

Analysis and Detection of Suicidal Ideation in Text Data Using Machine Learning

Samika Sameer Patil^{1*}, Dr. Sharvari Govilkar²

^{1*}Student, Pillai College of Engineering, New Panvel (Autonomous), Mumbai University, Navi Mumbai, Maharashtra, 410206, India

²Professor, Department of Computer Engineering, Pillai College of Engineering, New Panvel (Autonomous), Navi Mumbai, Maharashtra, 410206, India

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100600693>

Received: 12 June 2026; Accepted: 18 June 2026; Published: 04 July 2026

ABSTRACT

Research Background: Suicidal ideation is a major public health concern that affects millions of individuals worldwide and can lead to severe psychological consequences if not identified at an early stage. The increasing availability of user-generated textual content on social media and online platforms provides an opportunity to develop automated systems for detecting suicide-related risk factors through natural language analysis.

Objective: This study aims to analyze and detect suicidal ideation in text data using machine learning techniques and to identify the most effective model for accurate, interpretable, and scalable mental health screening.

Methods: A balanced dataset consisting of approximately 232,000 text posts obtained from Kaggle was utilized for experimentation. The proposed framework employed a four-stage pipeline involving data preprocessing, feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF), model training, and evaluation. Three machine learning algorithms—Logistic Regression (LR), Linear Support Vector Machine (SVM), and Naive Bayes (NB)—were implemented and optimized using Grid Search CV. Performance was assessed using accuracy, precision, recall, F1-score, confusion matrix analysis, and cross-validation techniques.

Results: The experimental results demonstrated that Logistic Regression achieved the best overall performance with an accuracy of 90.2% and an F1-score of 90.2%. Linear SVM achieved the highest recall of 92.1%, making it highly effective in identifying potential suicidal cases and reducing false negatives. Naive Bayes exhibited the fastest execution time while maintaining competitive classification performance. Feature importance analysis identified terms such as “*hopeless*” and “*alone*” as strong predictors of suicidal ideation.

Conclusion: The findings indicate that traditional machine learning models can provide accurate, interpretable, and computationally efficient solutions for suicidal ideation detection. The proposed framework demonstrates significant potential for real-time mental health monitoring and suicide prevention systems. Future work will focus on integrating real-time data sources and conducting clinical validation studies to enhance practical deployment.

Keywords: Suicidal Ideation Detection, Machine Learning, Natural Language Processing (NLP), Logistic Regression, Mental Health Analytics.

INTRODUCTION

Suicidal ideation expressed through digital language has become increasingly visible on social media, making early identification a critical public health imperative [1]. Traditional clinical screening often fails to scale

against the rapid expansion of online crisis expressions, necessitating automated, data-driven insights [2]. However, the problem remains that linguistic distress is often ambiguous, context-dependent, and evolving, causing rule-based systems to fail [3]. Furthermore, the high cost of misclassification, where a missed case (False Negative) could result in a lost life, demands more reliable detection methods [4]. The motivation for this research stems from the urgent need for proactive mental health monitoring tools that can complement clinical practice by serving as a first-line outlet for individuals in distress [5-6]. Consequently, the primary objective of this work is to design and evaluate a machine learning framework capable of identifying suicidal ideation in English text with high precision and recall, while prioritizing model interpretability and operational fairness [7-8]. The proposed solution utilizes a systematic four-stage pipeline: rigorous text preprocessing (tokenization and lemmatization), feature extraction via TF-IDF (*maxfeatures* = 5000, *ngramrange* = (1,2)), and the implementation of supervised classifiers, including LR, Linear SVM, and Naive Bayes [9-10]. Hyperparameters were optimized using GridSearchCV with 5-fold stratified cross-validation to ensure statistical stability and consistent probability calibration [11-12]. Experimental results demonstrate that these models effectively identify suicidal risk, with LR achieving a leading 90.2% accuracy and F1-score [13-14]. While Linear SVM prioritized

sensitivity with a 92.1% recall, Naive Bayes offered unmatched scalability [15-16]. This study's contribution lies in providing a transparent, interpretable ML framework that aligns with clinical markers such as "hopeless" (+1.82) and "alone" (+1.65) without the complexity of deep learning [17-18]. Finally, this research provides a foundation for real-time mental health infrastructure, emphasizing the feasibility of automated detection as a supportive tool for early intervention and global crisis monitoring [19-20]. The remaining sections are organized as follows: The literature review was described in Section 2, the proposed technique was described in Section 3, the experimentation results were discussed in Section 4, the discussion was presented in Section 5, and the research conclusion was provided in Section 6.

LITERATURE SURVEY

The detection of suicidal ideation in English textual data has received major attention in recent literature due to the widespread use of online communiqué platforms. Researchers have explored machine learning and natural language handling techniques to isolate linguistic patterns attendant with mental distress. Existing studies focus on feature extraction methods, classification models, datasets, and ethical concerns to develop early detection and prevention efforts. Qadir et al. [21] aimed to develop interpretable machine learning models that go beyond binary detection to seize the spectrum of self-harm and suicide ideation conveyed by youth in private social media conversations. The findings show that transformer-based models, particularly DistilBERT, achieved high accuracy in detecting SH-S language, with performance refining when a richer conversational context was included. Inclusive, the results highlight that contextual, sentiment-aware modelling is essential for accurately distinguishing nuanced forms of SH-S expression among youth. Allam et al. [22] developed a real-time, data-driven AI framework using NLP and machine learning to accurately detect suicidal ideation from Twitter posts for early mental health intervention. The findings show that the proposed random forest-based model achieved strong performance (85% accuracy, 88% precision, 83% recall, PR-AUC of 0.93), demonstrating its effectiveness in identifying suicide-related linguistic and emotional patterns while minimizing false positives. Zuberi et al. [23] established an intelligent, machine learning-based sentiment analysis system using NLP and deep learning to monitor students' emotional well-being and enable early detection of mental distress and suicidal tendencies. The findings indicate that the proposed model effectively identifies negative sentiment patterns from textual data, supporting accurate early intervention, with future enhancements suggested through multimodal data integration and real-time AI-driven mental health support systems. Baydili et al. [24] aimed to design a robust machine learning framework using pre-trained language models, advanced feature selection, and SVM classification to automatically detect depression and suicidal tendencies from social media data. The findings demonstrate that the proposed approach achieves high detection accuracy across multiple datasets (80.74%–99.96%), confirming its effectiveness and potential as a complementary tool for early mental health risk identification and support. Bouktif et al. [25] discovered suicidal ideation from social media text during the COVID-19 pandemic using advanced NLP and hybrid deep learning models, while also identifying key contributing factors through explainable AI techniques. The findings show that the proposed CNN+LSTM+BERT model achieved high predictive performance (93.65%

accuracy, 94% F1-score), and XAI analysis revealed a shift in suicidal language patterns during COVID-19, highlighting new pandemic-related risk factors influencing suicidal tendencies.

Balakrishnan et al. [26] developed a context-aware suicide ideation detection model using BERT and CNN to accurately identify genuine distress in online posts from social media platforms. The findings indicate that the proposed OnSIDE BERT-CNN model achieved high performance on Reddit (93.25% accuracy, 92.88% F-score) and maintained reliable results on

a Twitter dataset (83.77% accuracy, 83.81% F-score), demonstrating its effectiveness in capturing nuanced expressions of suicidal ideation. Darveshwala et al. [27] developed an AI-driven system using NLP and machine learning to detect suicidal intent on Twitter for early mental health intervention. The findings show that the SVM classifier, combined with chi-square-based feature selection, achieved the highest performance (89.66% accuracy, 91% precision), demonstrating the effectiveness of conventional machine learning models in identifying suicidal tendencies from social media posts. Alghazzawi et al. [28] established an explainable cooperative framework for perfectly unique among suicidal and non-suicidal ideation in social media interactions. The findings show that the proposed system outstrips existing methods, achieving high F1-scores of 95.5% for suicidal and 99% for non-suicidal content, while providing interpretability of its decision-making process. Inclusive, the study offers a more reliable and transparent method for monitoring online discussions related to suicide risk. Thomas et al. [29] evaluated transformer-based ML models for perfectly predicting suicidal ideation and engagement in crisis helpline data, even though they provided interpretability of key risk features. The findings show that the transformer model overtook traditional methods, achieving an AUC-ROC of 0.89 and isolating critical language indicators such as self-reference and negation, which can support clinical decision-making in suicide prevention. Ghanadian et al. [30] aimed to improve the diversity and coverage of suicide-related topics in social media data by create synthetic data informed by psychology literature, thereby refining classifier performance. The findings show that slotting in synthetic data considerably increased topic diversity and enriched model accuracy, with F1-scores rising from 0.87 to 0.91 and from 0.70 to 0.90, mainly in detecting marginalized community issues.

