarks: There is no standardized benchmark dataset or uniform set of evaluation metrics across studies. For instance, F1 scores are reported differently (stress vs. suicide detection), and some reviews call for standardized preprocessing to facilitate fair comparisons.
- Limited Explainability Beyond Local Features: While Bhuiyan et al. use SHAP for local explanations, most works (e.g., Tavčioski et al., Kaur et al.) do not prioritize model interpretability or transparent feature importance, which is crucial for clinical adoption and ethical accountability.
- Real-Time Deployment and Ethical Considerations Underexplored: None of the surveyed works address real-time monitoring or potential ethical pitfalls (e.g., privacy, false positives). The system pipeline from detection to intervention (e.g., alerting moderators or clinicians) remains largely unaddressed.

III. Methodology

A. Research Design

It employs a multi-stage experimental approach to investigate the promise of several linguistic and psychological features in discriminating suicide-risk posts from social media, i.e., Reddit. The methodology employs classical machine learning, hybrid construction of features as well as deep learning architectures to answer five research questions (Q1–Q5), and each phase is incrementally developed to create increasingly interpretable as well as precise system of classifying posts as suicide risk or not.

These research questions find the following answers:

1) Can intra-post emotional variance improve suicidal ideation detection?
2) How do different transformer architectures perform on suicide detection?
3) Can hybrid models using lexical and psychological features outperform deep learning models?
4) What are the most significant linguistic markers that distinguish suicidal from non-suicidal posts?
5) Can a lightweight model achieve near-transformer accuracy for real-time classification?

B. Dataset

1) Data Source: Suicide and Depression Detection. The dataset consists of posts gathered from the ”SuicideWatch” and ”depression” subreddits on Reddit, using the Pushshift API. Posts from ”SuicideWatch” span from its inception on December 16, 2008, to January 2, 2021, while those from the ”depression” subreddit were collected between January 1, 2009, and January 2, 2021. All entries in ”SuicideWatch” are labeled as suicide-related and those in ”depression” are labeled accordingly. For nonsuicidal content, additional posts were sourced from the ”r/teenagers” subreddit. There are 348,110 posts in the dataset. However, this version (v14) of the dataset has only suicide & non-suicide labels.

- https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch/data

2) Data Preprocessing: Text preprocessing was tailored depending on the experiment: Across all experiments, the initial cleaning process included lowercasing the text and removing all emojis and irrelevant metadata such as index numbers.

- For Q1: After initial cleaning, each post was split into sentences and passed through the VADER sentiment analyzer, which uses punctuation and casing to detect tone. Sentence-level compound scores were then aggregated to compute emotional variance features (standard deviation and range). After this step, punctuation was removed and the text was tokenized for NRC emotion lexicon processing. The resulting NRC and VADER features were combined and used as input for the model.
- For Q2: Minimal preprocessing was applied to preserve linguistic richness for contextual embeddings. Punctuation was retained, while emojis were removed and text was lowercased during the initial cleaning.
- Q3 and Q5: After the initial cleaning and punctuation removal, the text was first tokenized and processed using the NRC emotion lexicon to extract emotion counts and get emotion features. Simultaneously, the original cleaned post text (with punctuation removed but text preserved) was retained for TF-IDF vectorization using unigrams and bigrams. The resulting emotion features and TF-IDF vectors were then combined to create a hybrid feature set. These were saved alongside corresponding labels and TF-IDF vocabulary, which was later used in Q4 for interpretability.

3) Feature Construction:

1) VADER Sentiment Analysis: Applied only in Q1 to estimate emotional variance based on sentence-level sentiment scores.The features are Vader mean, Vader standard deviation, and vader Range.
2) NRC Emotion Lexicon: A widely used, lexicon-based resource developed by the National Research Council

of Canada [9], containing associations between English words and eight basic emotions (e.g., joy, sadness, anger) as well as positive/negative sentiment. Also, Variance features (e.g., standard deviation and range of emotion scores per post) were also computed. The only question that didn't use these features directly and indirectly is Q2 that focus on the transformers models.

3) TF-IDF Features: For Q3 and Q5, unigrams and bigrams were extracted using TF-IDF vectorization (max 10,000 features). These were combined with emotion features to form a hybrid feature space

C. Experimental Design

This section details the method and design of each experiment performed to address the five research questions (Q1–Q5). Each experiment was designed to disentangle and examine specific kinds of features or structures of models so as to establish their potential for suicide-risk prediction.

