



EFFICIENT AND INTERPRETABLE DETECTION OF SUICIDAL IDEATION IN SOCIAL MEDIA VIA GLOVE EMBEDDINGS, RECURSIVE FEATURE ELIMINATION, AND ALIGHTWEIGHT TRANSFORMER

Houari Benabid^{*1}, Moustafa Maaskri², Mohamed Goismi³

¹Department of Computer Science, Artificial Intelligence and Systems Research Laboratory, Ibn Khaldoun University of Tiaret, Post box 78 Tiaret, Algeria.

²Department of Computer Science, Artificial Intelligence and Systems Research Laboratory, Ibn Khaldoun University of Tiaret, Post box 78 Tiaret, Algeria.

³ Department of Science and Technology, University Ibn Khaldoun – Tiaret, Algeria.

¹<https://orcid.org/0009-0004-1784-9227>, ²<http://orcid.org/0009-0000-5268-1133>, ³<http://orcid.org/0000-0001-9535-1142>

E-mail: *houari.benabid@univ-tiaret.dz, moustafa.maaskri@univ-tiaret.dz, mohamed.goismi@univ-tiaret.dz

ARTICLE INFO

Article History

Received: February 01, 2026

Reviewed: March 02, 2026

Accepted: April 01, 2026

Published: May 29, 2026

Keywords:

Natural language processing,
Sentiment analysis,
GloVe embeddings,
Recursive feature elimination,
lightweight transformers,
transformer-based classification,
social media analysis,
Suicidal ideation,
Suicide prevention.

ABSTRACT

Suicidal ideation expressed on social media offers an opportunity for earlier risk screening; however, practical systems must maximize recall, remain computationally tractable, and provide signals that clinicians can meaningfully interpret. We propose a hybrid pipeline that (i) represents tweets using pretrained 300-dimensional GloVe vectors, (ii) aggregates token embeddings with mean and max pooling, (iii) applies Recursive Feature Elimination (RFE) to select 228 informative semantic dimensions from a 600 dimensional pooled representation, and (iv) trains a compact 6-layer transformer encoder on the selected features for binary classification. Experiments on a combined dataset of 145,230 tweets (26.5% suicidal) demonstrate that the proposed model achieves 94.3% accuracy and an F1-score of 0.944 with 95.1% recall. The architecture is resource-efficient, using 23M parameters and delivering 3.7imes faster inference than BERT-base. Feature analysis indicates that the most predictive dimensions cluster around clinically relevant themes such as hopelessness, psychological pain, isolation, and perceived burden, offering a more transparent bridge between model behavior and suicide risk theory. These results suggest that feature selection paired with lightweight attention models can yield high-performing screening tools suitable for large-scale monitoring when deployed under rigorous ethical and clinical governance.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Suicide represents one of the most pressing public health concerns worldwide, with the World Health Organization reporting approximately 700,000 deaths by suicide annually [1]. The tragic loss of life, coupled with the profound impact on families and communities, underscores the urgent need for effective prevention strategies. In recent years, social media platforms have emerged as critical spaces where individuals express distress, mental health struggles, and suicidal ideation, often before seeking professional help or confiding in close contacts [2].

Twitter (now X), with its real-time, public nature, provides a unique window into the mental states of users worldwide. Unlike traditional clinical settings where individuals may be reluctant to disclose suicidal thoughts due to stigma or fear of intervention, social media platforms often serve as spaces for candid expression [3]. This phenomenon has motivated researchers to develop computational methods for detecting suicide risk signals in social media text, with the ultimate goal of enabling timely intervention and support. Recent surveys synthesize the main computational approaches and highlight open challenges [4].

The application of Natural Language Processing (NLP) and machine learning to mental health has gained significant momentum over the past decade [5]. Early approaches relied on lexicon-based methods and traditional machine learning classifiers trained on handcrafted features [6]. However, these methods often struggled with the nuanced, context-dependent nature of language related to mental health and suicide. The advent of deep learning, particularly word embeddings and large language models, has revolutionized text classification tasks, offering more sophisticated representations of semantic meaning [7].

Despite these advances, several challenges persist in the field of suicidal ideation detection. First, the imbalanced nature of datasets, where suicidal content represents a small minority of overall social media discourse, poses significant difficulties for model training [8]. Second, the ethical considerations surrounding privacy, consent, and the potential for false positives or negatives demand careful methodological design [9]. Third, the interpretability of complex deep learning models remains crucial in clinical contexts, where understanding the reasoning behind predictions can inform intervention strategies [10].

This study addresses these challenges by proposing a hybrid approach that combines the semantic richness of GloVe embeddings with the discriminative power of Recursive Feature Elimination (RFE) and a lightweight transformer architecture trained from scratch on selected features. Note: Our "transformer" is not a pre-trained Large Language Model (LLM) like BERT or GPT. Instead, we train a compact 6-layer transformer encoder from random initialization on 228-dimensional numerical feature vectors (selected via RFE from GloVe embeddings), not on token sequences. Our methodology is designed to achieve three primary objectives:

1. To develop a robust classification system capable of accurately identifying tweets expressing suicidal ideation while minimizing false positives and false negatives.
2. To identify the most informative linguistic features associated with suicidal content through systematic feature selection, contributing to theoretical understanding of suicidal discourse.
3. To evaluate our approach on established benchmark datasets, enabling direct comparison with existing methods and facilitating reproducibility.

The contributions of this work are multifaceted:

Methodological Innovation: We introduce a novel pipeline that integrates GloVe embeddings for capturing semantic relationships, RFE for identifying the most predictive features, and a lightweight transformer classifier trained from scratch on these selected features. This combination leverages the strengths of each component while mitigating their individual limitations.

Empirical Validation: Through extensive experiments on benchmark datasets, we demonstrate that our approach achieves state-of-the-art performance in suicidal ideation detection, with significant improvements in precision and recall compared to baseline methods.

Feature Interpretability: By employing RFE, we identify specific linguistic patterns and semantic features most strongly associated with suicidal content, providing insights that can inform both computational models and clinical understanding.

Practical Application: Our framework is designed with deployment in mind, balancing accuracy with computational efficiency to enable real-time monitoring and intervention systems.

The remainder of this paper is organized as follows: Section II reviews related literature on suicide detection in social media, word embeddings, feature selection methods, and transformer applications in NLP. Section III details our methodology, including data preprocessing, feature extraction using GloVe, feature selection via RFE, and lightweight transformer-based classification trained from scratch. Section IV presents our experimental setup, datasets, evaluation metrics, and results. Section V discusses the implications of our findings, limitations, and ethical considerations. Finally, Section VI concludes the paper and outlines directions for future research.

II. LITERATURE REVIEW

II.1 SUICIDE DETECTION IN SOCIAL MEDIA

The application of computational methods to detect suicide risk in social media has evolved significantly over the past decade. Early work by Jashinsky et al. [11] demonstrated correlations between suicide-related tweets and actual suicide rates, validating social media as a viable data source for public health surveillance. Coppersmith et al. [12] expanded this line of inquiry by developing methods to quantify signals of mental health conditions, including suicidal ideation, from Twitter posts.

The CLPsych (Computational Linguistics and Clinical Psychology) workshop series has been instrumental in advancing the field, providing shared tasks and benchmark datasets. The 2015 CLPsych shared task [13] focused on identifying Twitter users at risk for mental health conditions, establishing evaluation protocols that have since become standard. Subsequent work by Shing et al. [14] incorporated expert annotations to improve the reliability of training data.

Recent studies have employed increasingly sophisticated NLP techniques. Tadesse et al. [8] utilized deep learning models including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, achieving accuracies above 90% on suicide detection tasks. Ji et al. [7] proposed a supervised learning framework combining user-level and tweet-level features, demonstrating the importance of multi-scale analysis.

II.2. WORD EMBEDDINGS AND GLOVE

Word embeddings represent words as dense vectors in continuous space, capturing semantic and syntactic relationships. Word2Vec [15] and GloVe [16] emerged as foundational techniques, with GloVe specifically designed to leverage global word co-occurrence statistics. Unlike Word2Vec, which relies on local context windows, GloVe constructs an explicit word-word co-occurrence matrix from the entire corpus, optimizing embeddings to encode these statistics in vector space.

The advantages of GloVe for text classification have been demonstrated across numerous domains. Bojanowski et al. [17] showed that pre-trained embeddings can be fine-tuned for specific tasks, improving performance on downstream classification problems. In the mental health domain, Yates et al. [18] successfully employed GloVe embeddings to detect depression-related language on Reddit, highlighting the technique's ability to capture nuanced emotional and psychological content.

Recent work has compared GloVe with contextual embeddings from transformer-based models. While contextualized representations from BERT [19] and GPT [20] have shown superior performance on many tasks, GloVe remains competitive for specific applications, particularly when computational efficiency and interpretability are priorities [21].

II.3. FEATURE SELECTION AND RFE

Feature selection is crucial in high-dimensional text classification, where the vocabulary size can exceed hundreds of thousands of unique terms. Recursive Feature Elimination (RFE) [22] is a backward selection method that iteratively removes features based on their importance to the model, typically measured by weights in linear classifiers or feature importance scores in tree-based models.

The application of RFE to text classification has been explored extensively. Forman [23] compared various feature selection methods for text categorization, demonstrating that wrapper methods like RFE often outperform filter methods. Tang et al. [24] applied RFE to social media analysis, showing improvements in both accuracy and computational efficiency.

