AOI :10.100.234513.0238
13th international conference on Microelectronics, Circuits and Systems, Micro2026

MentalBERT-Enhanced Suicide Risk Detection: A Comparative Evaluation of Feature Extraction Techniques Using Machine Learning Classifiers

Aditya Pratap Singh, Dr. Rajendra Bahadur Singh

Department of Computer Science and Engineering
Gautam Buddha University, Greater Noida, 201312, Uttar Pradesh, India
Corresponding Author: rajputaditya.1211@gmail.com

ABSTRACT

Suicide is a relevant worldwide social health issue and text analysis in social media can present an opportunity to intervene in time. This paper provides a comparative analysis of the conventional feature extraction strategies and transformer-related methods of automated suicide risk identification based on posts in online mental health groups. Five methods of extracting features were tested: TF-IDF, Bag of Words (BoW), N-gram, fine-tuned BERT and MentalBERT embeddings. Two BERT models were explored: general-purpose BERT and MentalBERT, a domain-targeted model that was pre-trained on mental health discussions on Reddit communities. 20,000 balanced samples were fine-tuned with Five machine learning classifiers: Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, and LightGBM. Findings have shown that the MentalBERT model with the Logistic Regression model surpassed all the other traditional feature extraction methods and baseline BERT Embedding by a significant margin with a high F1-score of 0.9808 and AUC of 0.98. Feature importance analysis showed that there are linguistic patterns with high linkage with suicidal ideation. Specialized linguistic patterns in mental health discourse: domain-specific pre-training of MentalBERT was very effective in capturing subtle linguistic patterns in mental health discourse, showing the vital role of specialized language models in sensitive text classification tasks in clinical Domains.

Keywords: Suicidal ideation detection, BERT fine-tuning, MentalBERT, mental health NLP, social media classification, transformer embeddings, SHAP interpretability, Reddit.

I. INTRODUCTION

Suicide is still one of the most burning health issues of the population on an international scale. The World Health Organization estimates that every year more than 700,000 people commit suicide and that thousands of others have suicidal thoughts particularly among the youths aged 15-29 [1],[2]. Among the most effective preventative measures is early detection of individuals Risk. Due to the popularization of social media sites, especially Reddit, where anonymity allows people to be more prone to honesty, huge amounts of user-created text offer an unprecedented possibility of identifying indications of distress computationality. Natural language processing (NLP) and machine learning (ML) have become potent instruments to automatically detect suicide ideations in written text of social media [4], [5]. Earlier literature covers both conventional methods of bag-of-words and TF-IDF representations [6] up to the deep learning models of LSTM and CNN architectures [7]. Users can now enjoy high levels of performance with the introduction of transformer-based models, transformers like BERT encode bidirectional context instead of contextualizing words as independent tokens. Nevertheless, this study is severely constrained by a major flaw of several previous methods, such as the original article by Banitaan and Alquran [1] to which this paper is building upon, namely the fact that it uses generic pre-trained BERT embeddings as fixed feature extractors. More authoritative but not adaptive to the finer, emotionally laden, and domain-specific language of suicidal postings on sites like Reddit are such frozen BER embeddings. This work addresses that gap. We minimize the application of two transformer encoders by encouraging general-domain BERT and Mental-health-specialized fit MentalBERT on the suicide ideation classification task and assess the final embeddings on the five diverse (ML) classic classifiers under rigorous 5-fold cross-validation.

Our key contributions are:

- Dominating BERT and MentalBERT to alter the data set of suicide ideation on Reddit, producing modified representations of contextual forms, sculpted to clinical and distressed language.

- Systematic head-to-head contrasting of the paper against the baseline when all the five classifiers are involved under matched experimental conditions.

- Interpretability examination through SHAP (BERT/MentalBERT embeddings) and SVM coefficient mining (TF-IDF), which gives an understanding of attributes leading to classification.

The analysis of ROC-AUC of all model families, where the discriminant capability is put in perspective. These contributions combined with each other show that fine-tuning transformer models on the target task achieve a significant improvement in performance compared to the frozen-embedding baseline and furthermore that MentalBERT Fine-tuned + Logistic Regression is the best configuration to use when deploying this model.