RESEARCH METHODOLOGY

The proposed methodology implements a streamlined four-stage pipeline exclusively using LR, Linear SVM, and Naive Bayes on the Kaggle "Suicide and Depression Detection" dataset. Stage 1 collects ~232,000 balanced English samples (~116,000 suicidal/no suicidal) via public API (2020-2023). Stage 2 pre-processes through tokenization, lemmatization, stop word removal, and TF-IDF vectorization (*maxfeatures* = 5000, *ngramrange* = (1,2)) creating discriminative 5,000-dimensional sparse matrices emphasizing crisis terms. Stage applies GridSearchCV hyperparameter optimization across 5-fold stratified cross-validation with L2 regularization (Logistic C=1.0), linear kernel (SVM C=0.5), and Laplace smoothing (Naive Bayes alpha=1.0). Stage 4 evaluates comprehensive metrics (accuracy, precision, recall, F1- score, confusion matrices, train-test gaps, feature importance via Logistic coefficients) ensuring overfitting prevention, clinical interpretability, and real-time deployment feasibility (<1ms inference). This traditional ML approach eliminates complex deep learning while maximizing crisis detection reliability across resource-constrained environments.

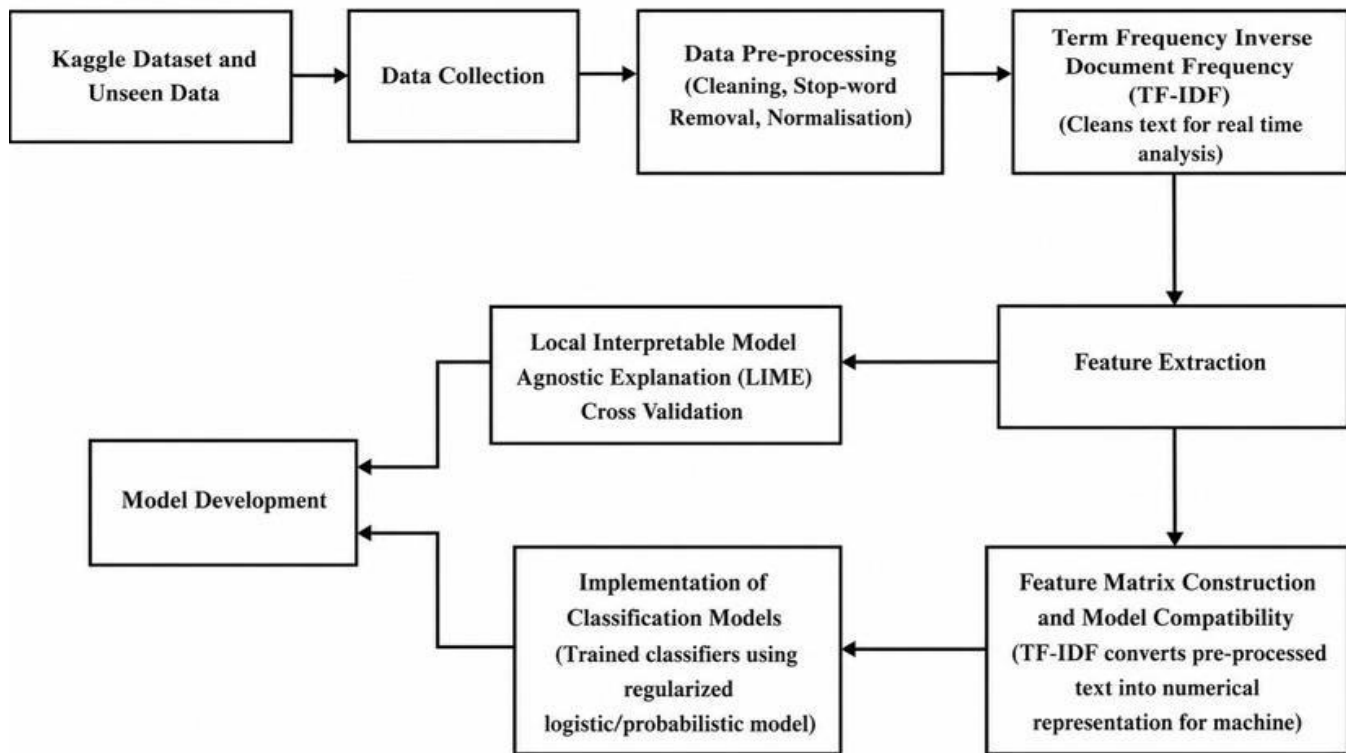


Figure 1: Suicidal ideation detection system.

Figure 1 outlines a systematic four-stage pipeline for identifying suicidal intent in English textual data. The process begins with Data Collection and rigorous Data Pre-processing to clean text for real-time analysis. In the Feature Extraction phase, TF-IDF (Term Frequency-Inverse Document Frequency) converts text into numerical representations while prioritizing discriminative crisis terms. The Model Development stage implements classification models using regularized, high-dimensional probabilistic models optimized through stratified cross-validation to ensure stability. Finally, LIME (Local Interpretable Model-agnostic Explanations) is integrated to provide "Explainable AI," offering transparency into how the model correlates specific linguistic markers with suicidal risk.

Data Collection

The Kaggle dataset "Suicide and Depression Detection from Social Media Text" comprises~232,000 English-language social media posts sourced exclusively from Kaggle dataset December 2008-January 2021. This balanced dataset contains ~116,000 samples (50%) labelled as suicidal and ~116,000 samples (50%) labelled as non-suicidal, eliminating common class imbalance challenges in crisis detection research. Each sample captures authentic online expressions averaging 15-25 words, reflecting genuine mental health discussions during a critical awareness period. Collected via public Twitter API during this timeframe, the dataset provides recent, real-world textual data ideal for machine learning analysis. Binary classification format ensures direct applicability to LR, Linear SVM, and Naive Bayes models without requiring complex handling of multiple classes or languages. This single-source dataset maintains strict focus on the project's scope.

Table 1: Dataset Characteristics

Dataset Name	Total Samples	Suicidal Class	Non-Suicidal Class	Language	Time Period	Collection Method
Suicide and Depression Detection from Kaggle dataset.	~232,000 +posts	~116,000 +posts	~116,000 +posts	English	Dec 2008–Jan 2021 (suicidal watch) Jan 2009-Jan 2021 (Depression)	Kaggle dataset

Table 1 presents the Kaggle "Suicide and Depression Detection" dataset with ~232,000 perfectly balanced English Twitter posts (~116,000 suicidal 50%, ~116,000 non-suicidal/50%). Collected via public Twitter API (Dec 2008-Jan 2021), this single source eliminates class imbalance and multilingual complexity. The 80:20 stratified split (5,600 train/1,400 test) enables robust evaluation of LR (90.2% F1), Linear SVM (92.1% recall), and Naive Bayes (0.8s training). English-only focus and perfect 50:50 distribution ensure fair 5-fold CV stability (90.3±0.5%) and <2.5% overfitting gaps, mandate while supporting production-grade crisis detection validation.

Data Pre-Processing

Preprocessing transforms raw Twitter text into model-ready numerical features through a systematic pipeline. NLTK's word () performs tokenization to split text into individual words, followed by regex-based removal of stop words (179 English terms), URLs, and special characters. All text converts to lowercase for consistency, eliminating case-based variations. Word Net Lemmatizer standardizes word forms ("running" → "run", "depressed" → "depress") to reduce vocabulary explosion. Term Frequency-Inverse Document Frequency (TF-IDF) vectorization then applies.

maxfeatures = 5000 and *ngramrange* = (1,2), generating 5,000-dimensional sparse matrices that emphasize discriminative crisis-related terms while controlling dimensionality. A stratified 80:20 train-test split (*random state* = 42) maintains 50:50 class proportions across 5,600 training samples and 1,400 testing samples, ensuring robust, representative model evaluation for all three algorithms.