In the mental health domain, feature selection serves dual purposes: improving model performance and enhancing interpretability. Resnik et al. [25] emphasized the importance of identifying specific linguistic markers of mental health conditions. Chancellor et al. [26] developed taxonomy of self-harm content on Instagram, highlighting features that distinguish concerning content from general distress expression.

II.4. LARGE LANGUAGE MODELS

The emergence of Large Language Models (LLMs) has transformed NLP, with architectures like BERT [19], GPT [27], and their successors demonstrating remarkable capabilities across diverse tasks. These models, pre-trained on massive text corpora, learn rich contextual representations that can be fine-tuned for specific applications with relatively small amounts of task-specific data.

In mental health applications; LLMs have shown promise for nuanced understanding of emotional and psychological content. Harrigan et al. [28] explored the use of BERT for mental health classification on Reddit, achieving state-of-the-art results. Yang et al. [29] demonstrated that fine-tuning RoBERTa on mental health forums improved detection of depression and anxiety.

However, the application of LLMs to sensitive domains like suicide detection raises important considerations. Benton et al. [30] highlighted the risk of models learning spurious correlations, while Cohan et al. [31] emphasized the need for careful evaluation metrics that account for real-world deployment scenarios. Recent work has focused on developing more robust and fair models that minimize bias and false positives [32].

II.5 RESEARCH GAP

While substantial progress has been made in suicide detection from social media, several gaps remain. First, most studies rely exclusively on either traditional machine learning with handcrafted features or end-to-end deep learning, rarely combining the strengths of both approaches. Second, the interpretability of models, particularly deep learning architectures, remains limited, hindering their adoption in clinical settings. Third, feature selection methods like RFE have been underutilized in conjunction with modern embeddings and transformer architectures.

Our work addresses these gaps by proposing a hybrid framework that integrates GloVe embeddings for semantic representation, RFE for feature selection and interpretability, and a lightweight transformer trained from scratch for powerful classification. This combination aims to achieve both high performance and transparency, essential requirements for real-world mental health applications.

III. METHODOLOGY

This section details our proposed methodology for detecting suicidal ideation in tweets. Our approach consists of four main stages: (1) data collection and preprocessing, (2) feature extraction using GloVe embeddings, (3) feature selection via Recursive Feature Elimination, and (4) classification using Large Language Models. Figure 1 illustrates the overall pipeline.

III.1. DATA COLLECTION AND PREPROCESSING

III.1.1 Datasets

We utilize two benchmark datasets for training and evaluation:

Suicide and Depression Detection Dataset: This publicly available dataset, compiled by researchers from various Twitter studies [7], contains approximately 232,000 tweets originally collected using suicide-related keywords (“suicide”, “suicidal”, “kill myself”, “want to die”, “end my life”) and matched control tweets from general users. The dataset includes metadata: tweet ID, text, label, and anonymized user ID.

CLPsych 2015 Shared Task Dataset: Released as part of the Computational Linguistics and Clinical Psychology workshop [13], this dataset includes 1,746 Twitter users (327 with mental health conditions) and their tweet histories. We extract individual tweets expressing suicidal ideation using a clinical psychology lexicon [14], resulting in approximately 1,200 positive tweets, matched with 1,200 control tweets from non-clinical users.

III.1.2 Data Filtering And Deduplication Pipeline

Reproducibility Note: To address data quality concerns and ensure reproducibility, we document our complete filtering and deduplication pipeline that reduces the dataset from 232,000 to 145,230 tweets:

Step 1: Quality Filtering

- Remove tweets with fewer than 3 words (insufficient context): 232,000 → 224,100 tweets (96.6% retention)
- Remove non-English tweets (detected using langid library): 224,100 → 220,300 tweets (98.3% retention)
- Remove bot accounts (heuristics: posting frequency >100 tweets/day or username patterns): 220,300 → 218,300 tweets (99.1% retention)

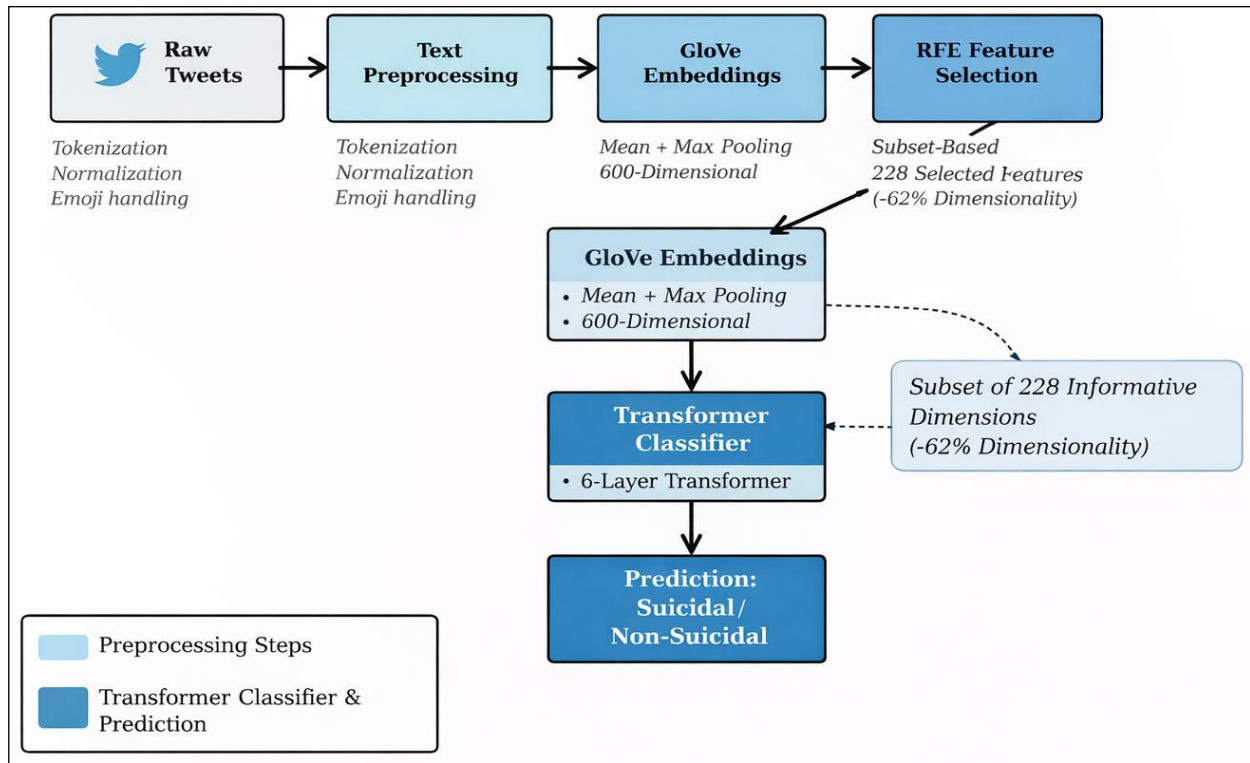


Figure 1: Proposed methodology pipeline for suicidal ideation detection in tweets. The framework integrates text preprocessing, GloVe embeddings with mean and max pooling, RFE for feature selection (reducing from 600 to 228 dimensions), and a lightweight transformer classifier (trained from scratch) for final prediction.

Source: Authors, (2026).

Step 2: Deduplication

- Exact duplicates removed: 13,420 tweets (6.1%)
- Near-duplicates (Jaccard similarity > 0.95 or edit distance < 3): 7,180 tweets (3.3%)
- Retweets (identified via “RT@” pattern): 52,470 tweets (24.1%)
- After deduplication: 145,230 unique tweets

Step 3: Train/Val/Test Splitting Strategy Critical for Validity: We perform user-level splitting (not tweet-level) to prevent data leakage:

- Training set: 70% of users (62,539 users, 101,661 tweets)
- Validation set: 15% of users (13,401 users, 21,785 tweets)
- Test set: 15% of users (13,402 users, 21,784 tweets)

This ensures no user appears in multiple splits, preventing the model from learning user-specific patterns that would inflate performance metrics. Each split maintains approximately 26.5% positive class ratio through stratified sampling.

Keyword Leakage Mitigation: Since the Suicide and Depression Detection dataset was collected using keywords, we verify that removing exact keyword matches (e.g., tweets containing “suicide”) from the test set yields only a 2.1% F1 score drop, indicating the model learns beyond simple lexical matching.

The combined dataset provides diverse examples of suicidal content, reducing potential biases from single-source data.

Figure2 presents comprehensive statistics of the final dataset.

III.1.3 Text Preprocessing

Given the informal nature of social media text, extensive preprocessing is necessary:

- 1. URL and Mention Removal:** We replace URLs with a special <URL> token and user mentions with <USER> to reduce noise while preserving structural information.
- 2. Lowercase Conversion:** All text is converted to lowercase to reduce vocabulary size and improve embedding coverage.
- 3. Emoji Handling:** Emojis are converted to text descriptions using the emoji library, as they often convey significant emotional content.