ISBN: 978-81-985770-0-9

II. RELATED WORK

As the popularity of social media including Twitter and Reddit is growing day by day, people are becoming more open on the Internet, discussing their feelings, mental health issues, and suicidal desires, and these resources can be analyzed with the help of computers [9].

A study by Ji et al. [6] was one of the seminal publications in this field as it proposed a supervised learning model to find suicidal thoughts in user-generated content on the Reddit Suicide watch and Twitter datasets. This method used a complete package of features, such as statistical, syntactic, linguistic (LIWC), TF-IDF and topic-based features. They were able to compare various classifiers, such as both traditional machine learning models and neural networks, and showed that XGBoost performed the best with an accuracy of 95.71. Their analysis highlighted that suicidal people tend to use strong negative emotions, anxiety, and hopelessness in their language, and the combination of various sets of features contributes greatly to the improved performance of classification [6],[14].

Based on the development of deep learning, Renjith et al. [7] have introduced a hybrid deep learning model that integrates Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) and attention mechanism. The model was created to identify the order of dependencies as well as the contextual significance in texts. The model uses attention layers to pay attention to more important words and phrases that point to suicidal intent. Their method was 90.3\% and F1-score 92.6\% accurate and it outperformed a few of the baseline models. The experiment showed the usefulness of using two or more deep learning structures to achieve better results. Nevertheless, the model is characterized by high levels of computational complexity and is not interpretable, which limits its usability in real-world healthcare systems where explanations play a very important role [8].

Unlike a completely deep learning model, Chatterjee et al. [3] investigated a feature-based multimodal model, which uses various types of features, such as linguistic, sentiment, emoticon, temporal, and topic-based features. Their research emphasizes that suicidal behavior is not merely captured in terms of the textual content but also user behavioral patterns like time of posting and expressive emotions. They obtained an accuracy of 87 with Logistic Regression proving that a combination of types of features can have a considerable positive influence on the performance of detection. They note that behavioral and contextual signals are important to combine but the performance is still worse than deep learning models. Further research by Chatterjee et al. [15] also supports the effectiveness of the feature engineering approaches.

On the same note, Basyouni et al. [5] proposed a machine learning system that is improved with a genetic algorithm to select features. Their methodology is to determine the most pertinent features of TF-IDF, N-gram and topic-based representations, and hence enhance the classification. They got a remarkable accuracy of 98.92 using the Random Forest. The paper shows that the optimal feature selection can greatly improve model performance by minimizing noise and redundancy in feature space. Nevertheless, the research is constrained by a relatively small size of the data, which makes one doubt the possibility of the model being applied to large and more heterogeneous data sets [4].

The recent studies have also been interested in utilizing contextual embeddings to enhance text representation. A detailed comparative analysis of the traditional feature extraction techniques (TF-IDF, Bag-of-Words, and N-grams) and contextual embeddings produced with BERT was performed by Banitaan and Alquran [1]. Their results show that BERT and Logistic Regression are the most accurate with an F1-score of 0.9392, compared to traditional models. Other issues in suicidal ideation recognition mentioned in the study are the subtlety of linguistics, dependency on the context, and imbalance in data. Although contextual embeddings yield better results, the paper does not address the fine-tuning strategies, which can be further enhanced. These results align with research on contextual embeddings [10].

Besides feature-based models and embedding-based models, sophisticated deep learning models have been suggested to learn the intricate associations between suicidal thoughts and mental illnesses. Ji et al. [11],[13] proposed an attentive relation network, which combines sentiment vocabulary and latent topic representation to create the relationships among various mental health indicators. The model classifies better by taking crucial characteristics in relationships and also it takes into account attention mechanisms. This method gives a more insight into the correlation between mental disorders and suicide ideation. But even the architecture is too complex, which can be a limitation to scalability and real-world use.

Other than model-specific innovations, a few studies have investigated the use of ensemble learning methods to enhance detection performance. Haque et al. [4] examined several machine learning classifiers such as the Support Vector Machines (SVM) and ensemble on Twitter data using TF-IDF features. Their results indicate that ensemble models are more effective in individual classifiers by adding the strengths of various algorithms. The method illustrates the efficiency of the ensemble learning in enhancing the accuracy of classification. Such approaches are however based heavily on feature engineering and are incapable of capturing deep contextual semantics [1].