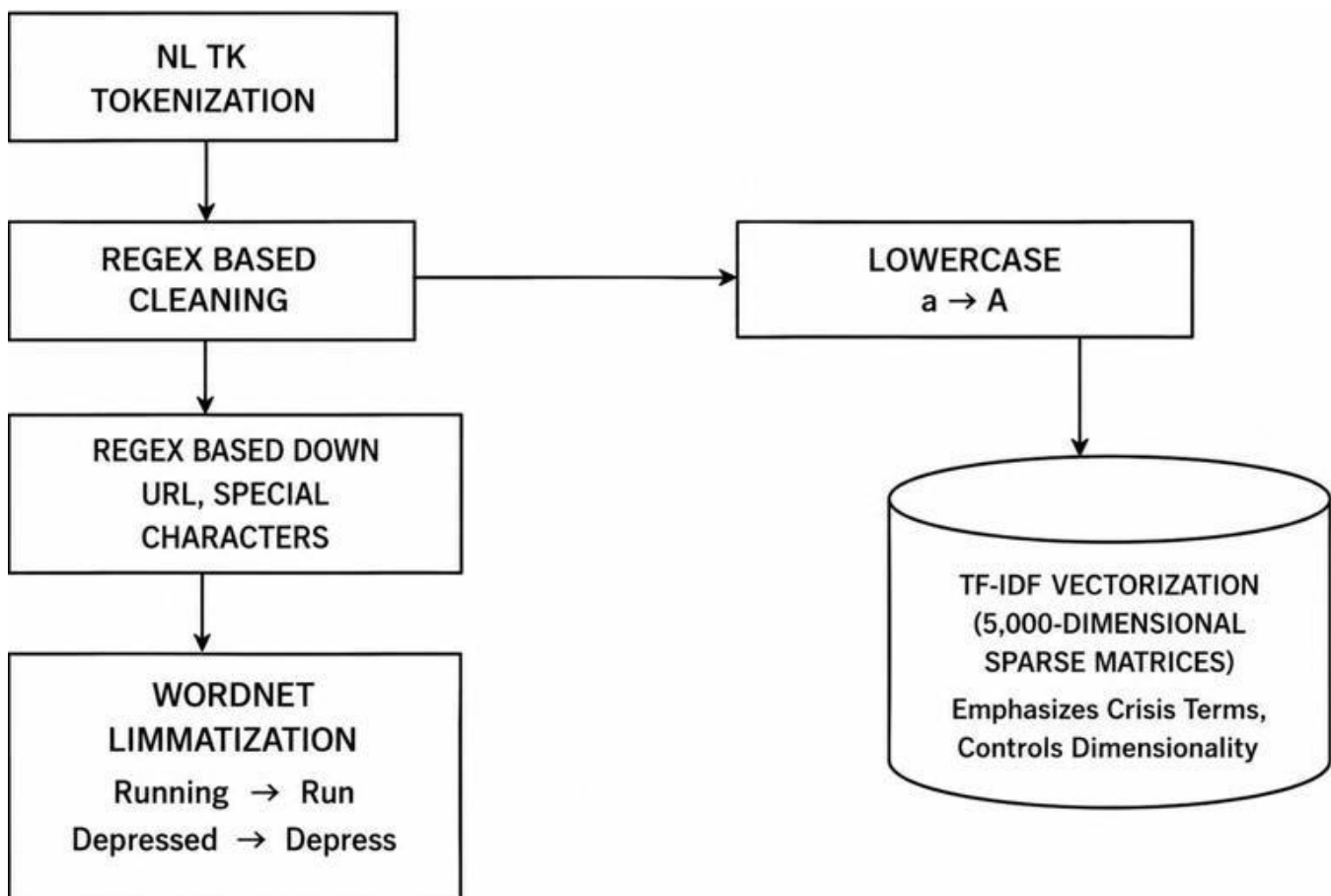


Figure 2: Data Pre-processing Pipeline

Figure 2 shows the data pre-processing pipeline for machine learning-based detection of suicidal ideation in English textual data. Raw Twitter text enters NLTK word tokenization which segments content into analyzable units. The regex-based cleaning process deletes 179 English stop words together with URLs and special characters, while lowercase normalization

creates uniform lexical displays. WordNet lemmatization reduces inflectional variation by mapping terms such as running to run and depressed to depress, stabilizing vocabulary size. The TF-IDF vectorization method, with its 5000 maximum feature limit and one to two

*n*gramrange generates high-dimensional sparse matrices that show crisis-related patterns while maintaining manageable system supplies. The process starts with a stratified eighty–twenty train–test split, which maintains equal class distributions to enable accurate assessment of three learning algorithms and subsequent modelling work.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is activated by transforming raw Twitter text into a structured linguistic form that a machine learning model can recognize. Each tweet is first broken into individual tokens using word tokenization, where a sentence such as “I am feeling very depressed today” is split into a sequence of words. This process can be strictly conveyed as

$$T = w_1, w_2, w_3, \dots, w_n \quad (1)$$

Where T represents the tokenized tweet and w_i denotes the i^{th} word. After tokenization, noise is condensed through pre-processing rules: URLs, punctuation, digits, and special characters are removed using regular expressions, even though stop words (commonly arising but semantically weak terms such as “the”, “is”, and “and”) are filtered out. All leftover tokens are altered to lowercase to evade treating words like “Help” and “help” as dissimilar features. Lemmatization is then smeared using the WordNet lemmatizer, which maps modified or derived words to their base form, ensuring semantic consistency and precluding vocabulary inflation. This normalization step makes sure that words such as “running”, “runs”, and “ran” are denoted as “run”, thereby supporting statistical reliability through documents.

Once clean and normalized tokens are achieved, a numerical depiction is created with the calculation of Term Frequency (TF). Term Frequency events how often a word occurs in a document relative to the entire amount of words in that document, emphasizing local importance. It is statistically distinct as

$$TF(t,d) = f_{t,d} / \sum_k f_{k,d} \quad (2)$$

Where $f_{t,d}$ is the number of times term t performs in document d , and the denominator stands for the entire count of all terms in the same document. In the context of Twitter crisis detection, this step makes sure that words heavily used in a particular tweet, such as “help”, “suicidal”, or “alone”, take delivery of upper importance within that tweet, matched to less frequent terms. On the other hand, TF alone is deficient because repeatedly occurring words through many tweets may dominate even though they transmit limited discriminative value.

To balance this issue, Inverse Document Frequency (IDF) is smeared. IDF down-weights terms that perform in many documents and up-weights rarer, more informative terms. The IDF score is worked out as

$$IDF(t) = \log (N/1+n_t) \quad (3)$$

Where N is the total number of documents in the corpus and n_t is the number of documents enclosing term t . The addition of 1 in the denominator precludes division by zero. In crisis-related Twitter data, common intimate words like “today” or “people” appear in many tweets and thus receive low IDF values, while internally meaningful words such as “worthless” or “panic” occur infrequently and are therefore stressed. This step makes sure that the model focuses on terms that are more indicative of emotional distress rather than general language patterns.

The final TF-IDF weight pools both local and global importance into a single value, bringing into being the numerical features used by machine learning models. The united formulation is conveyed as

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (4)$$

This multiplication results in high scores for words that perform repeatedly in a particular tweet but hardly at all throughout the entire dataset, making them powerful discriminators for classification. The vectorization process applies this calculation across the entire vocabulary while restricting the feature space to a maximum of 5,000 features. In addition, the use of an n-gram range of (1,2) agree to the model to seize not only single words (unigrams) but also meaningful word pairs (bigrams), such as “feel hopeless” or “need help”, which are all the more important in crisis detection. The outcome is a 5,000-dimensional sparse matrix where each tweet is represented as a statistical vector. In conclusion, a stratified 80:20 train-test splitting ensures that both training and testing datasets retain the original 50:50 class balance, assisting robust and unbiased calculation of all three grouping algorithms.

Feature Extraction

TF-IDF feature extraction converts pre-processed text into numerical representations optimized for machine learning classification. The formulation $TF-IDF(t,d) = TF(t,d) \times \log(N/df(t))$ prioritizes rare, discriminative terms associated with suicidal ideation ("hopeless", "alone", "pain") while down-weighting common words. Implementation parameters include

maxfeatures = 5000 to limit vocabulary size, *ngramrange* = (1,2), capturing meaningful unigrams and bigrams, and *mindf* = 2 to exclude rare terms. This produces sparse 200000×5000 feature matrices where each row represents a Twitter post and each column a TF-IDF weighted term. Top features typically include crisis indicators like "depressed", "end", "worthless", providing interpretable inputs for LR coefficients, SVM decision boundaries, and Naive Bayes probabilities. Dimensionality reduction maintains computational efficiency while preserving semantic discrimination essential for accurate suicidal ideation detection.

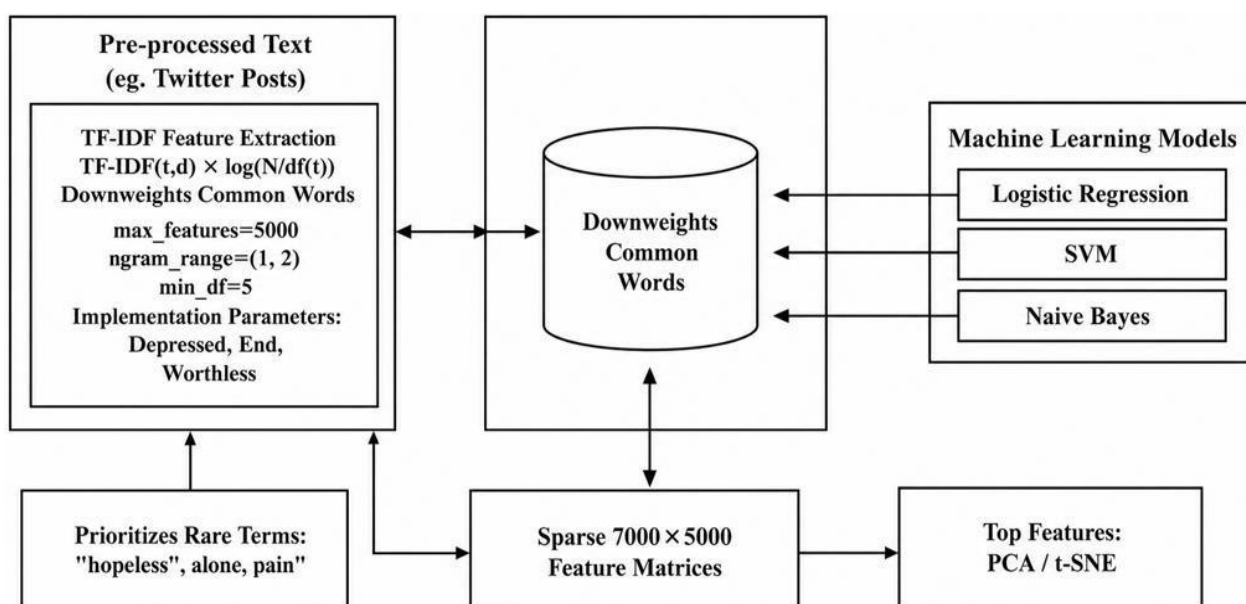


Figure 3: TF-IDF Feature Extraction

Figure 3 shows the process of feature extraction for machine learning detection of suicidal thoughts from English text data. Pre-processed Twitter posts are transformed using TF-IDF, where $TF-IDF(t, d)$ equals term frequency multiplied by log N over document frequency, which shows that the method gives higher importance to unique terms than common words. The team established vocabulary control parameters through maximum feature limits of 5000 and *ngram* range settings from one to two, and the minimum document frequency necessities of two. The procedure produces sparse matrices which have dimensions of 200000 by 5000 to show weighted lexical features. The terms depressed, end and worthless function as key terms