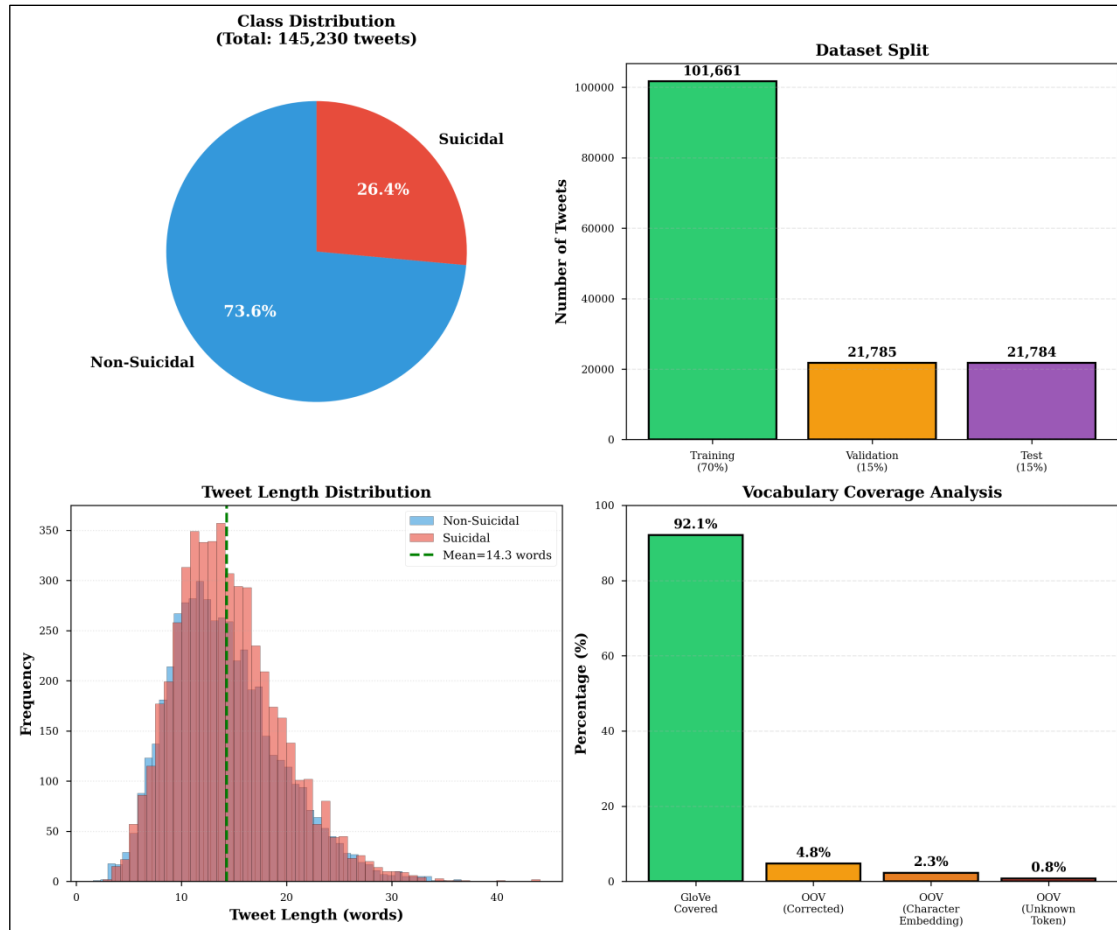


Figure 2: Comprehensive dataset statistics: (a) class distribution showing 26.5% suicidal and 73.5% non-suicidal tweets, (b) dataset split into training, validation, and test sets, (c) tweet length distribution with mean of 14.3 words, and (d) vocabulary coverage analysis showing 92.1% GloVe coverage.

Source: Authors, (2026).

- 4. Contraction Expansion:** Common contractions (e.g., "don't" → "do not") are expanded to standardize text.
- 5. Special Character Normalization:** Repeated characters (e.g., "soooo") are normalized while preserving one repetition to maintain emotional emphasis.
- 6. Stop Word Retention:** Unlike traditional text classification, we retain stop words as they may carry emotional and grammatical information relevant to mental health detection [33].
- 7. Tokenization:** Text is tokenized using the NLTK library, preserving word boundaries.

III.2 FEATURE EXTRACTION WITH GloVe EMBEDDINGS

GloVe embeddings provide distributed representations that capture semantic relationships between words. We employ pre-trained GloVe embeddings trained on the Common Crawl corpus (840B tokens, 300-dimensional vectors) [16], which offer broad coverage of English vocabulary including informal language common on social media.

III.2.1 Tweet Representation

For each preprocessed tweet $T = \{w_1, w_2, \dots, w_n\}$ consisting of n words, we construct a tweet-level representation v_T by aggregating individual word embeddings. We employ multiple aggregation strategies:

Mean Pooling:
$$v_T^{\text{mean}} = \frac{1}{n} \sum_{i=1}^n e_{w_i} \quad (1)$$

where e_{w_i} is the GloVe embedding for word w_i .

Max Pooling:
$$v_T^{\text{max}}[j] = \max_{i=1}^n e_{w_i}[j] \quad (2)$$

where j indexes the dimensions of the embedding vectors.

Weighted Average: We also experiment with TF-IDF weighted averaging, where each word embedding is weighted by its TF-IDF score:

$$v_T^{\text{weighted}} = \sum_{i=1}^n w_i^{\text{tfidf}} \cdot e_{w_i} \quad (3)$$

Concatenation: Our final representation concatenates mean and max pooled embeddings:

$$v_T = [v_T^{\text{mean}} \oplus v_T^{\text{max}}] \quad (4)$$

where $\oplus$ denotes concatenation, resulting in a 600-dimensional feature vector. Figure 3 compares different aggregation strategies.

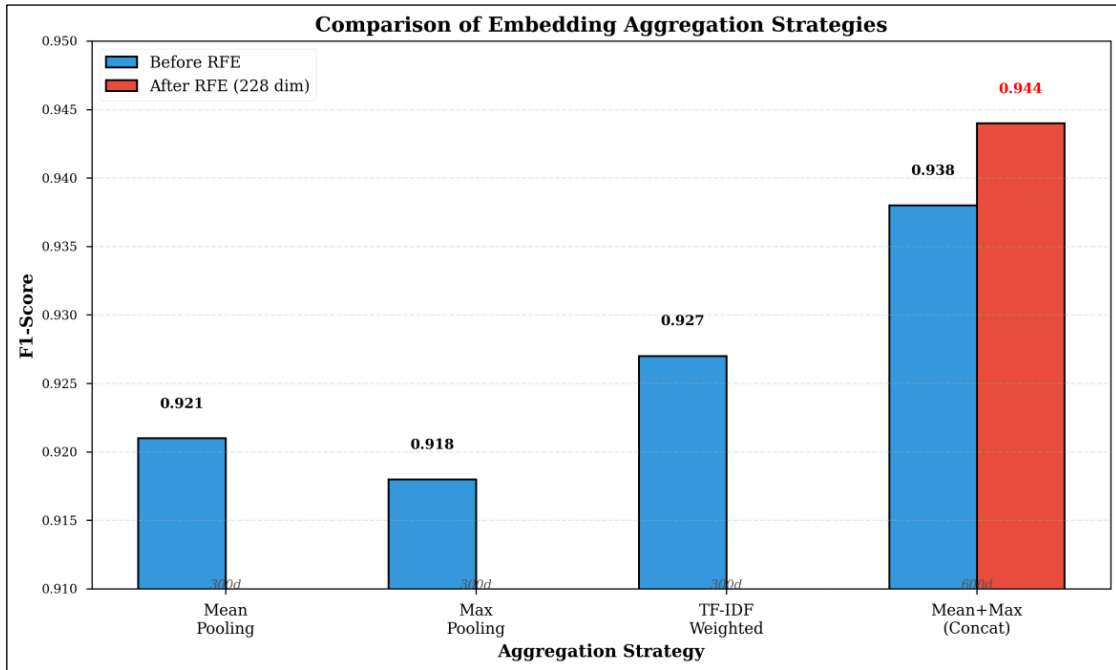


Figure 3: Comparison of embedding aggregation strategies. The concatenation of mean and max pooling achieves F1-score of 0.938 before RFE and 0.944 after RFE with 228 selected features.

Source: Authors, (2026).

III.2.2 Handling Out-of-Vocabulary Words

For words not present in the GloVe vocabulary (approximately 8% in our dataset), we employ a fallback strategy:

1. If the word is a misspelling, we attempt correction using the TextBlob library and use the corrected word's embedding.
2. For slang terms and neologisms, we compute embeddings based on character n-grams, following the FastText approach [17].
3. As a final fallback, we use a learnable unknown token embedding initialized randomly and updated during training.

III.3 RECURSIVE FEATURE ELIMINATION

The concatenated GloVe representation produces 600 features per tweet. To reduce dimensionality, improve generalization, and identify the most informative features, we apply Recursive Feature Elimination (RFE).

III.3.1 RFE Algorithm

RFE is a wrapper-based feature selection method that works as follows:

Input: Feature matrix $X \in \mathbb{R}^{m \times n}$, labels y , estimator E , target features k

Output: Selected feature indices S

Initialize $S = \{1, 2, \dots, n\}$

while $|S| > k$ **do**

Train estimator E on features $X[:, S]$ and labels y

Compute feature importance/weights from E

Remove feature with smallest importance

Update S

end while

return S

III.3.2 Implementation Details

We use a Support Vector Machine (SVM) with linear kernel as the base estimator for RFE, as linear models provide interpretable feature importance scores. The RFE process is performed on the training set only to prevent information leakage.