- Q1 Emotional Variance Feature Analysis: The Objective is to evaluate whether emotional signals and variance features—are sufficient for classifying suicide-risk posts. Posts were first segmented into sentences and scored using VADER. Sentence-level scores were aggregated (std, range), then combined with emotion counts from the NRC lexicon. These features were used to train a Logistic Regression model. Two version will be compare to prove whether the variance feature is helpful. which are the base-line (Logistic regression with only basic emotion counts) and variance-enhanced (base-line but add variance features) 1.
- Transformer Models performance comparison: The objective is to benchmark the performance of Transformer-based deep learning models in suicide-risk classification, using the Huggingface Transformers library with the Trainer API. This collecting performance metrics were also compared with the model in Q3 and Q5. Moreover, execution time and GPU memory usage were collected in this too for compare with the model in Q5. 2
- Q3 Hybrid Traditional ML vs Transformer: The objective is to test whether combining lexical (TF-IDF) and psychological (NRC) features improves the performance of traditional machine learning models enough to compete with Transformer models in Q2. The Traditional models to be used here are Logistic Regression, Random Forest, and Naive Bayes. After extracting and combining TF-IDF and emotion features, models were trained and evaluated so they can be compare with Transformers from Q2. SHAP values for Logistic Regression were also computed and saved for later interpretability (Q4). 3
- Linguistic marks: The objective is to identify the most influential linguistic markers that differentiate suicidal from non-suicidal posts. SHAP values of 200 samples were computed to identify the words that contributed most positively and negatively to the model's predictions. These words were visualized using word clouds. Also, the tables of top 40 of each class were made. 4
- Lightweight Model Trade-off Analysis: The objective is to compare the accuracy, execution time, and memory efficiency of lightweight traditional ML models against Transformer models. This experiment examines whether hybrid-enhanced lightweight models provide a viable trade-off for deployment scenarios, particularly in low-resource or real-time environments. The lightweight models to be used here are Logistic Regression (who has the best performance in Q3) and SDG Classifier. The progress werenot much different from Q3 but the execution time and memory usage were collected and used to compare with the Transformers. 6

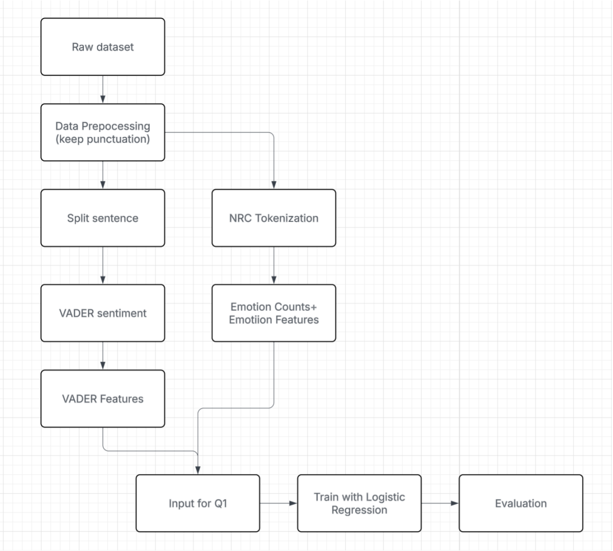
Raw dataset
Data Prepocessing (keep punctuation)
Split sentence
NRC Tokenization
VADER sentiment
Emotion Counts+ Emotiion Features
VADER Features
Input for Q1
Train with Logistic Regression
Evaluation

Fig. 1. Workflow for Q1

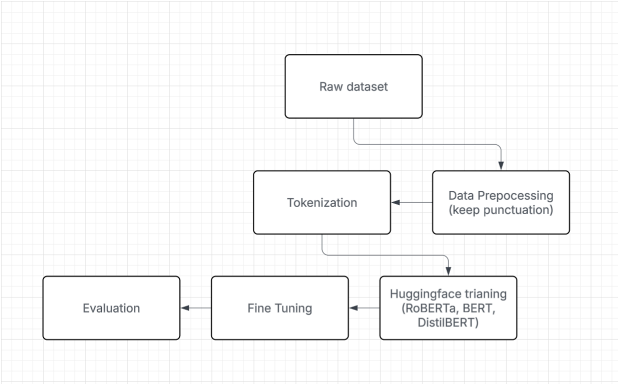
Raw dataset
Tokenization
Data Prepocessing (keep punctuation)
Evaluation
Fine Tuning
Huggingface trianing (RoBERTa, BERT, DistilBERT)

Fig. 2. Workflow for Q2

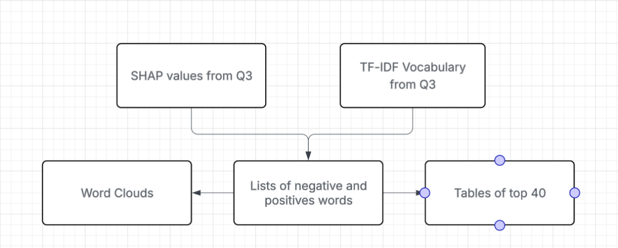
SHAP values from Q3
TF-IDF Vocabulary from Q3
Word Clouds
Lists of negative and positives words
Tables of top 40