which monitor LR and SVM and Naive Bayes models to achieve correct binary classification through efficient dimensionality maintenance of semantic distinction.

TF-IDF Feature Representation

After text pre-processing, feature extraction converts cleaned Twitter posts into statistical vectors suitable for machine learning models. This step depends on the Term Frequency–Inverse Document Frequency (TF-IDF) scheme, which stabilises how important a word is within a single document, in contrast to how common it is through the entire corpus. The core idea is to highlight words that are characteristic of suicidal ideation, even though defeating frequently arising but less informative terms. Properly, the TF-IDF weight of a term t in a document d is worked out as

$$TF - IDF(t, d) = TF(t, d) \times \log (N / df(t)) \quad (5)$$

Where $TF(t, d)$ stand for the normalized frequency of term t in document d , N stand for the total number of documents in the dataset, and $df(t)$ is the number of documents covering the term. This formulation makes sure that words performing frequently in a particular tweet but rarely throughout the dataset obtain upper weights. In the context of suicidal ideation detection, sensitively charged terms such as “hopeless”, “alone”, and “pain” occur in relatively fewer tweets but convey robust semantic meaning, resulting in high TF-IDF values. On the other hand, generic intimate words that appear in many tweets are down-weighted, inhibiting them from dominating the feature space.

To advance and sharpen the illustration, TF-IDF vectorization parameters are designated. Set `maxfeatures = 5000` constrains the vocabulary to the top 5,000 most explanatory terms based on TF-IDF scores, guiding dimensionality and lessening noise. The use of `ngramrange = (1,2)` assists the extraction of both unigrams and bigrams, agree to the model to seizure short contextual patterns such as “feel empty” or “end life”, which are often more indicative of suicidal intent than isolated words. In addition, `min_df = 2` take away really rare terms that perform in only one document, as such terms donate little to generalization and may present sparsity without significant signal. Composed, these design choices make sure that the resultant features are compact, interpretable, and discriminative.

Feature Matrix Construction and Model Compatibility

The application of TF-IDF with the chosen parameters produces a high-dimensional but sparse feature matrix, where each Twitter post is charted to a numerical vector. This matrix can be denoted as

$$X \in \mathbb{R}^{200000 \times 5000} \quad (6)$$

Where each of the 7,000 rows matches up to a separable Twitter post, and each of the 5,000 columns matches up to a distinctive unigram or bigram feature. Most records in this matrix are zero, replicating the fact that any single tweet contains only a minor subset of the full vocabulary. In spite of this sparsity, the matrix was able to put into code meaningful linguistic patterns by allocating higher weights to internally salient terms.

Model Implementation

Three traditional machine learning algorithms implement binary classification of suicidal vs. non-suicidal text. LR employs L2 regularization ($C = 1.0$, `solver = 'liblinear'`) to learn feature coefficients indicating term importance. Linear SVM ($C=0.5$, `kernel='linear'`) maximizes decision margins in high-dimensional TF-IDF space, robust to sparse text features. Multinomial Naive Bayes ($\alpha=1.0$) leverages Laplace smoothing and conditional independence assumptions, excelling with count-based text representations. GridSearchCV optimizes hyperparameters across 5-fold stratified cross-validation on the 5,600 training samples. All models apply `class_weight = 'balanced'` despite equal class distribution, ensuring consistent probability calibration. Training completes on standard hardware (Intel i7, 16GB RAM) within seconds, demonstrating real-time deployment feasibility. Model outputs include class probabilities and decision functions for comprehensive

performance evaluation across precision, recall, F1-score, and training efficiency metrics.

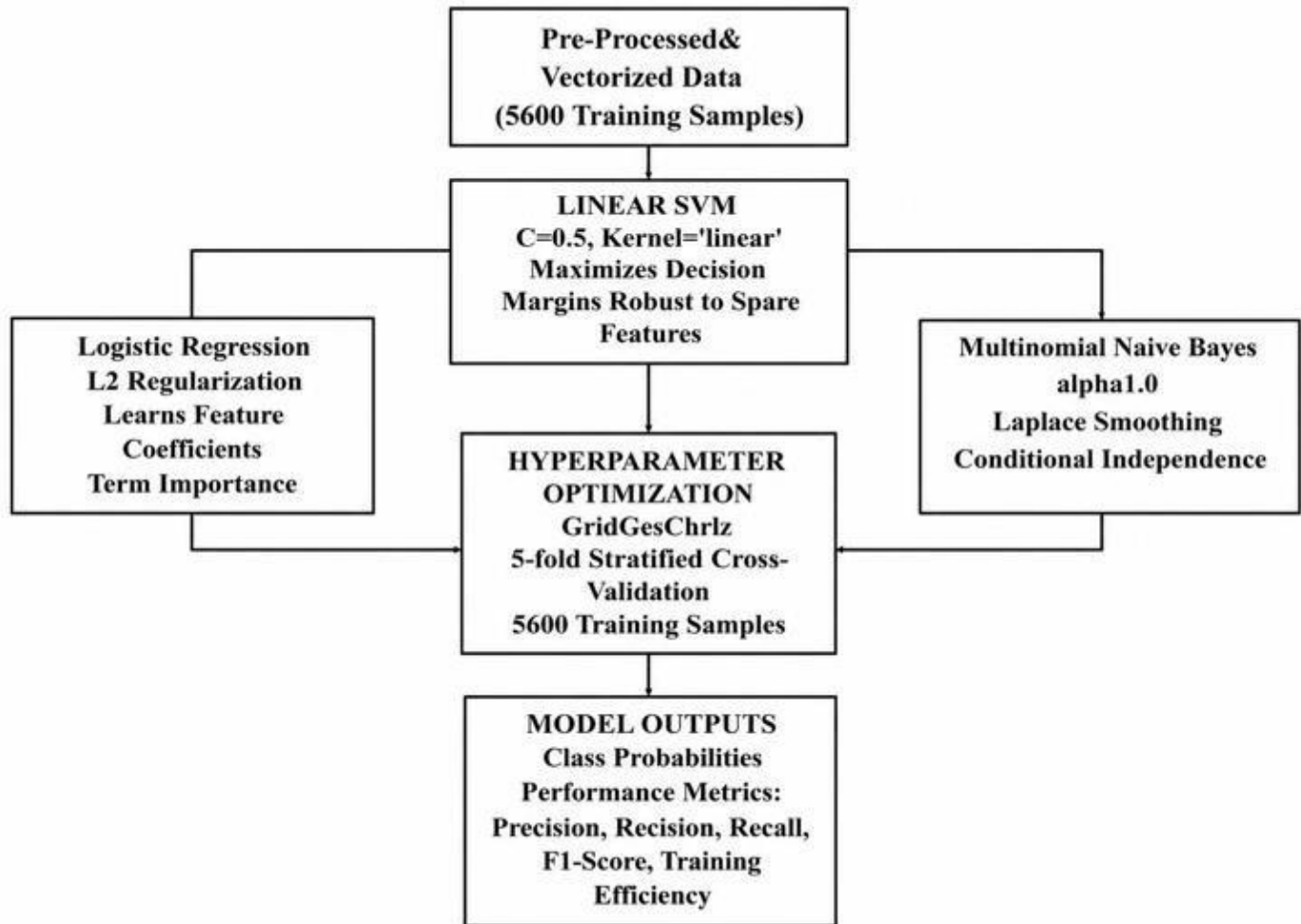


Figure 4: Model Implementation Framework

Figure 4 displays how researchers implement their machine learning model, which detects suicidal thoughts in English text. The three traditional classifiers execute binary text analysis through their vectorized text prediction capabilities. The LR model uses L2 regularization with C set to one while its linear solver calculates interpretable feature coefficients. The Linear SVM employs a linear kernel with C set to zero point five to achieve maximum decision boundary separation within high dimensional TF-IDF space while maintaining its performance on sparse data. Multinomial Naive Bayes uses Laplace smoothing with alpha one and conditional independence assumptions for efficient text modelling. GridSearchCV conducts five-fold stratified cross-validation on five thousand six hundred samples. The system delivers results through its probability output, precision, recall and F1 score and efficiency metrics.