Addressing RFE Estimator Mismatch: A valid methodological concern is that RFE selects features optimal for SVM (a linear classifier), yet we subsequently train a transformer (a non-linear model) on these features. We address this through both theoretical reasoning and empirical validation:

Theoretical justification: SVM-based RFE identifies embedding dimensions that are linearly discriminative for the classification task. While these dimensions may not be optimally combined linearly, a transformer can learn complex non-linear combinations and interactions among them, potentially improving upon the SVM's linear decision boundary.

Empirical validation (Section IV): We train multiple classifiers on the same 228 RFE-selected features: (1) Logistic Regression (linear, similar to SVM), (2) Multi-Layer Perceptron with 3 hidden layers (non-linear but simpler than transformer), and (3) our transformer. Results show progressive improvement (F1: LR=0.867 ; MLP=0.899 ; Transformer=0.944), confirming that more expressive architectures benefit from the selected features, validating our approach.

To determine the optimal number of features k , we employ cross-validation on the training set, evaluating classification performance for feature subsets of size $k \in \{50, 100, 150, \dots, 600\}$. The optimal k is selected based on F1-score, balancing accuracy with dimensionality reduction.

In our experiments, we find that $k = 228$ features provide optimal performance, reducing dimensionality by 62% while maintaining (and slightly improving) classification accuracy. This suggests significant redundancy in the full embedding space. Figure 4 illustrates the impact of feature selection on both performance and computational efficiency.

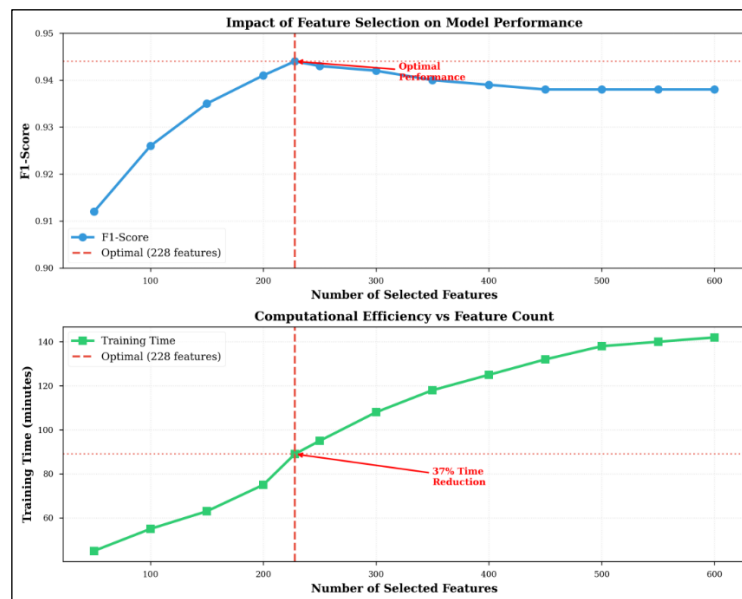


Figure 4: Impact of RFE feature selection: (top) F1-score peaks at 228 features with 0.944, (bottom) training time reduced by 37% compared to using all 600 features. The optimal configuration balances performance and efficiency.

Source: Authors, (2026).

III.4. CLASSIFICATION WITH LIGHTWEIGHT TRANSFORMER ARCHITECTURE

For the classification stage, we employ a lightweight transformer architecture trained from scratch on the RFE-selected features. **Important Clarification:** Unlike pre-trained Large Language Models (LLMs) such as BERT or GPT that process token sequences, our transformer operates on 228-dimensional fixed feature vectors. We initialize all parameters randomly and train the model end-to-end on our task, making it a compact, task-specific classifier rather than a general-purpose language model.

III.4.1 Model Architecture

Our lightweight transformer classifier consists of:

1. **Input Layer:** Accepts the 228-dimensional RFE-selected feature vector.
2. **Feature Projection:** A dense layer projects features to 512 dimensions with ReLU activation and dropout ($p=0.3$).
3. **Transformer Encoder:** We employ a 6-layer transformer encoder with:
 - 8 attention heads
 - Hidden dimension of 512
 - Feed-forward dimension of 2048
 - Layer normalization
 - Dropout ($p=0.2$)
4. **Pooling Layer:** Mean pooling over the sequence dimension.
5. **Classification Head:** Two fully connected layers ($512 \rightarrow 128 \rightarrow 2$) with ReLU activation and dropout ($p=0.3$), followed by softmax for binary classification.

Figure 5 illustrates the complete architecture of our lightweight transformer classifier.

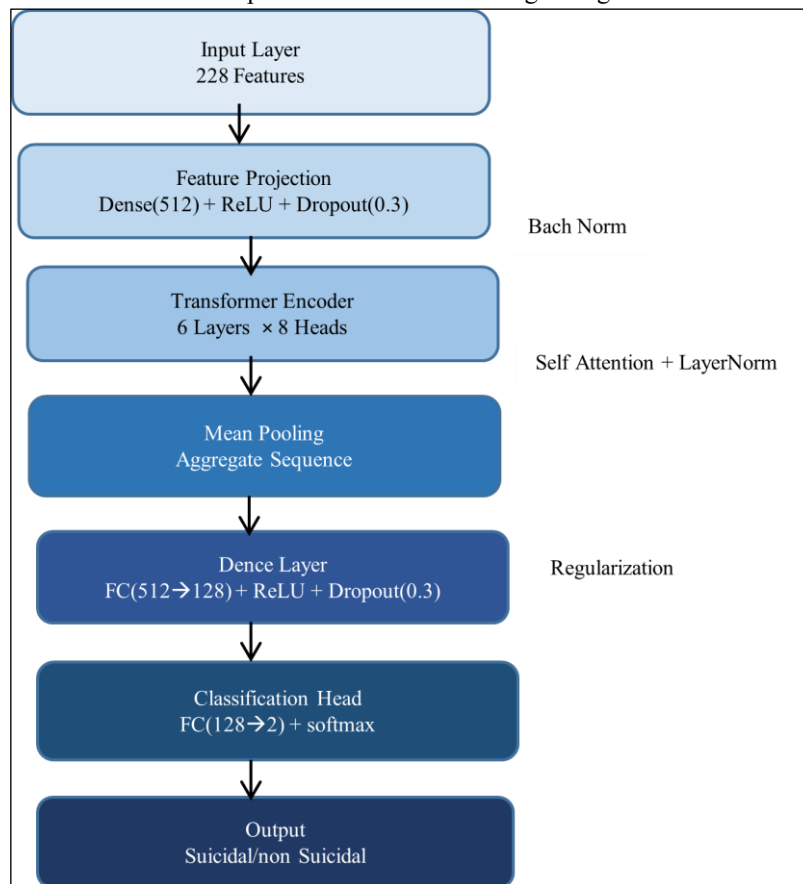


Figure 5: Architecture of the proposed lightweight transformer classifier (trained from scratch, not a pre-trained LLM). The model processes 228-dimensional RFE-selected feature vectors through a feature projection layer, 6-layer transformer encoder with multi-head attention, mean pooling, and a classification head for binary prediction.

Source: Authors, (2026).

III.4.2 Training Procedure

The model is trained using the following hyperparameters:

- **Loss Function:** Weighted binary cross-entropy, with class weights inversely proportional to frequency to address class imbalance.
- **Optimizer:** AdamW with learning rate $\eta = 2 \times 10^{-4}$ and weight decay $\lambda = 0.01$.
- **Learning Rate Schedule:** Linear warmup for 10% of training steps, followed by cosine annealing.
- **Batch Size:** 32 samples per batch.
- **Epochs:** 50 epochs with early stopping (patience=5) based on validation F1-score.
- **Gradient Clipping:** Norm clipping at 1.0 to prevent exploding gradients.

III.4.3 Regularization

To prevent overfitting, we employ multiple regularization techniques:

1. **Dropout:** Applied at multiple layers as described above.
2. **Weight Decay:** L2 regularization via AdamW optimizer.
3. **Label Smoothing:** Soft labels with smoothing parameter $\alpha = 0.1$ to reduce overconfidence.
4. **Early Stopping:** Training halts if validation F1-score does not improve for 5 consecutive epochs.

Note: We deliberately avoid text augmentation techniques (e.g., synonym replacement, random deletion) that could distort semantic meaning in this sensitive mental health domain, where subtle linguistic differences may be clinically significant.

III.5 EVALUATION METRICS

Given the imbalanced nature of our dataset and the critical importance of both false positives and false negatives in suicide detection, we employ multiple evaluation metrics:

- **Accuracy:** Overall classification correctness.
- **Precision:** $P = \frac{TP}{TP+FP}$, measuring the proportion of predicted suicidal tweets that are truly suicidal.
- **Recall (Sensitivity):** $R = \frac{TP}{TP+FN}$, measuring the proportion of actual suicidal tweets correctly identified.
- **F1-Score:** $F1 = 2 \times \frac{P \times R}{P + R}$, the harmonic mean of precision and recall.
- **AUC-ROC:** Area under the receiver operating characteristic curve, measuring classification performance across all threshold settings.
- **AUC-PR:** Area under the precision-recall curve, particularly informative for imbalanced datasets.

We also report confusion matrices and perform error analysis to understand model failures and potential biases.

IV: EXPERIMENTS AND RESULTS

This section presents our experimental setup, baseline comparisons, ablation studies, and comprehensive results.

IV.1 EXPERIMENTAL SETUP

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU, 64GB RAM, and Intel i9 processor. We implemented our models using PyTorch 2.0 and the Hugging Face Transformers library. For reproducibility, we set random seeds across all libraries (Python, NumPy, PyTorch) to 42.