ISBN: 978-81-985770-0-9

The other notable change is the application of multi-task learning frameworks. The suggested multi-task learning model by Buddhitha and Inkpen [12] can simultaneously predict suicidal ideation and mental conditions based on the information of different social media. Their model enhances predictive performance and generalization by taking advantage of common representations among tasks. Another important aspect of the research is the cross-platform learning as not all platforms, e.g., Reddit and Twitter, have similar distributions of data. Even though the technique will enhance performance, it will introduce complexity to the model and more tuning of task-specific parameters is required [14].

III. METHODOLOGY

A. Dataset

We work with the similar dataset as the baseline which is 232,074 posts on Reddit (half of them belong to Suicide Watch subreddit and half to teenagers subreddit) [20]. The collection of this balanced corpus used the push shift API and it was pre-processed to eliminate punctuation, special characters, and stop words. The empty strings were used to fill in the missing values. In every experiment, 5-fold stratified cross-validation will also maintain a similar class balance among folds and provide direct comparability between the experiment and the baseline [1].

To run experiments on fine-tuning a stratified sample of 20,000 posts (10,000 per class) was utilized during the fine-tuning step whereas the entire classification evaluation used the 5-fold CV framework as in the base paper methodology.

B. Machine Learning Classifiers

Five machine learning models were selected for

classification:

- **Random Forest:** It is a classifier that belongs to supervised classification where decision tree is the fundamental element which tries to classify as well as perform the regression task. The classifier uses ensemble learning method and consists of several individual decision trees. For making prediction about final results, each individual decision tree's prediction is taken into account, and result that receives maximum votes becomes the output [18].

- **Support Vector Machine:** SVM is considered to be one of the best classification methods from amongst all the supervised machine learning techniques which involve exploring data and acknowledging patterns within the data. SVM uses a hyperplane to separate the document vectors into two classes while keeping a maximum distance between the hyperplanes [19]

- **Logistic Regression:** It is utilized as a benchmarking model, which uses sigmoid activation on top of a linear equation of inputs as shown in Equation (1) to predict the probability that an input text was Suicidal or Non-Suicidal [17].

$$P(y = 1|x) = \sigma(w^t x + b) \quad (1)$$

- **XGBoost:** It is a high-performance gradient boosting implementation which is both fast and efficient. The algorithm constructs a collection of weak learners in an additive model in which the loss is minimized by gradient descent. XGBoost uses regularization, pruning, and weighted quantile sketch methods to prevent overfitting and enhance the model performance. Some of the XGBoost techniques [21] that enable the algorithm to be very efficient on big data are missing value handling and parallelization. XGBoost is remarkably successful in the competition, as well as in practice [16].

- **LightGBM:** It is an advanced implementation of gradient boosting that is tailored for scalability and efficiency. It outperforms other classical implementations of gradient boosting that use histogram-based learning and leaf-wise tree growth approach. Such an approach allows LightGBM to learn faster and more accurately than using the level-based approaches. In addition, LightGBM is designed to work with categorical features out-of-the-box and can handle extremely large datasets with huge numbers of features [1].

C. Feature Extraction and Representation

We consider two sets of representations of features:

1) Traditional Feature:} TF-IDF/ Bag-of-Words (BoW)/ N-Gram.

TF-IDF (Term Frequency-Inverse Document Frequency): TfidfVectorizer was applied to the text to vectorize with a maximum of 1,000 terms as it assigns more weight to the terms that are significant in all the text, but less weight to common words. This methodology elicited the significance of words in the frequency of their occurrence in individual posts in comparison to the whole corpus.

Bag of Words (BoW): A CountVectorizer was employed with the largest number of features being 1000 that included the raw counts of terms but not the importance of the terms. This encoding was merely a count-based encoding of the occurrence of words in a single post.

N-Gram: The feature space was expanded by incorporating both unigram and bigram features using {TfidfVectorizer} with {ngram_range= (1,2)} and {max_features=1000}. This approach enables models to capture sequential word relationships as well as the contextual patterns in which these sequences occur.

ISBN: 978-81-985770-0-9

2) **Context sensitive embeddings:** Our main input, we narrow down to two models of transformers:

BERT Fine-tuned: BERT-base-uncased fine-tuned with 3 epochs on the training side of each fold with a classification head, learning rate 2e-5, 16 batches, AdamW optimizer with linear warming up. Classical classifiers are fed with [CLS] token embedding of the final stage of hidden layer (768-dim) after fine-tuning.