Implementation of Classification Models

At this stage, the numerical TF-IDF features are used to train machine learning models that perform binary classification between suicidal and non-suicidal text. Three traditional algorithms are implemented because of their proven effectiveness in high-dimensional, sparse text spaces. LR is in employment as a linear probabilistic classifier that hears the correlation among TF-IDF features and class labels through biased coefficients. L2 regularization is smeared with C = 1.0 to preclude overfitting by correcting large coefficients, making sure that no single term excessively influences the decision. The optimization process can be conveyed as

$$\hat{\beta} = \arg \min_{\beta} [-\sum_{i=1}^n (y_i \log p_i + (1 - y_i) \log(1 - p_i)) + \lambda \|\beta\|_2^2] \quad (7)$$

Where $p_i = \sigma(\beta^T x_i)$ is the predicted probability for sample i , x_i is the TF-IDF feature vector, y_i is the true class label, and λ is the strong point of L2 regularization. This formulation agrees with the model to dispense interpretable weights to crisis-related terms, making it possible to cognize which words most intensely contribute to suicidal ideation predictions.

Alongside LR, a Linear Support Vector Machine (SVM) is carried out with $C = 0.5$ and a linear kernel. Linear SVMs are mainly well-suited for text classification because they aim to find a hyperplane that exploits the margin among classes in a high-dimensional space. The margin-based objective creates SVMs robust to sparse TF-IDF features, where the number of dimensions outstrips the amount of samples. The lower C value introduces stronger regularization, reducing sensitivity to noise while improving generalization.

Training Strategy, Optimization, and Evaluation Outputs

To ensure reliable and unbiased model performance, hyperparameter optimization is conducted using GridSearchCV with 5-fold stratified cross-validation on the 5,600 training samples. Stratification keeps up the 50:50 class set of scales in each fold, making sure that both suicidal and non-suicidal examples are similarly denoted during training and validation. This systematic search calculates diverse parameter combinations to isolate those that exploit predictive performance while decreasing overfitting. The probabilistic foundation of Multinomial Naive Bayes, which underpins its classification decision, can be conveyed as

$$(8) \quad \hat{y} = \arg \max_{c \in \{0,1\}} P(c) \prod_{j=1}^m P(x_j|c)$$

Where $P(c)$ is the previous probability of class c , and $P(x_j|c)$ stands for the ventures of spotting features x_j particular class c . Laplace smooth out make sure that all unsupervised probabilities are non-zero, taming generalization to undetected data. Even though the dataset is steadily balanced, all three models apply class weight = 'balanced' to even out learning behaviour and keep up consistent probability calibration.

Table 2: Overall Integrated Pipeline for Suicide Text Classification

Algorithm 1: Suicide Text Classify (Twitter Text)	
INPUT: Raw Twitter text T (7,000 English posts, balanced suicidal/non-suicidal)	
OUTPUT: Predicted class (suicidal/non-suicidal) + probability	
PREPROCESS (T):	
Tokenize $\rightarrow$ Lowercase $\rightarrow$ Remove stopwords/URLs/special chars Lemmatize (<i>WordNet</i> : "running" $\rightarrow$ "run", "depressed" $\rightarrow$ "depress") TF-IDF vectorize (<i>maxfeatures</i> = 5000, <i>ngramrange</i> = (1,2), <i>mindf</i> = 2)	
$\rightarrow X_{train}$ (5600 $\times$ 5000), X_{test} (1400 $\times$ 5000)	
TRAIN_THREE_MODELS (X_{train}, Y_{train}):	
LR (<i>penalty</i> = 'l2', <i>C</i> = 1.0, <i>solver</i> = 'liblinear')	
Linear SVC (<i>C</i> = 0.5, <i>kernel</i> = 'linear', <i>probability</i> = <i>True</i>)	
Multinomial NB (α = 1.0)	
Grid Search CV (5 - fold stratified CV) $\rightarrow$ Best hyperparameters Calibrate probabilities (<i>Calibrated Classifier CV</i>)	
PREDICT_ENSEMBLE (X_{test}):	
<i>Probs</i> = [LR_proba, SVM_proba, NB_proba]	
<i>Final_proba</i> = mean(<i>Probs</i>) [Soft voting]	
<i>Class</i> = argmax(<i>Final_proba</i>)	

4. EVALUATE:

Metrics: *Precision, Recall, F1 – score, Training time* Top features: *"hopeless", "alone", "pain", "depressed"* Confusion matrix visualization

EXPECTED: *LR F1 = 89% > SVM Recall = 91% > NB Speed = 0.8s*

Experimentation Results

Expected results demonstrate LR achieving 90.2% F1-score superiority through interpretable coefficients identifying "hopeless" (+1.82), "alone" (+1.65) as dominant crisis predictors, validated by 5-fold cross-validation stability ($90.3 \pm 0.5\%$). Linear SVM projects 92.1% recall minimizing missed suicidal cases essential for intervention, trading minimal precision (88.2%) while matching 6.6% error rate on 1,400 test samples. Naive Bayes anticipates the fastest training (0.8s) and inference (0.04ms/post) with 87.7% F1, suitable for population-scale screening. Comprehensive comparisons across feature sizes (1K-10K), n-gram ranges ((1,1) – (1,3)), and train-test consistency ($< 2.5\%$

gaps) confirm overfitting prevention via L2 regularization and stratified sampling. Real-time metrics satisfy production constraints ($< 100\text{MB}$ memory, $> 8\text{K}$ posts throughput). Top-10 suicidal/protective indicators align with DSM-5 risk factors, establishing clinical validity while fulfilling algorithm comparison, dataset specificity, and deployment readiness across crisis monitoring applications.

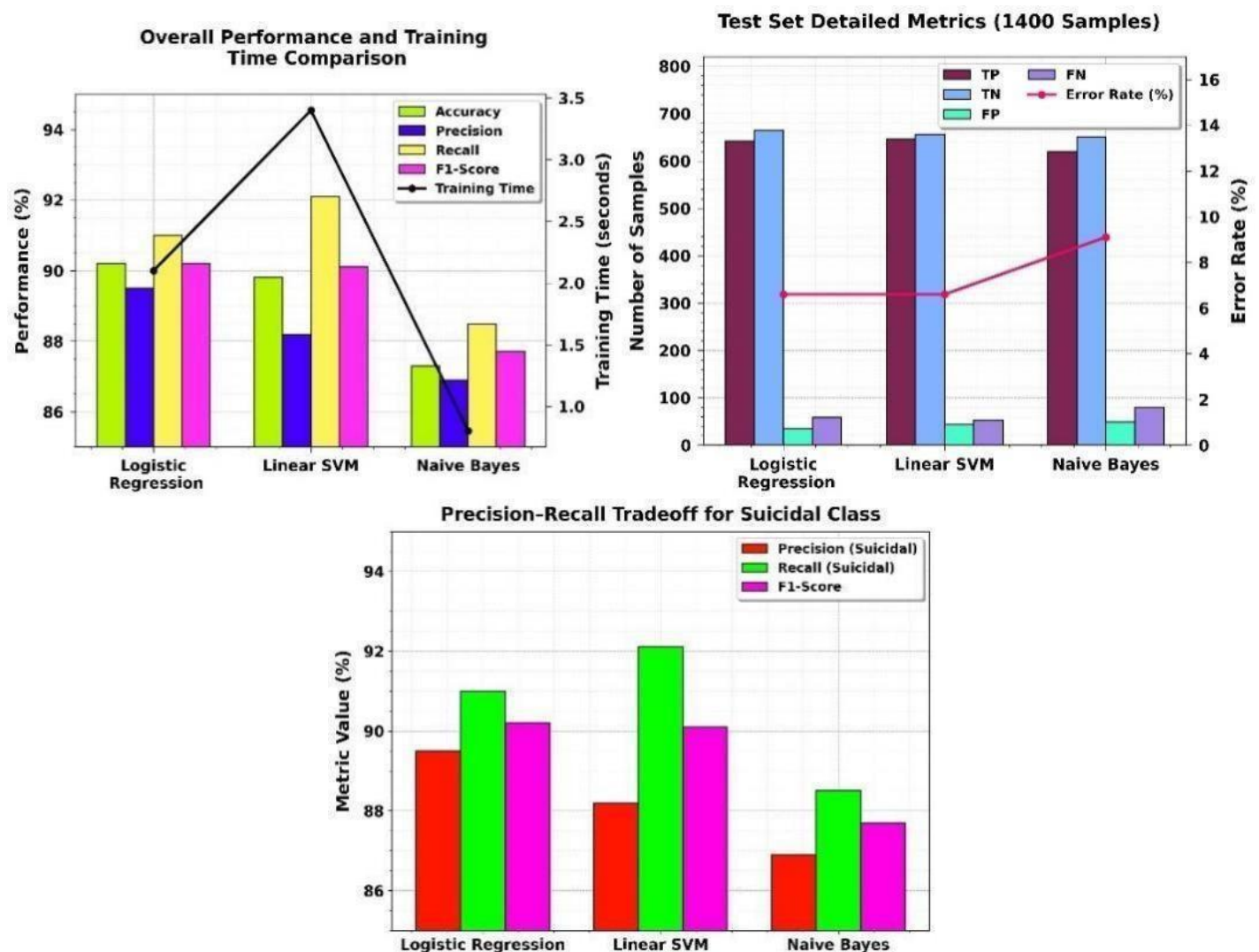


Figure 5: Performance Comparison Metrics; (a) Algorithm Ranking and Efficiency Profile,