Fig. 3. Workflow for Q3

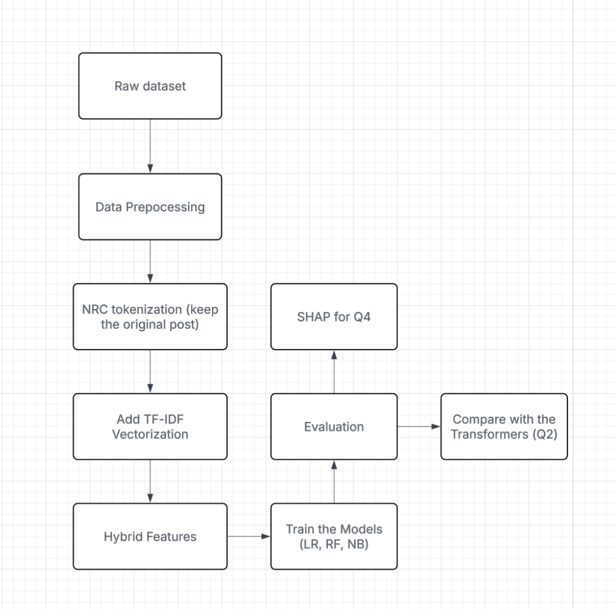
Raw dataset
Data Prepocessing
NRC tokenization (keep the original post)
SHAP for Q4
Add TF-IDF Vectorization
Evaluation
Compare with the Transformers (Q2)
Hybrid Features
Train the Models (LR, RF, NB)

Fig. 4. Workflow for Q4

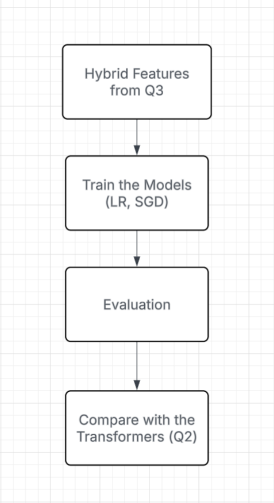
Hybrid Features from Q3
Train the Models (LR, SGD)
Evaluation
Compare with the Transformers (Q2)

Fig. 5. Workflow for Q5

D. Evaluation Metrics

1) Classification Metrics:

- Accuracy: The proportion of total predictions that were correct.
- Precision: The proportion of predicted suicidal posts that were actually suicidal. High precision indicates fewer false positives.
- Recall: The proportion of actual suicidal posts that were correctly identified. High recall indicates fewer false negatives.
- F1 Score: a measure of the harmonic mean of precision and recall.

2) Confusion Metrix: to visualize the number of true positives, false positives, true negatives, and false negatives, allowing a deeper understanding of model behavior.

3) Learning curve: Learning curves of plotting training size vs F1 score (for standard ML models) or vs steps (for Transformers) were used to examine the scaling of the performance with data size as well as with model generalization or overfitting.

4) Resource Usage Metrics (Only for Q5:

- Execution Time: Time taken for training and inference was recorded to assess deployment feasibility.
- Memory Usage: RAM usage, GPU usage, and Peak GPU usage were monitored during model training to compare resource demands between lightweight models and Transformer models.

E. Models

- Logistic Regression: A linear classifier used for its interpretability and efficiency, especially with high-dimensional sparse feature sets.
- Random Forest: An ensemble of decision trees used to capture nonlinear relationships in hybrid features.
- Naive Bayes: A probabilistic baseline suited for text classification tasks.
- SGDClassifier: A linear classifier optimized using stochastic gradient descent, selected for its lightweight footprint.
- BERT, DistilBERT, RoBERTa: Pre-trained Transformer models fine-tuned on the classification task to benchmark deep contextual language understanding against traditional approaches

IV. Results

A. RQ1: Emotional Variance Feature Analysis

TABLE I
Performance With/Without Emotional Variance

Model	Accuracy	Precision	Recall	F1-Score
Baseline (no variance)	0.79241	0.862074	0.696212	0.770316
With Emotional Variance	0.809738	0.842286	0.762184	0.800235