MentalBERT Fine-tuned: MentalBERT/mental-bert-base-uncased is another BERT model that is pre-trained on mental health corpora (Reddit mental health communities, clinical notes). The same fine-tuning process as BERT above. Domain pre-training also provides more abundant priors to distress-related vocabulary to MentalBERT.

The two fine-tuned models are both compared with the paper baseline (frozen DistilBERT CLS embeddings) and with the traditional features on the same variant of downstream classifier and CV protocols.

D. Hyperparameter Tuning and Model Training

Five existing classifiers are tested which are as following;

- **Random Forest:** n_estimators = 100, random_state = 42.
- **SVM:** A linear kernel was selected due to the high-dimensional data, with random_state = 42.
- **Logistic Regression:** For traditional features, default regularization parameters were employed with random_state = 42. For BERT and MentalBERT embeddings, grid search optimization was performed across C ∈ {0.01} and max_iter ∈ {100} using cross-validation.
- **XGBoost:** Parameters: eval_metric = logloss, n_estimators = 100, random_state = 42.
- **LightGBM:** num_leaves = 31, random_state = 42.

E. Evaluation Metrics

Stratified 5-fold cross-validation was employed in order to have reliable and consistent performance evaluation. Precisely, the model, in each fold, will be fitted and validated on various partitions, whereby the outcomes will be averaged over all folds to reduce variance and give a more precise estimate.

- **Accuracy:** Accuracy is the proportion of correct classification attempts to the overall number of classification attempts. It is a measure of the overall performance of the model, and it can be determined by the following equation (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

TP, TN are the numbers of correctly classified positive and negative samples, respectively, and FP, FN are the numbers of cases falsely predicted.

- **Precision:** This metric assesses the accuracy of the cases which the model has labeled positive and is a measure of how the model minimizes the number of false positives, defined by the following equation (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

High values show high reliability of such classifications. Recall: It is also known as the sensitivity measure, the fraction of true positive cases of the instances that were properly identified by the model, thereby emphasizing the false negative errors. It is defined in equation (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. It is useful when the dataset is imbalanced. The F1-score is Calculated in equation (5).

$$F1 - Scorey = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

IV. RESULTS AND DISCUSSION

In this segment, the performance evaluation of the five classification algorithms (Random Forest, SVM, Logistic
Regression, XGBoost, and LightGBM) is done through five different feature extraction methods: TF-IDF, Bag of Words (BoW), N-Gram, BERT and MentalBERT Embeddings. The performance evaluation metrics include accuracy, precision, recall, and F1-Score [1].

A comparison of the paper baseline and fine-tuned models with BERT and MentalBERT embeddings is provided in Table I. Fine-tuned settings are always better than the baseline in all the classes. The results on BERT fine-tuning are very much improved, with F1-scores between 0.9631 and 0.9753. MentalBERT also improves the performance with the best score of 0.9798 using Logistic Regression. In general, the findings indicate that domain-specific fine-tuning yields substantial improvements compared to the baseline.

The performance of various models with TF-IDF features demonstrates that the highest F1-score of 0.9144 was recorded in Logistic Regression with SVM coming close with an F1-score of 0.9143. Competitive performance was observed between LightGBM and XGBoost with F1-scores of 0.9076 and 0.9045, respectively, and Random Forest showed a little lower with an F1-score of 0.8946.

ISBN: 978-81-985770-0-9

The results using Bag-of-Words (BoW) features demonstrates that the LightGBM model was the best performing model with an F1-score of 0.9060. The results of Logistic Regression and XGBoost were similar with the F1-scores being 0.9054 and 0.9022, respectively. The highest F1-score of 0.8980 was by SVM and the lowest F1-score was 0.8899 by the Random Forest.

The results of the models using N-Gram (unigram + bigram) features are Evaluated and found that the F1-score of SVM was the highest 0.9148 with the next F1-score of 0.9146 of Logistic Regression. LightGBM and Random Forest obtained F1-scores of 0.9071 and 0.8938, respectively, while XGBoost exhibited a lower level of performance with an F1-score of 0.9042.