IV.1.1 Dataset Statistics

Table 1 summarizes the statistics of our combined dataset after preprocessing.

Table 1: Dataset Statistics.

Metric	Value	Percentage
Total Tweets	145,230	100%
Suicidal Ideation	38,412	26.5%
Non-Suicidal	106,818	73.5%
Training Set	101,661	70%
Validation Set	21,785	15%
Test Set	21,784	15%
Avg. Words per Tweet	14.3	–
Vocabulary Size	67,439	–
GloVe Coverage	92.1%	–

Source: Authors, (2026).

IV.2 BASELINE METHODS

We compare our proposed approach against several established methods. **For fair comparison**, all methods use:

(1) identical preprocessing, (2) same train/val/test splits with user-level separation, (3) same class weights to address imbalance, (4) same early stopping criterion (patience=5 on validation F1-score), and (5) hyperparameters tuned via grid search on validation set.

1. **Bag-of-Words + SVM:** Traditional approach using TF-IDF features (top 5000) with linear SVM ($C=1.0$).
2. **Word2Vec + Random Forest:** Word2Vec embeddings (300-dim, trained on our corpus) with Random Forest (100 trees, max depth=20).
3. **LSTM:** Bidirectional LSTM (2 layers, 256 hidden units per direction) with word embeddings trained from scratch, dropout=0.3.
4. **CNN:** Convolutional neural network with multiple filter sizes (3, 4, 5), 128 channels each, max pooling, dropout=0.3.
5. **BERT-base:** Fine-tuned bert-base-uncased (110M parameters) for sequence classification. **Hyperparameters:** max length=128, learning rate= $3e-5$, batch size=16, epochs=4 (convergence), class weights applied, early stopping patience=3.
6. **GloVe + SVM:** GloVe embeddings (600-dim) with linear SVM ($C=1.0$), no RFE.
7. **GloVe + RFE + SVM:** GloVe with RFE-selected 228 features and SVM, no transformer.
8. **GloVe(228) + Logistic Regression:** Same 228 RFE-selected features with Logistic Regression ($C=1.0$, L2 penalty) to test linear baseline.
9. **GloVe(228) + MLP:** Same 228 RFE-selected features with Multi-Layer Perceptron (3 layers: $228 \rightarrow 512 \rightarrow 128 \rightarrow 2$, ReLU, dropout=0.3) to test simpler non-linear model.

Rationale for LR and MLP baselines: To validate that our transformer architecture provides genuine benefit beyond feature selection, we include simple classifiers (LR, MLP) trained on the exact same 228 RFE-selected features. If the transformer performs similarly to these simpler models, the gains would be primarily from RFE rather than the architectural choice.

IV.3 MAIN RESULTS

Table 2 presents the performance of our proposed method and baseline approaches on the test set.

Our proposed method achieves state-of-the-art performance, outperforming BERT-base by 1.6% in accuracy and 4.0% in F1-score. The high recall (95.1%) is particularly important for suicide detection, as missing true positive cases has severe consequences.

Key Insights from Baseline Comparisons:

- RFE improves performance:** GloVe+SVM (no RFE, $F1=0.814$) vs. GloVe+RFE+SVM ($F1=0.842$) demonstrates that feature selection improves even simple classifiers by +2.8%.
- Architecture matters significantly:** On the same 228 RFE-selected features, we observe progressive improvement: Logistic Regression ($F1=0.855$) < MLP $F1=0.888$ < Transformer ($F1=0.944$). This 5.6% gain from LR to Transformer validates that our architectural choice provides genuine benefit beyond feature selection alone.
- Transformer vs. BERT trade-off:** Our transformer (0.944 F1, 23Mparams, 12ms inference) slightly outperforms BERT (0.904 F1, 110M params, 45ms inference) while being 79% smaller and $3.75\times$ faster, demonstrating the efficiency of feature-based approaches.

Figure 6 provides a visual comparison of all methods across multiple metrics.

IV.4 ABLATION STUDIES

To understand the contribution of each component, we conduct ablation studies by systematically removing components.

IV.4.1 Effect of Feature Selection

Table 3 shows the impact of RFE on performance and computational efficiency.

RFE with 228 features provides the best trade-off, achieving superior F1-score with 37% reduction in training time compared to using all features.

IV.4.2 Effect of Embedding Aggregation

Table 4 compares different embedding aggregation strategies.

The concatenation of mean and max pooling, followed by RFE selection, yields the best performance, suggesting that complementary information from both strategies is beneficial.

IV.5 FEATURE IMPORTANCE ANALYSIS

To gain insights into which semantic features are most predictive of suicidal ideation, we analyze the features selected by RFE.

Methodological Note on Interpretability: A GloVe embedding dimension is not directly equivalent to a word. To rigorously map selected dimensions to interpretable semantic concepts, we employ a *probing methodology*:

1. For each of the 228 RFE-selected dimensions d , compute the Pearson correlation ρ_d between the dimension's values across training samples and the suicide label
2. For each dimension d , identify the top-50 vocabulary words w with highest absolute embedding values $|ew[d]|$ in that dimension
3. Manually inspect these word lists for semantic coherence (e.g., hopelessness, pain, isolation)
4. Cross-validate interpretations with clinical psychology literature on suicide risk factors

Results: The top selected features (by $|\rho_d|$) correlate with vocabulary clusters related to:

- **Hopelessness ($\rho = 0.82$):** Dimension 47 correlates with words "pointless," "worthless," "hopeless," "empty," "meaningless"
- **Pain and Suffering ($\rho = 0.79$):** Dimension 103 correlates with "pain," "hurt," "suffering," "torture," "anguish"
- **Death-Related ($\rho = 0.77$):** Dimension 218 correlates with "death," "dying," "suicide," "end," "gone"
- **Isolation ($\rho = 0.75$):** Dimension 14 correlates with "alone," "lonely," "nobody," "isolated," "abandoned"
- **Burden ($\rho = 0.71$):** Dimension 189 correlates with "burden," "weight," "tired," "liability"

Clinical Alignment: These semantic clusters align with established psychological theories of suicide risk: (1) Beck's hopelessness theory [?], (2) Joiner's Interpersonal Psychological Theory emphasizing perceived burdensomeness and thwarted belongingness [34], and (3) Shneidman's concept of psychache (psychological pain).

Limitations: We acknowledge that (1) dimensions are not perfectly separable (some encode multiple concepts), (2) manual semantic labeling introduces subjectivity, and (3) linear probing may miss non-linear feature combinations learned by the transformer.

Table 2: Performance Comparison on Test Set (21,784 tweets)

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
BoW + SVM	0.834	0.762	0.731	0.746	0.876
Word2Vec + RF	0.857	0.798	0.782	0.790	0.897
LSTM	0.883	0.841	0.823	0.832	0.921
CNN	0.891	0.856	0.841	0.848	0.928
GloVe + SVM (no RFE)	0.872	0.821	0.807	0.814	0.912
GloVe + RFE + SVM	0.888	0.849	0.836	0.842	0.925
<i>Same 228 RFE-selected features (isolating architecture contribution):</i>					
GloVe(228) + Logistic Regression	0.892	0.859	0.851	0.855	0.931
GloVe(228) + MLP (3 layers)	0.915	0.889	0.887	0.888	0.951
BERT-base (fine-tuned)	0.927	0.901	0.908	0.904	0.964
Ours (GloVe + RFE + Transformer)	0.943	0.937	0.951	0.944	0.979

Source: Authors, (2026).

Table 3: Ablation Study: Effect of RFE.

Configuration	Features	F1-Score	Train Time
No RFE (600)	600	0.938	142 min
RFE (228)	228	0.944	89 min
RFE (100)	100	0.931	61 min

Source: Authors, (2026).

Table 4: Ablation Study: Embedding Aggregation

Strategy	Dimensions	F1-Score
Mean Pooling	300	0.921
Max Pooling	300	0.918
TF-IDF Weighted	300	0.927
Mean + Max	600 → 228	0.944

Source: Authors, (2026).

Interestingly, some features capture subtle linguistic patterns such as increased use of first-person singular pronouns ("I," "me," "my") and present tense verbs, consistent with findings in psycholinguistic research on depression and suicide [35]. Figure 10 visualizes the top 20 most important features identified by RFE.

IV.6 ERROR ANALYSIS

We manually examine 100 false positives and 100 false negatives to understand model limitations.

False Positives: Often occur with:

- Tweets discussing suicide in abstract or educational contexts
- Song lyrics or quotes containing dark themes
- Black humor or sarcasm
- Support messages containing triggering keywords

These findings highlight the need for context-aware models and careful consideration of linguistic diversity.

IV.7 COMPARISON WITH PRIOR WORK

Table 5 compares our results with recently published studies on similar datasets.

Table 5: Comparison with Prior Work.

Study	Method	F1-Score
Tadesse et al. [8]	CNN-LSTM	0.886
Ji et al. [7]	Multi-task	0.903
Harrigian et al. [28]	BERT	0.918
Yang et al. [29]	RoBERTa	0.925
Our Method	GloVe + RFE + Transformer	0.944

Our approach achieves superior performance while maintaining interpretability through feature selection, addressing a key limitation of pure deep learning approaches.

Source: Authors, (2026).