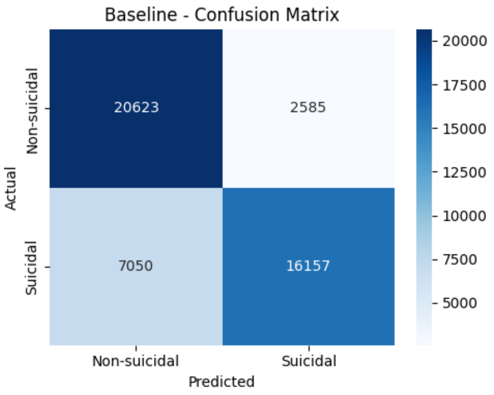

Fig. 6. Confusion Metric for Baseline

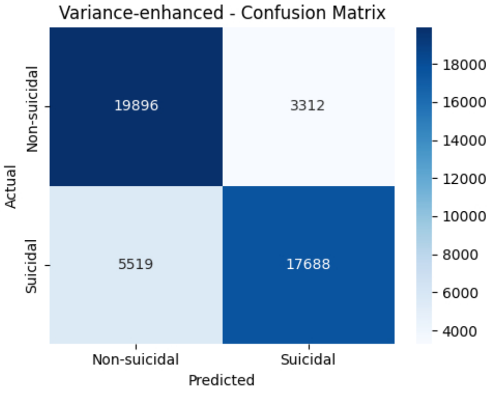

Fig. 7. Confusion Metric for variance-enhanced

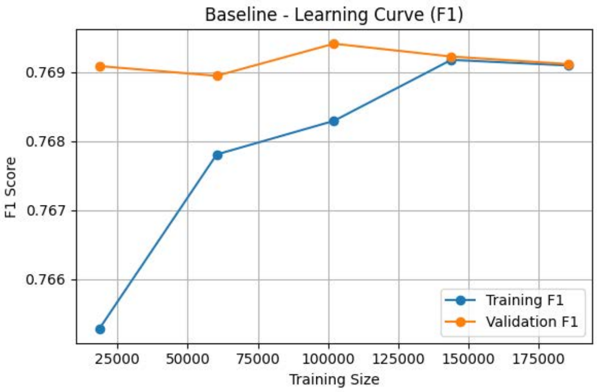

Fig. 8. Learning Curve for Baseline

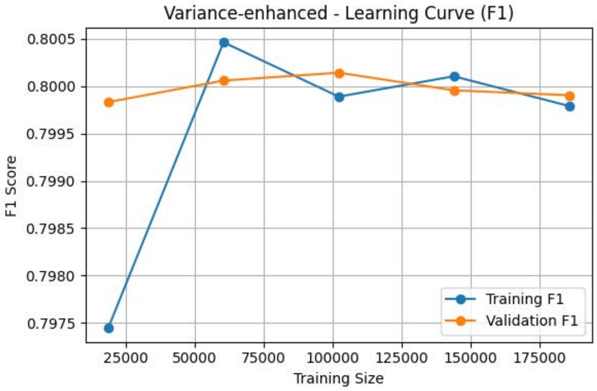

Fig. 9. Learning Curve for variance-enhanced

B. RQ2: Transformer Architecture Comparison

TABLE II
TRANSFORMER MODEL'S PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
RoBERTa	0.98173	0.97959	0.98498	0.98228
BERT	0.97652	0.97373	9.97929	0.9765
DistilBERT	0.97585	0.97276	0.97894	0.97584