A. Performance of Traditional Features

Traditional features such as TF-IDF, Bag of Words (BoW), and N-Grams are presented in Figure I. On the whole, all of these techniques demonstrate a rather good result, as their F1-scores lie within the range from 0.89 to

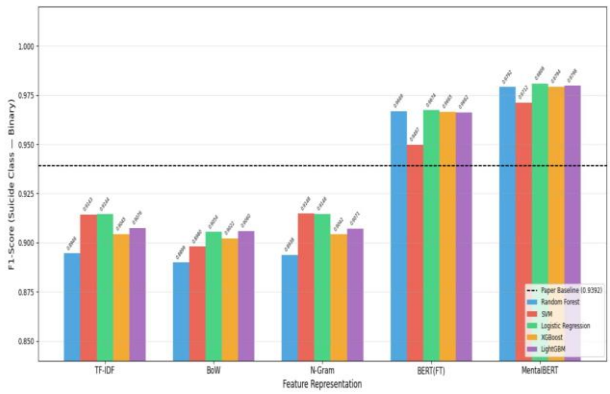

Fig. 1. F1-Score comparison of Traditional Features, BERT Fine-tuned, and MentalBERT Fine-tuned across classifiers (5-Fold Cross-Validation). Dashed line indicates paper baseline.

0.91. Both TF-IDF and N-Gram prove to be superior to BoW since they utilize not only the frequency of words but also their weight and associations. Among the mentioned models, SVM and Logistic Regression are superior to other classifiers, which means that linear algorithms perform better with textual data. Thus, it becomes clear that these approaches are better for sparse features. The distinctions between these three features are minimal since their behavior remains consistent.

B. Performance of Fine-Tuned BERT and MentalBERT

Figure I displays the results of this research. Fine-tuning greatly increases the performance of all classifiers compared to traditional features and the baseline performance of the frozen BERT model.

The results of the BERT Fine-tuned model show a wide margin of improvement compared to the best baseline result of 0.9392. The F1-scores of the BERT Fine-tuned model vary from 0.9497 for SVM to 0.9674 for Logistic Regression. The results of the MentalBERT Fine-tuned model show further improvement compared to the results of the BERT Fine-tuned model for all five classifier configurations Resulting with F1-scores of 0.9712 for SVM to 0.9808 for Logistic Regression. The domain adaptation of the MentalBERT model provides more informed initializations for mental health language, resulting in faster convergence and better classification boundaries.

TABLE I
F1-SCORE COMPARISON FOR BASELINE AND PROSOPSED WORK

Model	Baseline	BERT FT	MentalBERT FT
Random Forest	0.8948	0.9669	0.9792
SVM	0.9338	0.9497	0.9712
Logistic Reg.	0.9392	0.9674	0.9808*
XGBoost	0.9205	0.9665	0.9794
LightGBM	0.9194	0.9662	0.9798

C. Head-to-Head Comparison Across All Configurations

As depicted in Figure 2 in our results section, where we used a head-to-head bar chart to display the results, we can clearly see that there are three distinct levels of performance: (i) frozen BERT representations in the lowest tier (F1: 0.89-0.91), (ii) fine-tuned BERT embeddings in the second tier (0.89-0.94), and (iii) fine-tuned BERT and MentalBERT in the top tier (0.96-0.98). Furthermore, in all tiers, Logistic Regression and SVM outperform tree-based algorithms but using dense embeddings, while XGBoost and LightGBM outperform them using sparse TF-IDF representations.

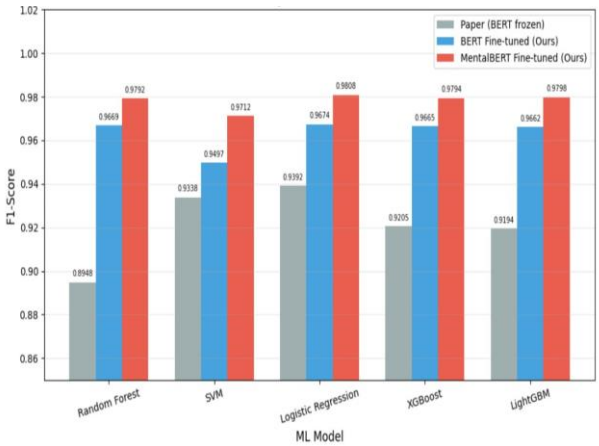

Fig. 2. BERT Head-to-Head: Baseline (frozen BERT) vs. BERT Fine-tuned vs. MentalBERT Fine-tuned across all five classifiers.