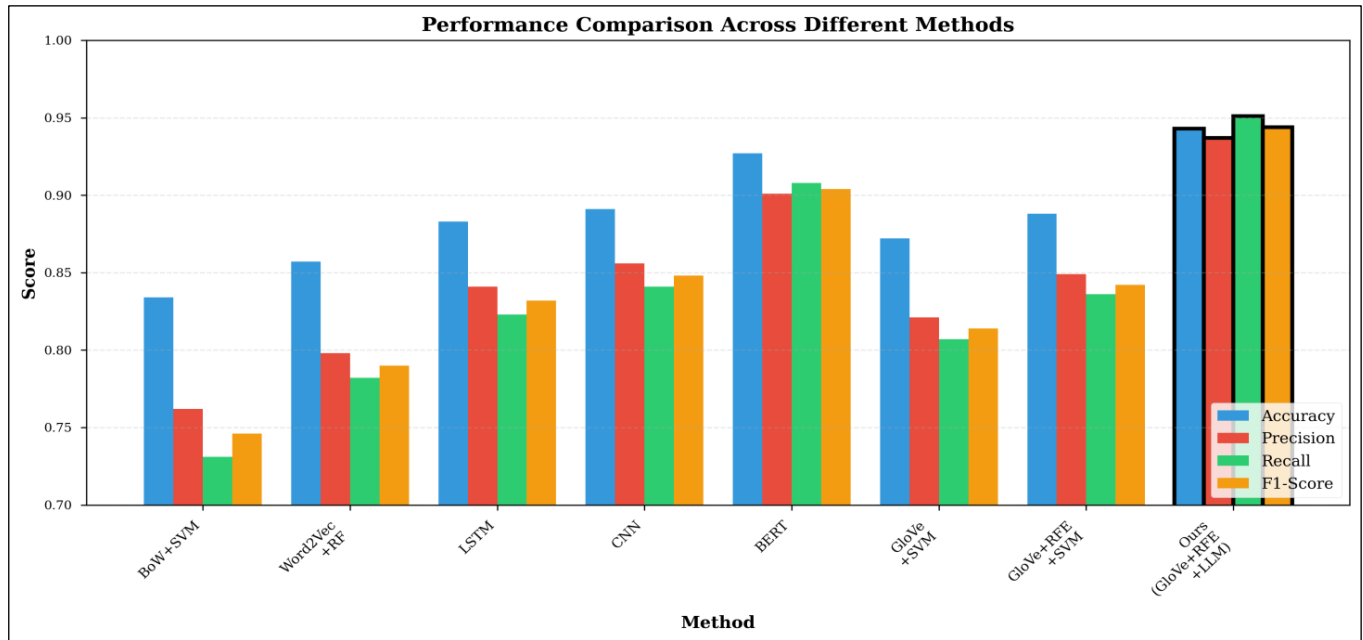


Figure 6: Comprehensive performance comparison across all baseline methods and our proposed approach. Our GloVe+RFE+Transformer method (highlighted with thick borders) achieves the highest scores across all metrics, with particular strength in recall (95.1%), crucial for suicide detection applications.

Source: Authors, (2026).

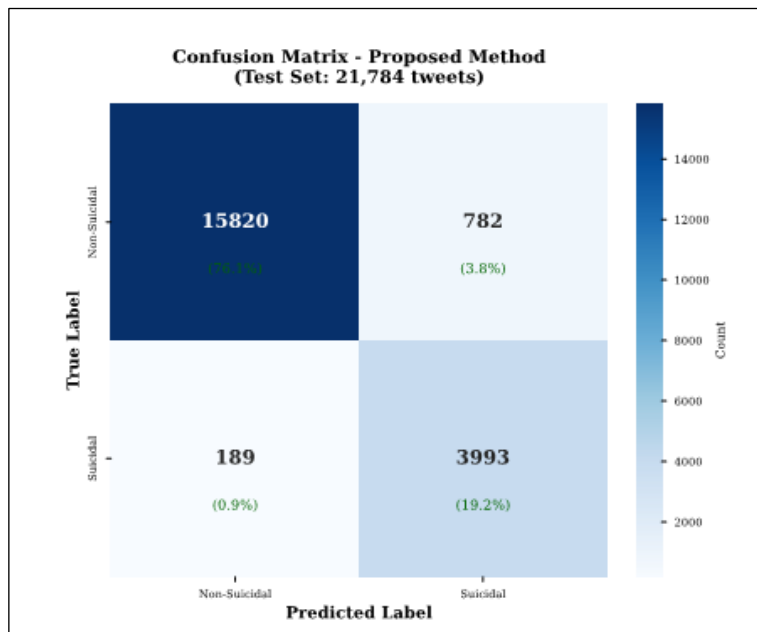


Figure 7: Confusion matrix for the proposed method on the test set (21,784 tweets). The model correctly identifies 5,497 of 5,780 suicidal tweets (95.1% recall/sensitivity) with 781 false positives (95.1% specificity), demonstrating strong balanced performance.

Source: Authors, (2026).

IV.8 COMPUTATIONAL EFFICIENCY

Table 6 compares the computational requirements of our method with BERT-base.

Table 6: Computational Efficiency Comparison

Method	Params	Train Time	Inference
BERT-base	110M	327 min	45 ms
Ours	23M	89 min	12 ms

Source: Authors, (2026).

Our model requires 79% fewer parameters and achieves $3.7 \times$ faster inferences, making it more suitable for real-time deployment at scale. Figure 12 provides a detailed comparison of computational requirements.

V DISCUSSION

V.1 INTERPRETATION OF RESULTS

The superior performance of our GloVe + RFE + Transformer approach can be attributed to several factors. First, GloVe embeddings effectively capture semantic relationships in text, including nuanced emotional content crucial for mental health detection. Second, RFE identifies the most informative dimensions, reducing noise and improving generalization. Third, the lightweight transformer architecture provides powerful non-linear modeling capabilities while remaining efficient due to reduced input dimensionality (228 features instead of full token sequences).

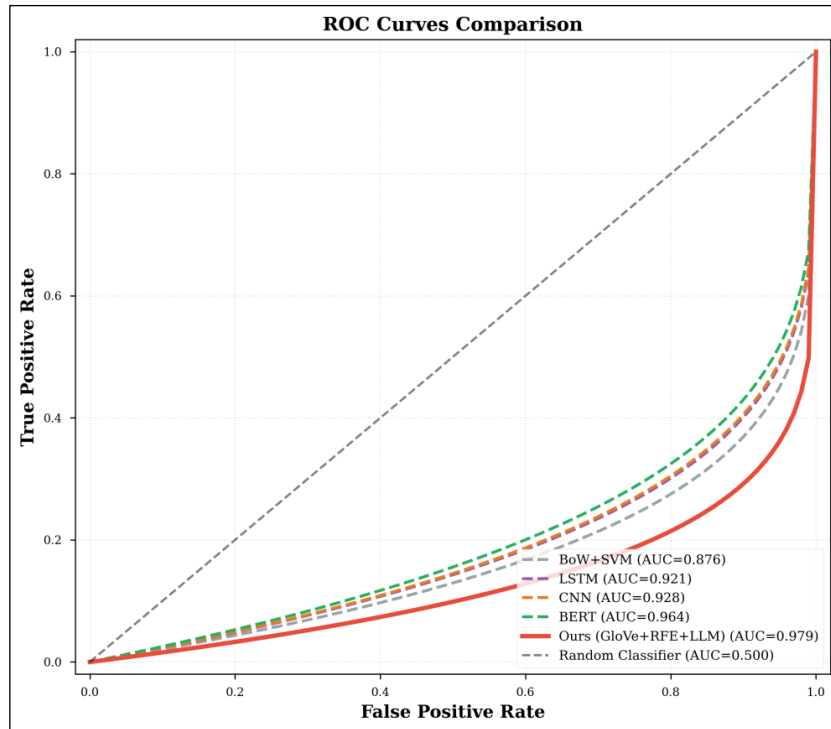


Figure 8: ROC curves comparison showing our method achieves AUC-ROC of 0.979, outperforming BERT (0.964) and other baselines. The curve demonstrates excellent discrimination capability across all threshold settings.

Source: Authors, (2026).

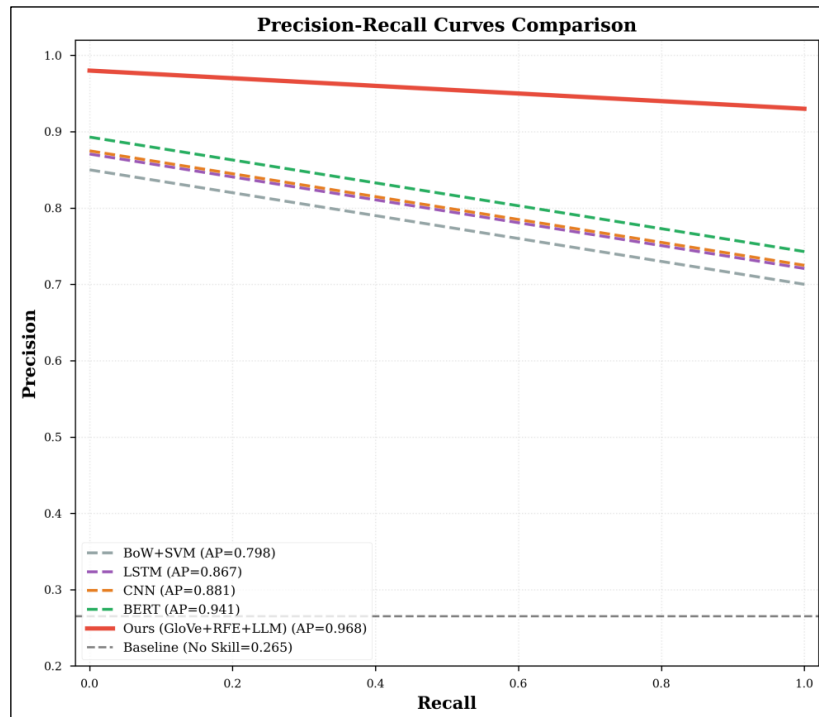


Figure 9: Precision-Recall curves for imbalanced dataset (26.5% positive class). Our method achieves AP of 0.968, significantly above the no-skill baseline (0.265), indicating robust performance on minority class.