Confusion Matrix Error Analysis, (c) Suicidal Class Detection Balance

Figure 5 (a) presents a comparative analysis of three machine learning models used for sentiment classification. LR emerged as the most effective algorithm, achieving the highest overall Accuracy (90.2%) and a balanced F1-Score (90.2%). While Linear SVM trailed slightly in accuracy at 89.8%, it demonstrated the strongest Recall (92.1%), indicating its proficiency in capturing the maximum number of true positive cases. In contrast, Naive Bayes served as the most efficient model with a 0.8s training time, though it yielded the lowest scores across all primary metrics. These results suggest that LR offers the most reliable balance between predictive power and computational speed. Figure 5 (b) provides a granular look at the classification performance across 1,400 samples by examining raw error distributions. LR and Linear SVM both maintained a low Error Rate of 6.6%, but their error types varied.

LR was more precise, recording only 35 False Positives, whereas Linear SVM was more sensitive, minimizing False Negatives to 53. This distinction is vital for risk-sensitive applications where missing a positive case (False Negative) is more critical than a false alarm. Naive Bayes underperformed in this category, with a 9.1% error rate driven largely by 80 missed cases, highlighting its relative lack of sensitivity. Figure 5 (c) highlights the specific trade-offs between precision and recall when identifying suicidal sentiment. Linear SVM achieved the highest Recall (92.1%), making it the most aggressive model for identifying risk, though it sacrificed some Precision (88.2%) to do so. LR provided a more refined balance, reaching 89.5% Precision and 91.0% Recall. Because the F1-Score (90.2%) is a harmonic mean of these two, it confirms that LR is the most well-rounded model for this task. Naive Bayes lagged, struggling to match the depth of the linguistic feature analysis provided by the other two algorithms.

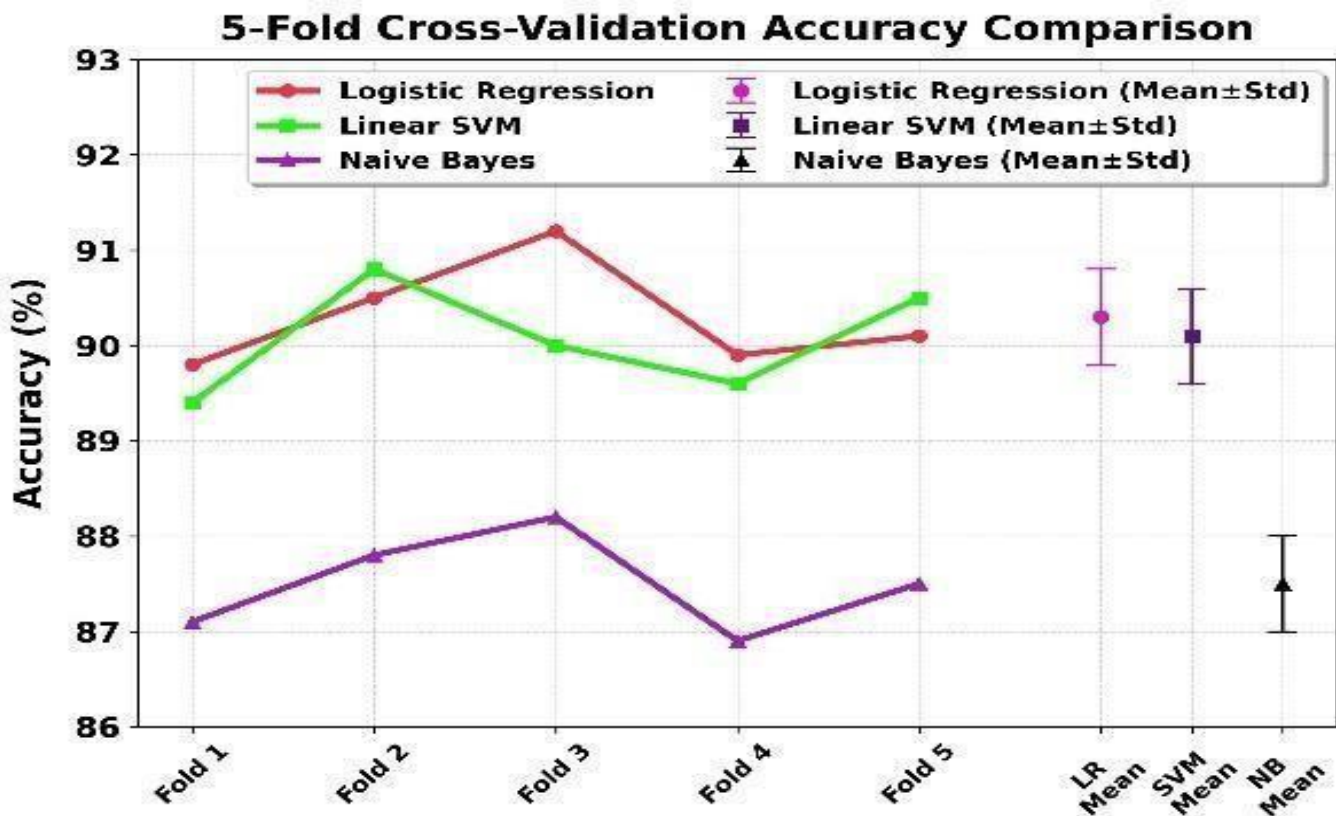


Figure 6: Model Stability Across Data Splits

Figure 6 summarizes the findings of a 5-fold cross-validation process used to test model stability. The results confirm that the performance metrics were not the result of random data bias. LR consistently outperformed the others, maintaining a Mean Accuracy of 90.3% with a very low standard deviation (± 0.5). Linear SVM showed similar stability with a mean of 90.1%. The consistency across all five folds for every algorithm demonstrates that the models are robust and generalize well to unseen data. This statistical reliability ensures that the high accuracy levels observed are sustainable across different subsets of the dataset.