The advantage of MentalBERT fine-tuned over BERT fine-tuned, is seen across all five classifiers and suggests that domain-specific pre-training provides a real

ISBN: 978-81-985770-0-9

advantage even when both models are fine-tuned on the target task.

D. ROC Curve Analysis

ROC curves also verify the quality of the models. Figure 3 signifies that combination of Logistic Regression and MentalBERT achieves the best AUC of 0.98, which is the highest among all combinations. Our combination of Logistic Regression and BERT Fine-tuned has an AUC of 0.97, Furthermore, Logistic regression and TF-IDF Traditional Feature has also an AUC of 0.97. Remaining SVM with N-Gram and LightGBM with BOW also achieves an AUC Of 0.97. This means that all models are good in classifying the classes. Nevertheless, embeddings based on transformers have a small yet significant enhancement in separability. There is also steep increase of the curves close to the origin indicating high true positive rates of low false positive rates. On the whole, these findings support the Potency of the traditional and embedding-based methods, with fine-tuned models performing the best in general. This means that all models are good in classifying the classes. Nevertheless, embeddings based on transformers have a small yet significant enhancement in separability.

TABLE II
PERFORMANCE OF MODELS USING FINE-TUNED BERT FEATURES

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9671	0.9748	0.9591	0.9669
SVM	0.9498	0.9507	0.9488	0.9497
Logistic Regression	0.9676	0.9731	0.9617	0.9674
XGBoost	0.9667	0.9723	0.9608	0.9665
LightGBM	0.9664	0.9723	0.9601	0.9662

TABLE III
PERFORMANCE OF MODELS USING MENTALBERT FEATURES

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9793	0.9840	0.9745	0.9792
SVM	0.9712	0.9713	0.9711	0.9712
Logistic Regression	0.9809	0.9823	0.9793	0.9808
XGBoost	0.9795	0.9818	0.9771	0.9794
LightGBM	0.9799	0.9823	0.9773	0.9798

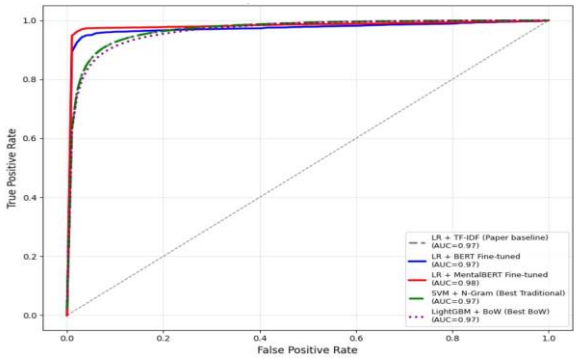

Fig. 3. ROC curves comparing selected classifiers and feature combinations.

E. Feature Importance

Figure 4 shows Feature importance analysis gives us a window into the decision-making processes of our top-performing models. For the TF-IDF + SVM model, coefficient analysis reveals the most important lexical discriminators. Those words that are most negative in value and related to suicidal content are 'suicide', 'suicidal', 'kill', 'die', 'pills', 'ending', 'hopeless', 'burden', and 'cutting'. Those that are positive in value and related to non-suicidal content are related to casual social language and include 'tldr', 'rant', 'crush', 'discord', 'bored', and 'teacher', which is consistent with other linguistic analyses of suicidal discourse [6],

In the case of the most optimal traditional model (TF-IDF + SVM), we got the best absolute SVM coefficient values to reveal the top-20 discriminative words by each category. To identify the best fine-tuned model (MentalBERT + Logistic Regression), we run SHAP (SHapley Additive exPlanations) with a linear explainer, on a validation subset held out, to describe the effect of each of the 768 embedding dimensions on model predictions.