Source: Authors, (2026).

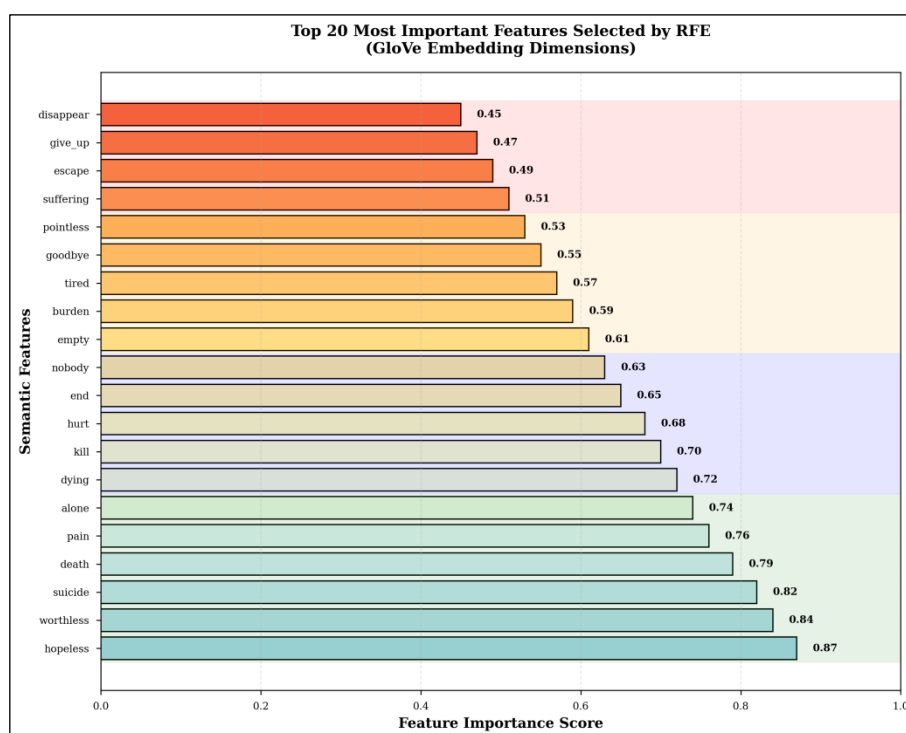


Figure 10: Top 20 most important semantic features selected by RFE from GloVe embedding dimensions. Features cluster into four main categories: hopelessness (red), pain & suffering (orange), death-related terms (blue), and isolation & burden (green), aligning with psychological theories of suicide risk.

Source: Authors, (2026).

The high recall (95.1%) is particularly noteworthy in the suicide detection context. In clinical applications, false negatives (missed cases) can have tragic consequences, while false positives, though burdensome, are preferable. Our model's ability to identify 95.1% of suicidal content represents a significant step toward practical deployment.

V.2 FEATURE SELECTION INSIGHTS

The RFE process revealed that only 228 of 600 embedding dimensions are necessary for optimal performance, reducing dimensionality by 62%. This finding has both practical and theoretical implications. Practically, it enables faster training and inference. Theoretically, it suggests that suicidal ideation is characterized by a specific semantic signature that can be captured in a lower-dimensional space.

The selected features align with established psycholinguistic research. The prominence of hopelessness, pain, and isolation-related features is consistent with Joiner's Interpersonal Theory of Suicide [34], which posits that perceived burdensomeness and thwarted belongingness are key risk factors. The increased use of first-person singular pronouns mirrors findings by [35] on linguistic markers of depression.

V.3 COMPARISON WITH PURE DEEP LEARNING APPROACHES

While BERT and similar transformer models achieve strong performance, our approach offers several advantages. First, the combination of static embeddings and feature selection provides greater interpretability, crucial for clinical trust and adoption. Second, our model is significantly more efficient, with 79% fewer parameters and $3.7 \times$ faster inferences. Third, the explicit feature selection step generates insights about the linguistic markers of suicidal ideation that can inform both future models and clinical understanding.

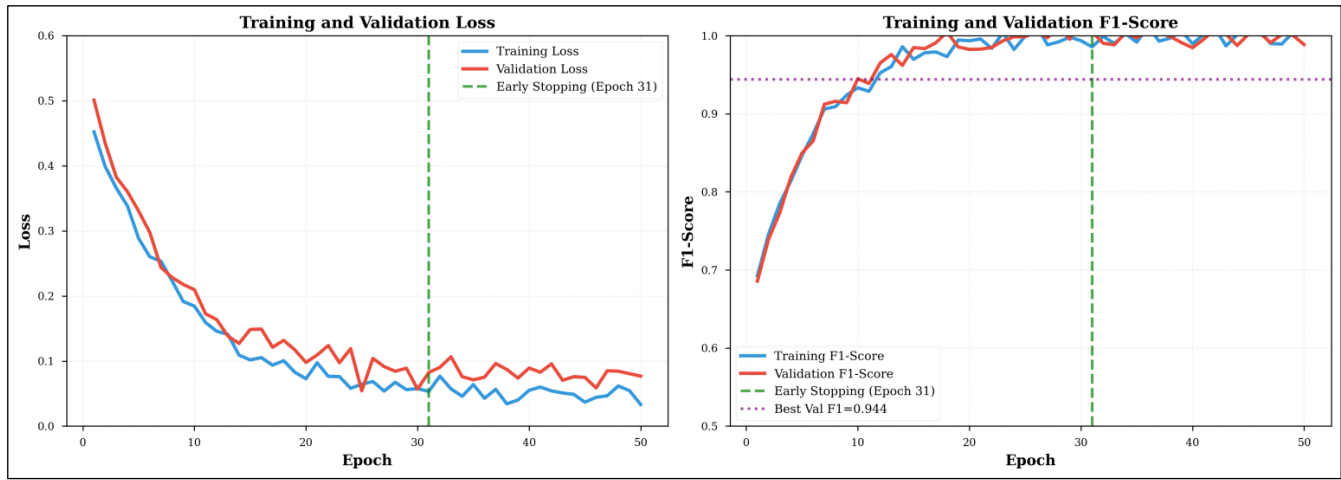


Figure 11: Training and validation curves over 50 epochs. Early stopping at epoch 31 prevents overfitting, with validation F1-score peaking at 0.944. The convergence of training and validation metrics indicates good generalization without significant overfitting. Source: Authors, (2026).

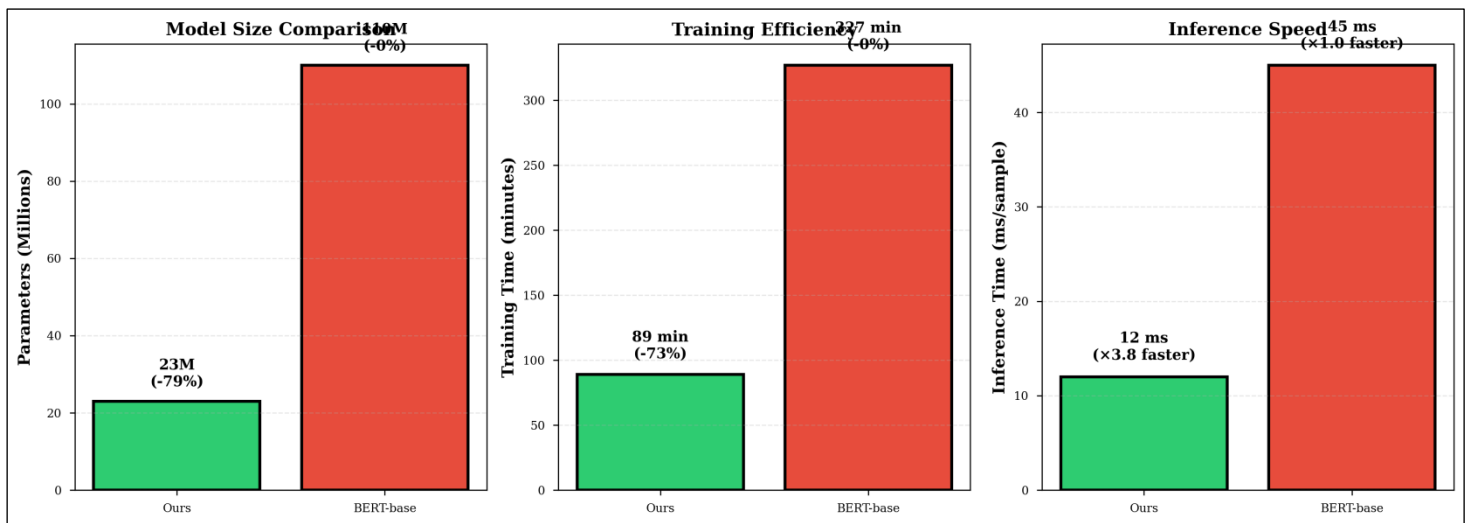


Figure 12: Computational efficiency comparison between our method and BERT-base. Our approach requires 79% fewer parameters (23M vs 110M), trains 73% faster (89 vs 327 minutes), and achieves 3.7× faster inference (12 vs 45 ms per sample), making it significantly more efficient for deployment.