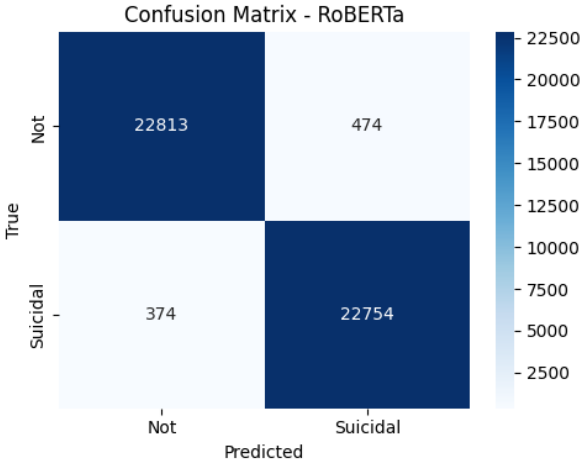

Fig. 10. Confusion Metric for RoBERTa

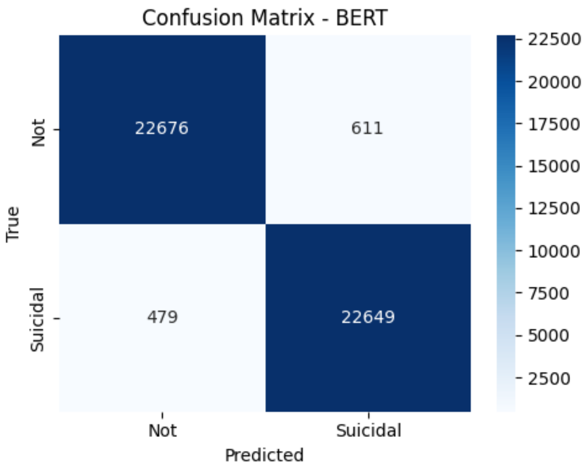

Fig. 11. Confusion Metric for BERT

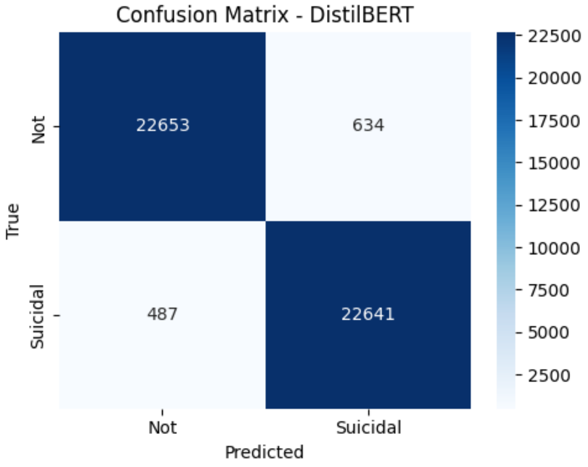

Fig. 12. Confusion Metric for DistilBERT

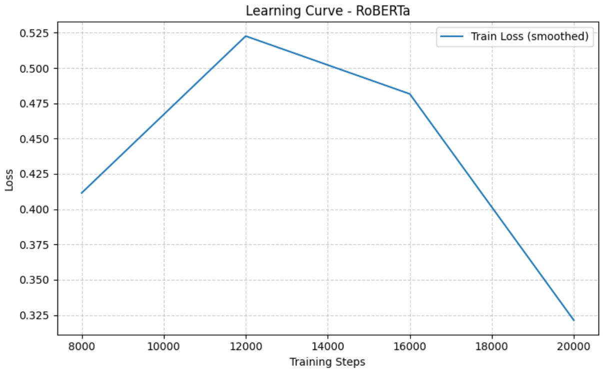

Fig. 13. Learning Curve for RoBERTa

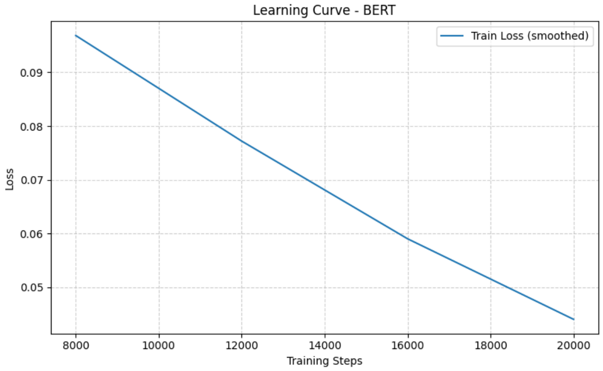

Fig. 14. Learning Curve for BERT

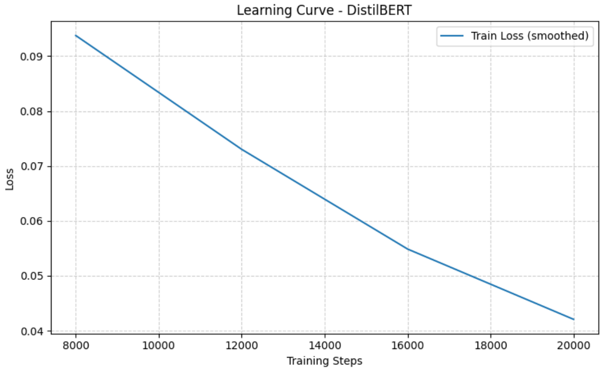

Fig. 15. Learning Curve for DistilBERT

C. Hybrid Models vs Transformer Model

for Confusion Matrices and Learning curves of Transformer models see from 10

TABLE III
HYBRID VS. DISTILBERT PERFORMANCE

Model	Accuracy	Precision	recall	F1-Score
Logistic Regression	0.93498	0.94288	0.92606	0.93439
Random Forest	0.89581	0.89448	0.89749	0.89598
Naive Bayes	0.87319	0.88255	0.86095	0.87161
RoBERTa	0.98173	0.97959	0.98498	0.98228
BERT	0.97652	0.97373	9.97929	0.9765
DistilBERT	0.97585	0.97276	0.97894	0.97584