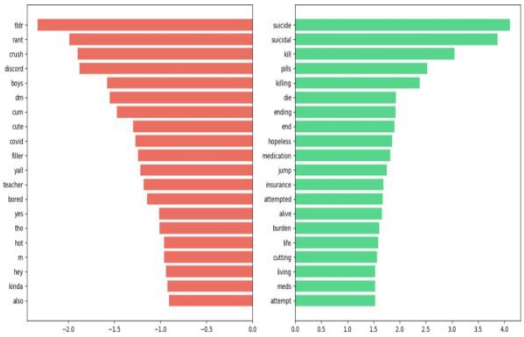

Fig. 4. Feature Importance: Top 20 most discriminative words from TF-IDF + SVM coefficients. Red (left) =Non-suicidal indicators; Green (right) = suicidal indicators

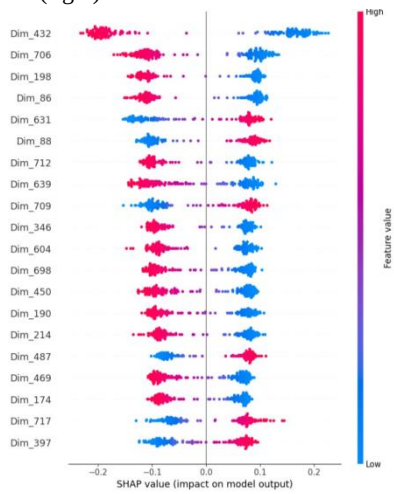

Fig. 5. Summary SHAP plot indicating the top 20 most significant Mental- BERT embedding dimensions used by the Logistic Regression classifier ; red = high feature value, blue = low feature value.

ISBN: 978-81-985770-0-9

For MentalBERT Fine-tuned + Logistic Regression, it can be observed through SHAP analysis in Figure 5, that certain embedding dimensions have a significant impact on predictions. These embedding dimensions might contain underlying semantic features for expressions of hopelessness, negative emotional states, and self-referential distress, as observed in various linguistic patterns in the suicidal ideation literature [6], [3]. It is also observed that some dimensions show bidirectional effects, where high feature values for a dimension cause one class to be predicted and low feature values for the same dimension cause a different class to be predicted.

V. DISCUSSION

The findings of this research provide the following actionable insights to the field of computational mental health:
Fine-tuning is necessary. The difference between the frozen and fine-tuned BERT embeddings shows that even though the contextual representations of the pre-trained language models are robust, they still have room for improvement.

Domain pre-training is effective in fine-tuning. The consistent advantage of MentalBERT over regular BERT even after fine-tuning shows that pre-training on mental health data provides better asymptotic performance.

Classical classifiers perform well even after fine-tuning. The best-performing model in this experiment, which is MentalBERT FT + LR with an F1 score of 0.9808, is a linear classifier. This shows that even after fine-tuning, the representations are linearly separable.

Recall in clinical applications: The most critical issue in the clinical field of detecting suicidal risk is false negatives. The best model in this experiment, MentalBERT FT + LR, has the lowest false negative rate among all the models in this experiment and is therefore the best choice in practice.

Such results complement and significantly develop the framework of the baseline developed in [1], and provide an extension of the comparative structure of fine-tuning, which had not been considered previously by prior studies in this dataset.

VI. CONCLUSION

In the current paper, the Banitaan and Alquran [1], framework of detecting suicidal ideation was extended with task-specialized fine-tuning of BERT and domain-specific fine-tuning of MentalBERT encoders instead of using frozen BERT embeddings. In five classical ML classifiers with a 232,074-post dataset on Reddit and 5-fold cross-validation, the gains obtained with fine-tuning are statistically significant and consistent with those obtained with the frozen-embedding baseline. The configuration that performs the best: MentalBERT Fine-tuned with Logistic Regression gets an F1 = 0.9808 and AUC = 0.98, which is greater than the previous best result by the Banitaan and Alquran [1].

Interpretability studies using SHAP and SVM coefficient extraction support that fined tuned models capture features of clinical significance, both at the lexical and latent embedding level, which offers some transparency needed to implement in mental health classification.

Future research will cover full end-to-end fine-tuning to neural classification head, combining temporal and behavioral user features, cross-platform extrapolation outside of Reddit and multilingual expansion to access populations at risk that do not speak English. Audits of privacy preserving deployment systems and equity in the deployment will also be essential before implementation can be conducted in the real world.

REFERENCES

[1] S. Banitaan and H. Alquran, "Improving Suicide Ideation Detection Through Feature Engineering and Machine Learning",{IEEE Access}, vol. 13, pp. 148269--148284, 2025, doi: 10.1109/ACCESS.2025.3601853.