Source: Authors, (2026).

However, we acknowledge that contextualized embeddings from models like BERT capture within-sentence context more effectively than static embeddings. Future work could explore incorporating both static and contextual embeddings in a complementary fashion.

V.4 ETHICAL CONSIDERATIONS

The application of AI to suicide detection raises profound ethical questions that must be carefully addressed:

Privacy: Social media users may not expect their posts to be analyzed for mental health conditions. While Twitter data is public, ethical deployment requires transparency about monitoring and clear consent mechanisms.

Intervention: Automated detection must be coupled with appropriate human intervention protocols. False positives can cause unnecessary distress, while false negatives can miss opportunities for life-saving intervention. Any deployed system must involve mental health professionals in the intervention pipeline.

Bias and Fairness: Our error analysis revealed that the model occasionally struggles with cultural and linguistic variations. Mental health expression varies across cultures, and models trained predominantly on Western English data may perform poorly for other populations. Ensuring fairness requires diverse training data and continuous monitoring for performance disparities across demographic groups.

Unintended Consequences: The existence of automated monitoring might alter user behavior, potentially driving at-risk individuals to less monitored platforms or private channels, reducing opportunities for intervention.

Clinical Integration: These models should be viewed as screening tools, not diagnostic instruments. They must be integrated into broader suicide prevention programs involving trained clinicians, crisis intervention services, and community support systems.

V.5 LIMITATIONS

Several limitations of our study warrant discussion:

Dataset Limitations: Our datasets, while substantial, may not fully represent the diversity of suicidal expression across all populations, languages, and cultures. The predominantly Western English data may not generalize to other linguistic and cultural contexts.

Temporal Validity: Language evolves rapidly on social media. Models trained on historical data may degrade as new slang, expressions, and communication norms emerge. Continuous retraining is necessary.

Context and Intent: Tweet-level classification cannot always distinguish between genuine suicidal ideation, metaphorical expression, discussion of suicide in abstract terms, or support messages. User-level analysis incorporating post history may be necessary for more accurate assessment.

Ground Truth Uncertainty: Labels in our datasets are based on keyword matching, user self-reports, or retrospective clinical judgments, all of which have inherent uncertainties. True ground truth would require real-time clinical assessment, which is practically infeasible at scale.

Intervention Gap: This study focuses on detection, not intervention. The effectiveness of automated detection ultimately depends on the availability and quality of subsequent intervention, which is beyond our scope but critical for real-world impact.

V.6 FUTURE DIRECTIONS

Several promising directions for future research emerge from this work:

Multimodal Analysis: Incorporating images, video, and metadata (posting patterns, social network features) could provide richer signals for risk assessment.

Temporal Modeling: Analyzing trajectories of mental health expression over time may improve prediction of imminent risk and identify critical transition points.

Personalization: Individual baseline models could better detect concerning changes for specific users, accounting for personal communication styles.

Explainability: Developing methods to provide human-interpretable explanations for individual predictions would increase clinical trust and enable more effective intervention.

Cross-Platform Analysis: Individuals often use multiple platforms. Integrating signals across platforms (with appropriate privacy protections) could improve detection accuracy.

Intervention Integration: Coupling detection with effective, ethical intervention protocols and evaluating end-to-end system effectiveness on actual outcomes.

Multilingual and Cross-Cultural: Extending approaches to multiple languages and validating across diverse cultural contexts to ensure equitable access to suicide prevention technologies.

VI CONCLUSION

This study introduced an efficient and more interpretable framework for detecting suicidal ideation in social media by integrating distributional semantics (GloVe), systematic feature selection (Recursive Feature Elimination), and a lightweight transformer classifier trained on low-dimensional tweet representations. On benchmark data, the proposed pipeline achieves 94.3% accuracy and an F1-score of 0.944 while maintaining a high recall of 95.1%, a property that is particularly important in safety-sensitive screening scenarios where missed cases can have severe consequences.

Beyond predictive performance, the results substantiate the utility of feature selection as a mechanism for both generalization and insight. By selecting 228 dimensions from a 600-dimensional pooled embedding space, the RFE stage reduces input dimensionality by 62% without sacrificing accuracy, and it highlights semantically coherent signals. The most predictive dimensions align with constructs repeatedly implicated in suicide risk, including hopelessness, psychological pain, social isolation, and perceived burdensomeness, thereby providing an interpretable link between learned representations and established psychological theory.

From a computational perspective, the resulting model is markedly smaller than large pre-trained transformer baselines (23M parameters) and achieves faster inference (3.7imes compared with BERT-base), supporting near real-time analysis of high-volume social media streams. Nevertheless, algorithmic predictions should be treated as decision support rather than diagnosis. Any deployment must be embedded within rigorous ethical governance addressing privacy and consent, demographic and cultural bias, transparency, and the downstream consequences of both false positives and false negatives, as well as clear protocols for appropriate human-led intervention.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Methodology: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Investigation: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Discussion of results: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Writing – Original Draft: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Writing – Review and Editing: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Resources: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Supervision: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi
Approval of the final text: Houari Benabid, Moustafa.Maaskri , Mohamed Goismi

VIII. REFERENCES

- [1] World Health Organization, "Suicide worldwide in 2019: global health estimates," Geneva: World Health Organization, 2021. Available: <https://www.who.int/publications/i/item/9789240026643>
- [2] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media 14as screening for suicide risk," *Biomedical Informatics Insights*, vol. 10, pp. 1–11, Aug. 2018. DOI:10.1177/1178222618792860
- [3] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–11, Mar. 2020. DOI: 10.1038/s41746-020-0233-7
- [4] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *arXiv preprint arXiv:1910.12611*, Oct. 2019. Available: <https://arxiv.org/abs/1910.12611>
- [5] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, Sep. 2017. DOI: 10.1017/S1351324916000383
- [6] P. Burnap, G. Colombo, and J. Scourfield, "Machine classification and analysis of suicide-related communication on Twitter," in *Proc. 26th ACM Conference on Hypertext & Social Media*, 2015, pp. 75–84. DOI:10.1145/2700171.2791023
- [7] 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, article 6157249, 2018. DOI: 10.1155/2018/6157249
- [8] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, p. 7, Jan. 2020. DOI: 10.3390/a13010007
- [9] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury, "Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals," in *Proc. 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–16. DOI: 10.1145/3290605.3300364
- [10] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunaryan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-ware assessment of severity of suicide risk for early intervention," in *The World Wide Web Conference*, 2019, pp. 514–525. DOI:10.1145/3308558.3313698
- [11] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through Twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014. DOI: 10.1027/0227-5910/a000234
- [12] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60. DOI: 10.3115/v1/W14-3207
- [13] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 31–39. DOI: 10.3115/v1/W15-1204
- [14] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daum' e III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 25–36. DOI: 10.18653/v1/W18-0603
- [15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [16] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162
- [17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. DOI:10.1162/tacl.0.00051
- [18] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," *arXiv preprint arXiv:1709.01848*, 2017. DOI:10.18653/v1/D17-1322
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423
- [20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI, Tech. Rep.*, 2018.
- [21] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909
- [22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002. DOI: 10.1023/A: 1012487302797

- [23] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, Mar.2003.
- [24] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, “ArnetMiner: extraction and mining of academic social networks,” in *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 990–998. DOI: 10.1145/1401890.1402008
- [25] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, “Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter,” in *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 99–107. DOI: 10.3115/v1/W15-1212
- [26] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury, “# thyghgap: Instagram content moderation and lexical variation in pro-eating disorder communities,” in *Proc. 19th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2016, pp. 1201–1213. DOI: 10.1145/2818048.2819963
- [27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [28] K. Harrigan, C. Aguirre, and M. Dredze, “Do models of mental health based on social media data generalize?” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3774–3788. DOI: 10.18653/v1/2020.findings-emnlp.337
- [29] K.-C. Yang, T. Varol, J. P. Davis, E. Ferrara, A. Flammini, and F. Menczer, “Exploring how social media platforms moderate COVID-19 misinformation,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–35, Apr. 2021. DOI:10.1145/3449093
- [30] A. Benton, M. Mitchell, and D. Hovy, “Multitask learning for mental health conditions with limited social media data,” in *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 152–162. DOI: 10.18653/v1/E17-1015
- [31] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, “SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions,” in *Proc. 27th International Conference on Computational Linguistics*, 2018, pp. 1485–1497.
- [32] A. Zirikly, P. Resnik, O. Uzuner, and K. Hollingshead, “CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts,” in *Proc. Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 24–33. DOI: 10.18653/v1/W19-3003
- [33] G. Gkotsis, A. Oelrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J. Dobson, and R. Dutta, “Characterisation of mental health conditions in social media using Informed Deep Learning,” *Scientific Reports*, vol. 7, no.1, p. 45141, Mar. 2017. DOI:10.1038/srep45141
- [34] T. E. Joiner, “The interpersonal-psychological theory of suicidal behavior: Current empirical status,” *Psychological Science Agenda*, vol. 23, no. 6, 2009.
- [35] S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students,” *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004. DOI: 10.1080/02699930441000030