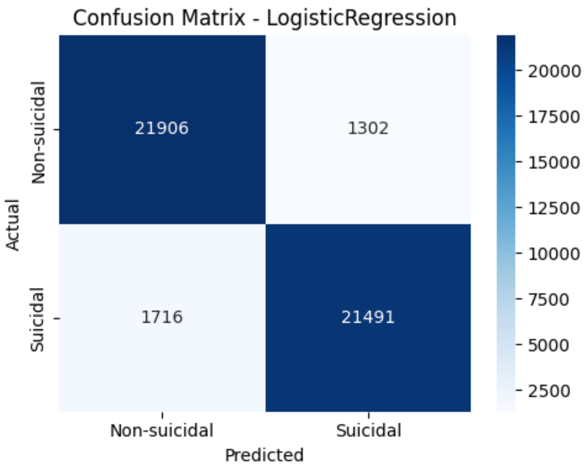

Fig. 16. Confusion Metric for Logistic Regression

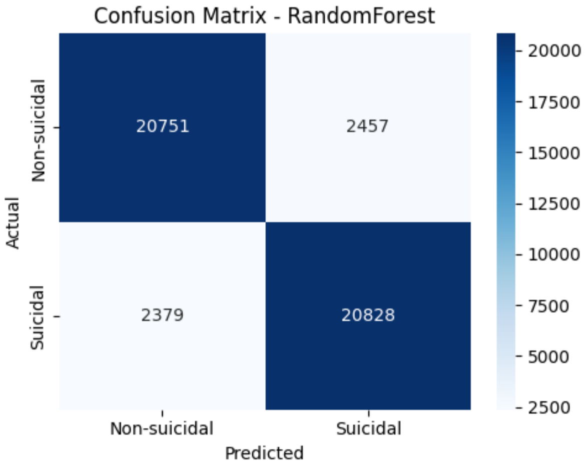

Fig. 17. Confusion Metric for Random Forest

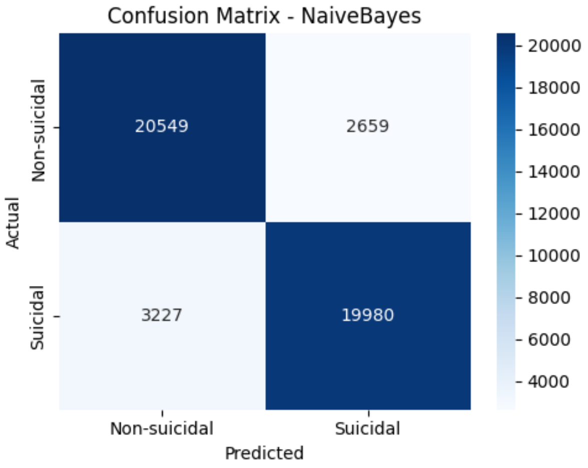

Fig. 18. Confusion Metric for Naive Bayes

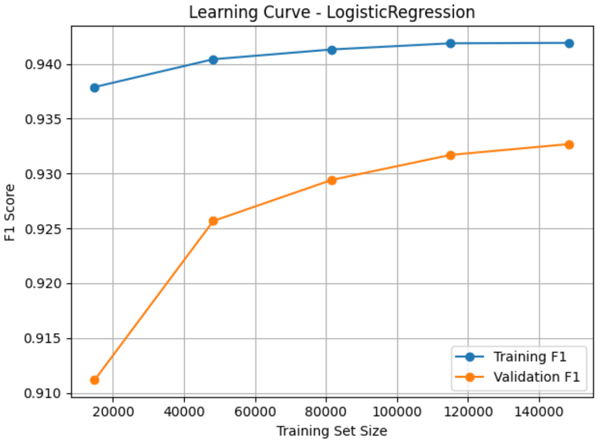

Fig. 19. Learning Curve for Logistic Regression

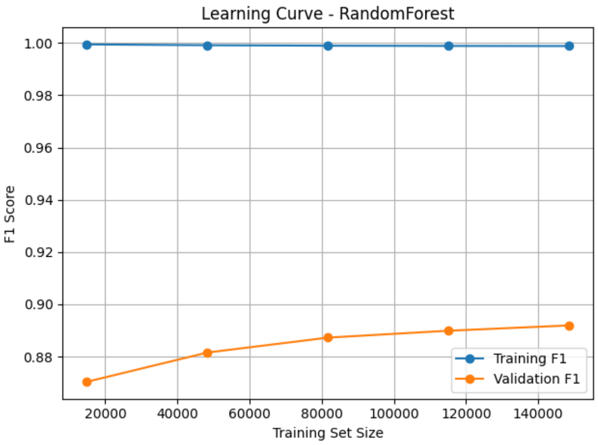

Fig. 20. Learning Curve for Random Forest

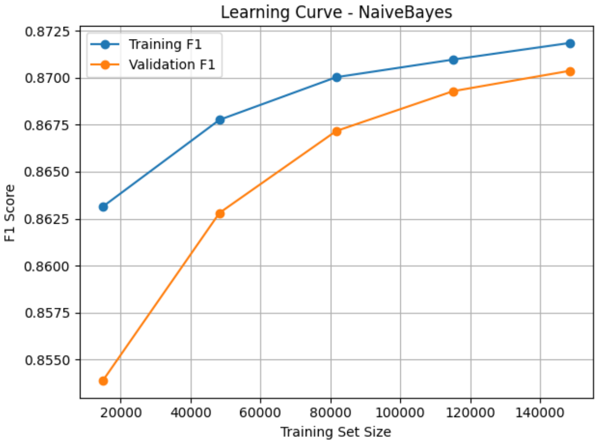

Fig. 21. Learning Curve for Naive Bayes

D. RQ4: Linguistic Marker Importance

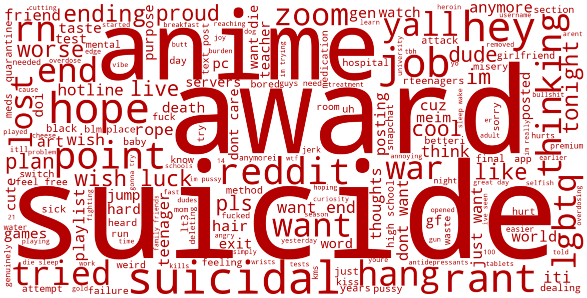

Fig. 22. Word cloud for Positive words

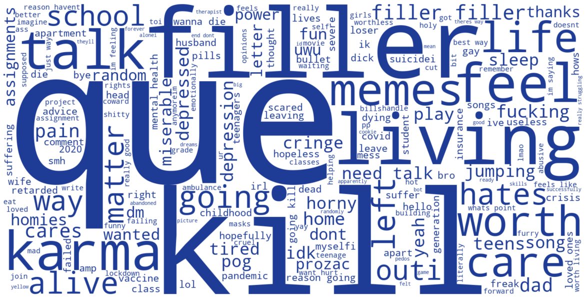

Fig. 23. Word cloud for Negative words

E. RQ5: Lightweight vs. Transformer Models

for Confusion Matrices and Learning curves of Transformer models see from 10

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1
RoBERTa	0.98173	0.97959	0.98498	0.98228
BERT	0.97652	0.97373	9.97929	0.9765
DistilBERT	0.97585	0.97276	0.97894	0.97584
Logistic Regression	0.93498	0.94288	0.92606	0.93439
SGD Classifier	0.92265	0.92059	0.92511	0.92284

TABLE V