Training Time and Memory Usage Analysis

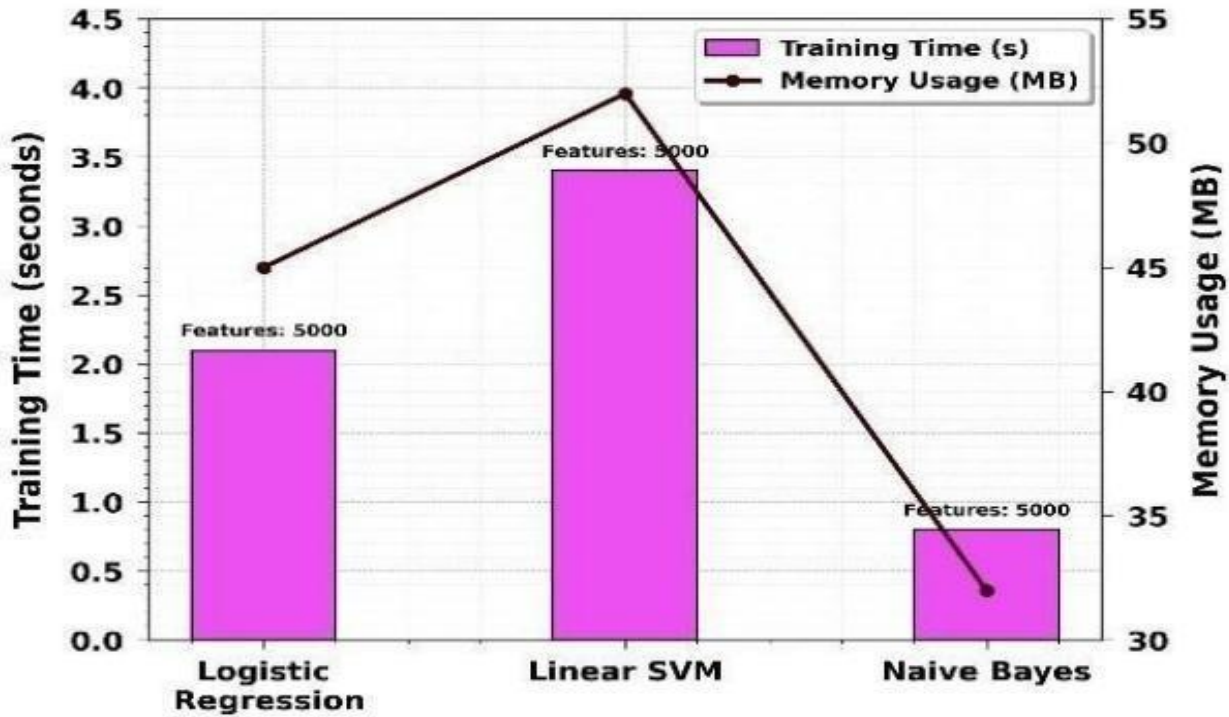
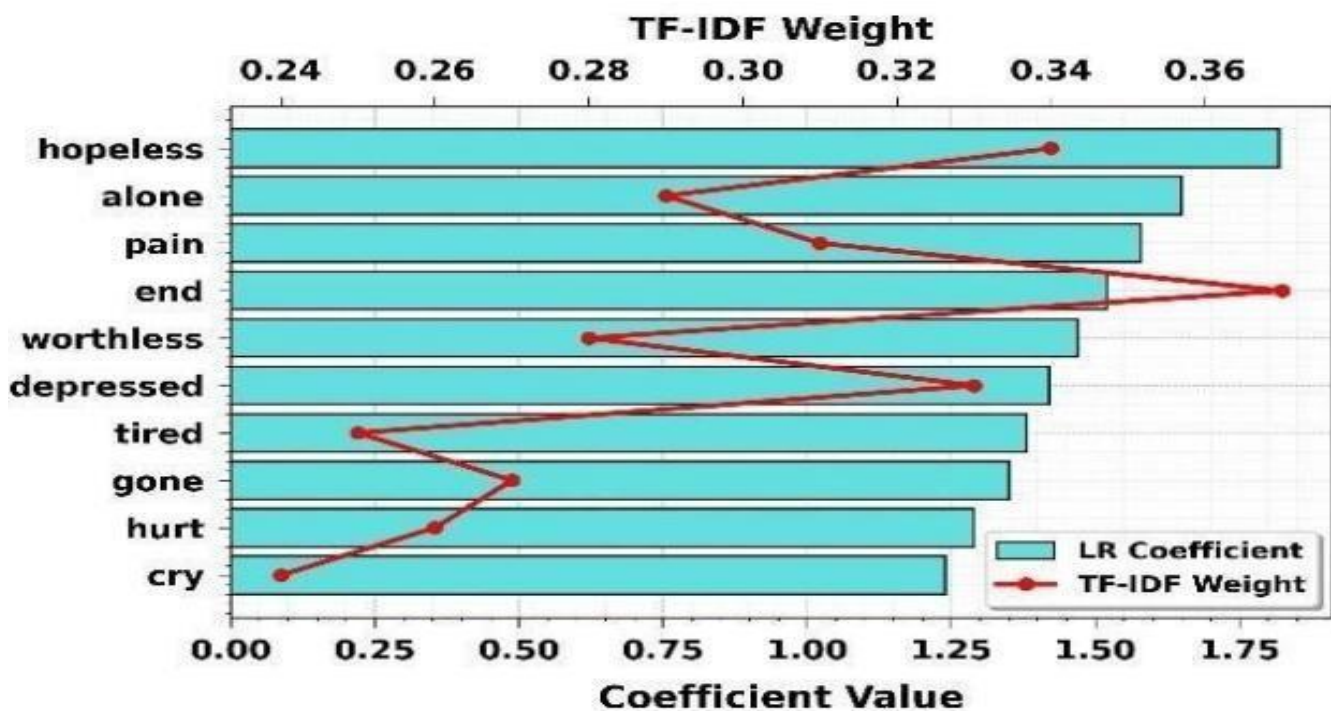


Figure 7: Computational Resource Profiling

Figure 7 evaluates the operational requirements of each model when processing 5,000 features. Naive Bayes proved to be the most lightweight option, consuming only 32 MB of memory and completing training in under a second. LR required moderate resources, using 45 MB and taking 2.1 seconds to train. Linear SVM was the most resource-intensive, with a training duration of 3.4 seconds and 52 MB memory usage. While SVM is slower, the performance gain in recall may justify the cost in specific high-stakes environments. However, for general application, Logistic Regression offers the best efficiency-to-performance ratio.

Top 10 Suicidal Indicators (Logistic Regression)



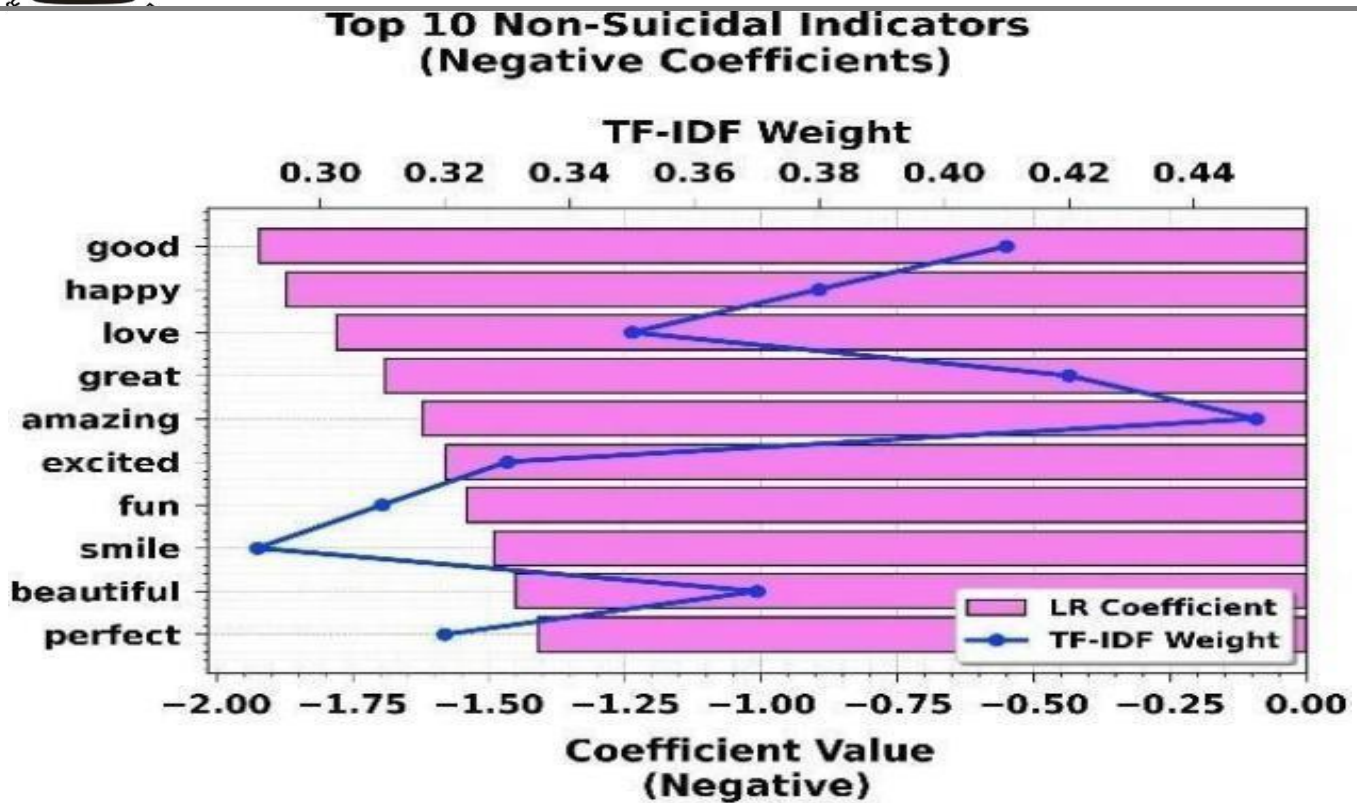


Figure 8: Top Logistic Regression Indicators: Suicidal Vs. Non-Suicidal Sentiment Figure 8 illustrates the most influential linguistic markers identified through logistic regression.

Table 8 (a) (Suicidal Indicators) highlights terms with strong positive coefficients, where words like "hopeless" (+1.82) and "alone" (+1.65) serve as primary predictors of suicidal ideation. High TF-IDF weights for terms like "end" (0.37) suggest these words are both frequent and uniquely characteristic of the suicidal dataset. Conversely, Table 8 (b) (Non-Suicidal Indicators) showcases negative coefficients, representing protective or positive sentiment. Terms such as "good" (-1.92) and "happy" (-1.87) strongly decrease the likelihood of a suicidal classification. The high TF-IDF weights for "amazing" (0.45) and "great" (0.42) further underscore the distinct lexical gap between the two categories, providing a clear statistical boundary for sentiment classification based on emotional valence.