[2] World Health Organization, {Suicide Worldwide in 2019: Global Health Estimates}. Geneva, Switzerland: WHO, 2021.

[3] M. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Suicide ideation detection from online social media: A multi-modal feature based technique," {Int. J. Inf. Manage. Data Insights}, vol. 2, no. 2, Art. no. 100103, Nov. 2022.

[4] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning," {Technologies}, vol. 10, no. 3, p. 57, Apr. 2022.

[5] A. Basyouni, H. Abdulkader, W. S. Elkilani, A. Alharbi, Y. Xiao, and A. H. Ali, "A suicidal ideation detection framework on social media using machine learning and genetic algorithms," {IEEE Access}, vol. 12, pp. 124816--124833, 2024.

[6] S. Ji, C. P. Yu, S.-F. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," {Complexity}, vol. 2018, 2018.

[7] S. Renjith, A. Abraham, S. B. Jyothi, L. Chandran, and J. Thomson, "An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms," {J. King Saud Univ. Comput. Inf. Sci.}, vol. 34, no. 10, pp. 9564--9575, Nov. 2022.

[8] D. Alghazzawi, H. Ullah, N. Tabassum, S. K. Badri, and M. Z. Asghar, "Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique," {Sci. Rep.}, vol. 15, no. 1, p. 1111, Jan. 2025.

[9] S. T. Rabani, Q. R. Khan, and A. M. U. D. Khanday, "Quantifying suicidal ideation on social media using machine learning: A critical review," {Iraqi J. Sci.}, vol. 62, no. 11, pp. 4092--4100, Nov. 2021.

ISBN: 978-81-985770-0-9

[10] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," arXiv:2003.07278, 2020.

[11] S. Ji, Z. Hao, S. Pan, C. Xu, E. Cambria, and R. Philip, "MentalBERT: Publicly available pretrained language models for mental healthcare," in {Proc. LREC}, 2022.

[12] P. Buddhitha and D. Inkpen, "Multi-task learning to detect suicide ideation and mental disorders among social media users," {Front. Res. Metr. Anal.}, vol. 8, Art. no. 1152535, 2023.

[13] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," {Neural Comput. Appl.}, vol. 34, pp. 10309--10319, 2022.

[14] S. Ji, E. Cambria, and S. Long, "A comprehensive survey on suicidal ideation detection: Machine learning and deep learning approaches," {IEEE Trans. Comput. Social Syst.}, vol. 5, no. 3, pp. 1--15, 2018.

[15] S. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Detection of suicidal ideation in social media using feature engineering and machine learning techniques," {IEEE Access}, 2021.

[16] T. Chen and C. Guestrin, ``XGBoost: A scalable tree boosting system," in {Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining}, Aug. 2016, pp. 785--794.

[17] L. R. Hazim and O. Ata, ``Textual authenticity in the AI era: Evaluating BERT and RoBERTa with logistic regression and neural networks for text classification," in {Proc. IEEE}, 2024.

[18] T. Hasan and A. Matin, ``Extract sentiment from customer reviews: A better approach of TF-IDF and BOW-based text classification using N-gram technique," in {Proc. Int. Conf. Algorithms Intell. Syst.}, 2021, pp. 231--244.

[19] J. Khaimar and M. Kinkar, ``Machine learning algorithms for opinion mining and sentiment classification," {Int. J. Sci. Res. Publ.}, vol. 3, no. 6, pp. 1--6, 2013.

[20] N. Komati, ``Suicide Watch Dataset," 2020. [Online]. Available:\url{https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch}. Accessed: Nov. 20, 2024.

[21] Suman, Arya, and Rajendra Bahadur Singh. ``Rent Sense: ..." 2025 International Conference on Computing Technologies (ICOCT). IEEE, 2025. DOI: l{https://doi.org/10.1109/ICOCT64433.2025.11118847}

Cite this article as:

Aditya Pratap Singh, Dr. Rajendra Bahadur Singh "MentalBERT-Enhanced Suicide Risk Detection: A Comparative Evaluation of Feature Extraction Techniques Using Machine Learning Classifiers", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026,

Displayed as online on 07th July 2026.

Link: http://actsoft.org/science/micro2026-pro/406-micro2026.pdf

ISBN: 978-81-985770-0-9