MODEL TIME AND MEMORY USAGE COMPARISON

Model	Time (s)	Memory (MB)	GPU (MB)	Peak GPU (MB)
RoBERTa	36644	507.67	1038.82	8757.71
BERT	32118	447.01	915.85	8390.71
DistilBERT	14349	276.28	546.13	8757.71
LR	23.169	9.2226	-	-
SGD	1.4928	4.1713	-	-

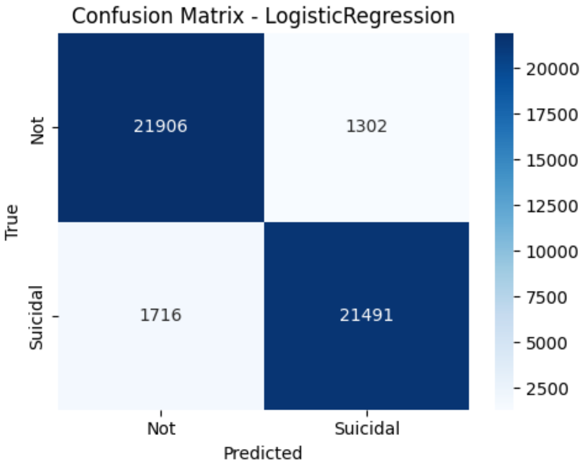

Fig. 24. Confusion Metric for Logistic Regression

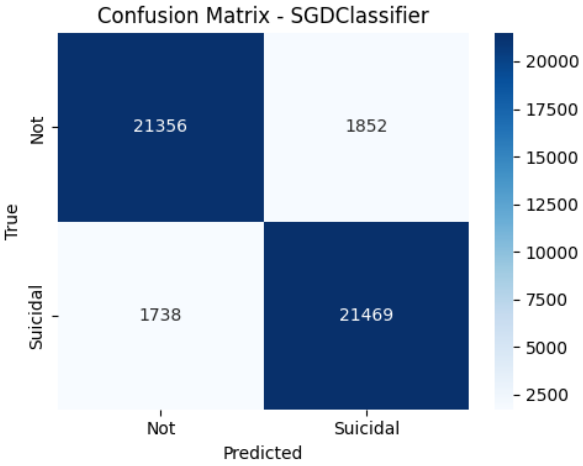

Fig. 25. Confusion Metric for SGD Classifier

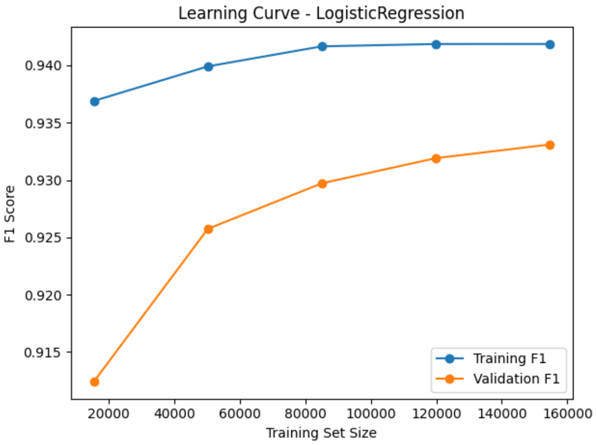

Fig. 26. Learning Curve for Logistic Regression

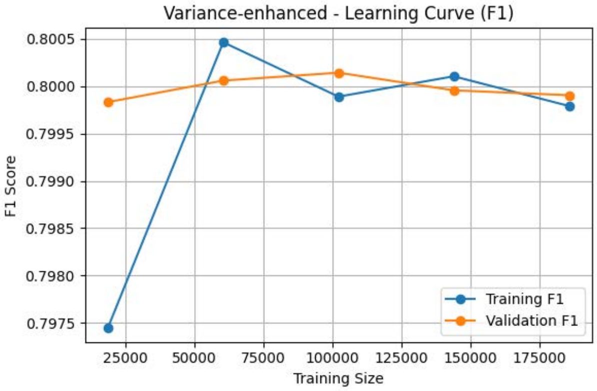

Fig. 27. Learning Curve for SGD Classifier

V. Discussion

A. Emotional Variance Feature Analysis

when comparing two confusion Matrices, the Variance-enhanced model significantly reduced False Negatives (FN) from 7,050 to 5,519. This improvement led to a noticeable increase in Recall and F1 Score, which is crucial for suicide detection tasks where missing a positive case (suicidal post) is highly undesirable. True Positives (TP) also increased by more than 1,500 instances with the addition of variance features. These should be enough prove that intra-post emotional variance can improve prediction.

B. Transformer-Based Models

All three models have magnificent performance, achieving F1 scores above 0.98 especially, RoBERTa who performed best, minimizing both false positives and false negatives. The learning curves for all three Transformer models show effective learning behavior, with steadily decreasing loss values throughout the training process. Just RoBERTa had a brief increase in loss during mid-training, likely due to learning rate adjustments or model sensitivity, but recovered and achieved the lowest final loss among all models.

C. Hybrid Traditional ML Vs Transformers

Logistic Regression is the best performance Traditional Models. And even it's improving a lot from in Q1. However, despite the performance boost, these models still lagged behind the Transformers, like RoBERTa has 5 time less false negatives. For Learning curve, Logistic Regression shows steady improvement and effective generalization, suggesting its suitability for the tasks. In contrast, Random Forest showed severe overfitting. Naive Bayes, although less prone to overfitting, was limited by its simplistic probabilistic assumptions.

D. Linguistic Marker Analysis

In SHAP analysis from the Logistic Regression Hybrid model, I found that some words that humans interpret as high-risk , like depressed, kill, and pain, actually appear in the low-risk section. But the context of the post could be ‘kill time’, ‘not depressed anymore’. Or it could be a sample bias since I only used 200 posts in this.

E. Lightweight Efficiency vs. Transformer Trade-off

Logistic Regression achieved F1 approx. 0.93, while SGD-Classifier performed slightly lower, which showed that their accuracy is at acceptable level. Both models trained within a few seconds and used minimal memory, compared to Transformers that need good GPU and large amount of time (4-8 hours or more). In spite of the performance gap persisting (RoBERTa approx. 0.98 vs. LogReg approx. 0.93), the computational efficiency trade-off is worthwhile for low-resource or real-time deployments. These findings reinforce the hypothesis that compact models, aided with hybrid features, offer an efficient substitute for scalable suicide-risk detection.

VI. Conclusion

Findings show that emotion-variance features (Q1)improve prediction and hybrid models with TF-IDF and NRC emotion features (Q3) provide large performance gains even though they still can't beat Transformers. Transformer models (Q2), and RoBERTa in particular, provide nearly perfect accuracy with the cost of high computational requirements. A SHAP analysis (Q4) of the Logistic Regression model reveals strong linguistic features and reveals the limitations of context-insensitive models. Finally, lightweight hybrid models (Q5) provide an outstanding trade-off between performance and efficiency and are therefore perfect for real-time or low-budget application.

References

[1] M. I. Bhuiyan, N. S. Kamarudin, and N. H. Ismail, “Enhanced suicidal ideation detection from social media using a cnn-bilstm hybrid model,” *arXiv preprint arXiv:2501.11094*, 2025.

[2] I. Tavchioski, M. Robnik-Šikonja, and S. Pollak, “Detection of depression on social networks using transformers and ensembles. arxiv 2023,” *arXiv preprint arXiv:2305.05325*, 2023.

[3] E. Yeskuatov, S.-L. Chua, and L. K. Foo, “Detecting suicidal ideations in online forums with textual and psycholinguistic features,” *Applied Sciences*, vol. 14, no. 21, p. 9911, 2024.

[4] R. Kaur, S. Kaur, and A. Kumar, “Sentiment analysis of twitter for detection of depression using machine learning algorithms,” *Tujin Jishu/Journal of Propulsion Technology*, vol. 44, no. 04, 2023.

[5] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, “A review of machine learning and deep learning approaches on mental health diagnosis,” in *Healthcare*, vol. 11, no. 3. MDPI, 2023, p. 285.

[6] E. Yeskuatov, S.-L. Chua, and L. K. Foo, “Leveraging reddit for suicidal ideation detection: A review of machine learning and natural language processing techniques,” *International journal of environmental research and public health*, vol. 19, no. 16, p. 10347, 2022.

[7] M. Garg, A. Shahbandegan, A. Chadha, and V. Mago, “An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts,” *arXiv preprint arXiv:2305.18727*, 2023.

[8] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, “Machine learning driven mental stress detection on reddit posts using natural language processing,” *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 80–91, 2023.

[9] S. M. Mohammad and P. Turney, “Nrc word-emotion association lexicon (aka emolex),” http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm, 2011, national Research Council Canada. Last updated: August 2022.