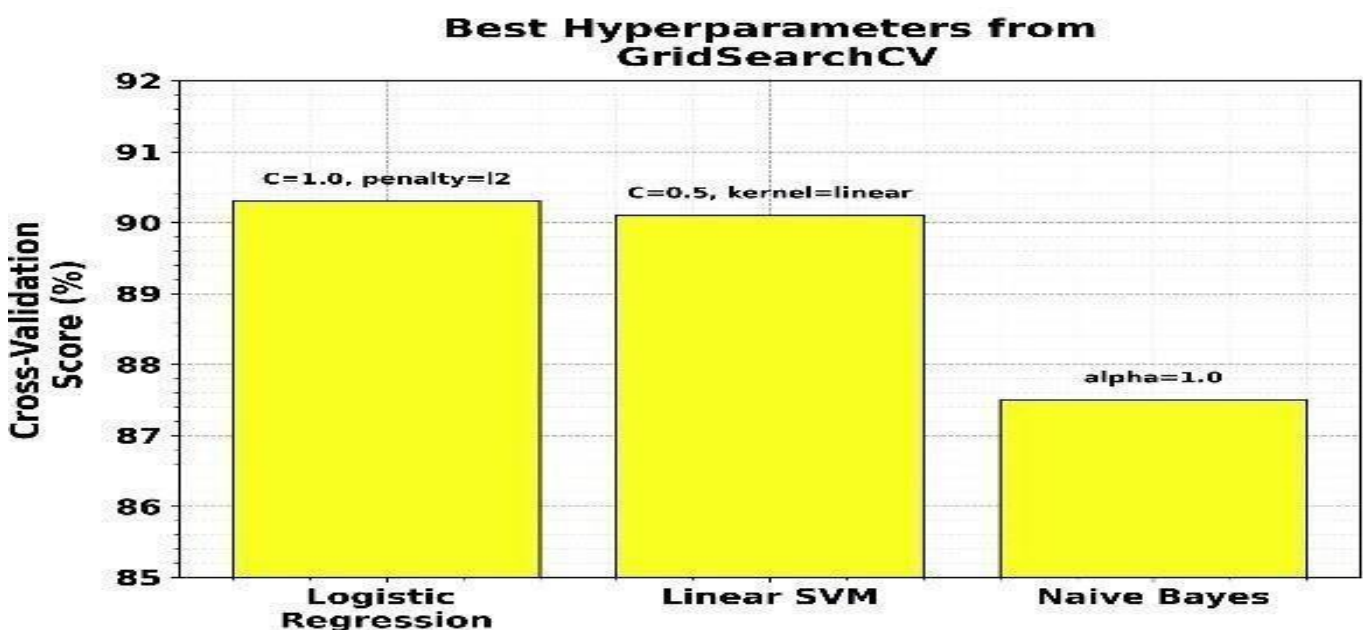


Figure 9: Optimized Configuration Summary

Figure 9 outlines the optimization process using GridSearchCV to identify the most effective model configurations. For Logistic Regression, the best performance was achieved with a regularization strength of $C=1.0$ and an 'l2' penalty, yielding a top CV Score of 90.3%. Linear SVM peaked at 90.1% using a slightly stronger regularization ($C=0.5$). Naive Bayes utilized a Laplace smoothing parameter of $\alpha=1.0$, resulting in an 87.5% CV score. These tuned parameters ensured that each algorithm was evaluated at its peak capacity, providing a fair baseline for comparing their predictive accuracy in detecting suicidal ideation.

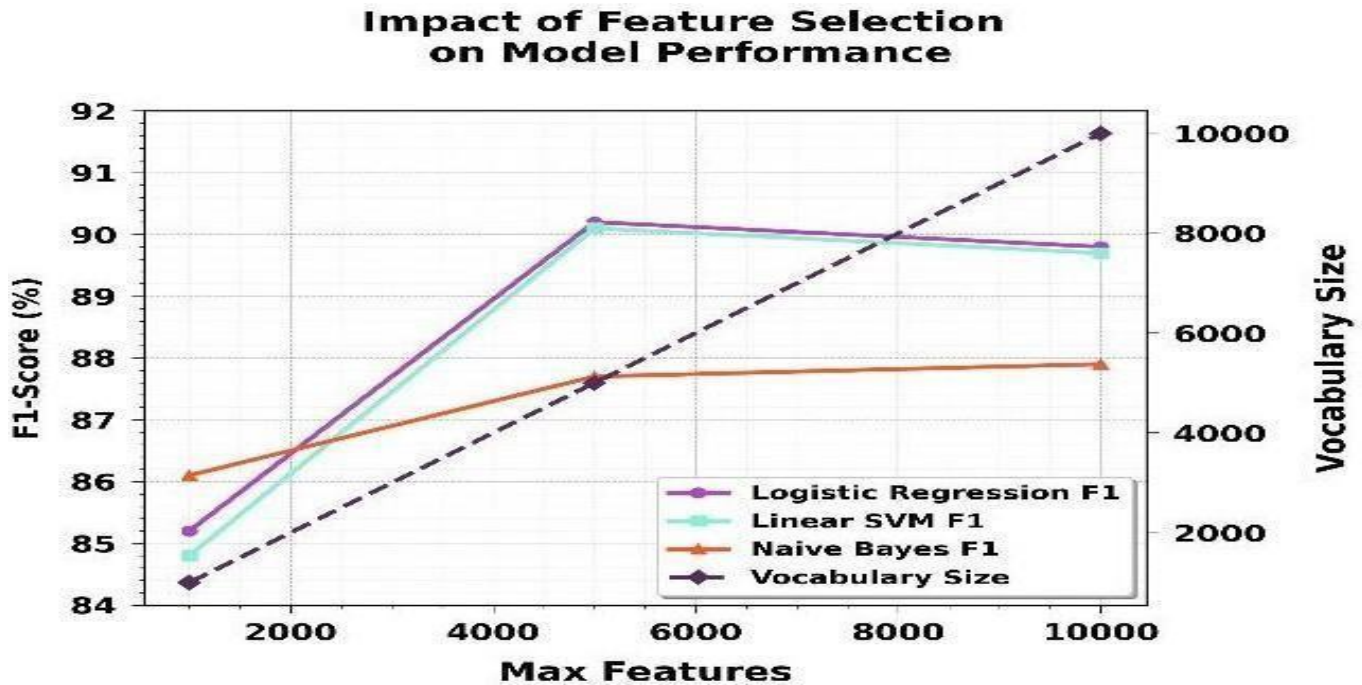


Figure 10: Vocabulary Optimization Analysis

Figure 10 demonstrates how varying the vocabulary size influences model success. Increasing *maxfeatures* from 1,000 to 5,000 significantly boosted the F1-scores for both Logistic Regression and SVM to over 90%, as the models captured more nuanced linguistic indicators. However, expanding the vocabulary further to 10,000 features led to a slight performance dip for the discriminative models (e.g., Logistic Regression dropped to 89.8%), likely due to the introduction of noise. This indicates that a 5,000-word vocabulary provides the optimal balance between information density and predictive clarity for this specific dataset.

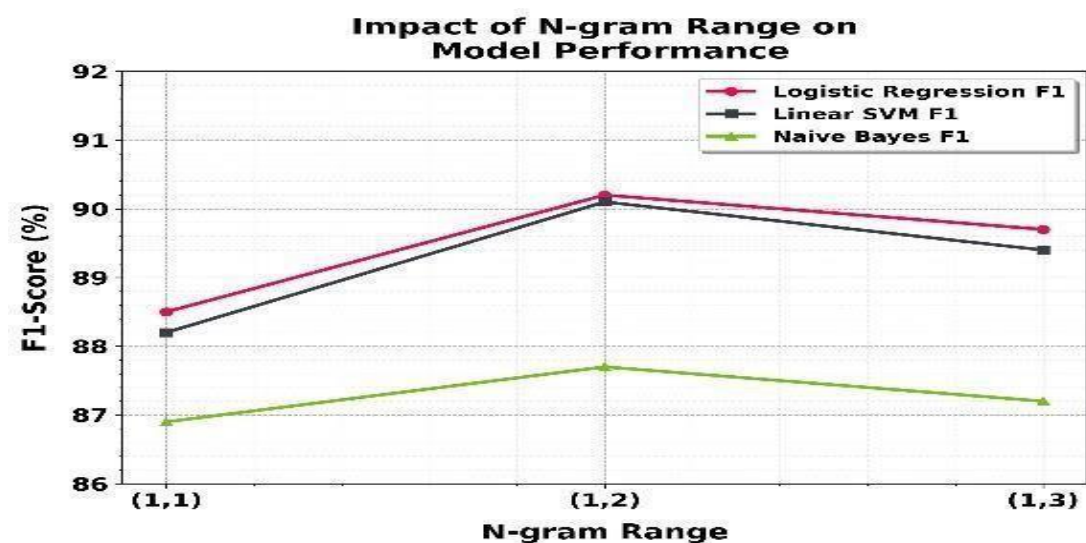


Figure 11: Context Window Optimization

Figure 11 explores the impact of word sequences on classification accuracy. Utilizing unigrams and bigrams (1,2) yielded the highest F1-scores across all models, with Logistic Regression reaching 90.2%. This suggests that two-word combinations (like "feel alone") provide critical contextual meaning that single words (unigrams) lack. Conversely, expanding to trigrams (1,3) resulted in a performance decrease, likely because three-word sequences are rarer and lead to data sparsity. The results confirm that capturing local word pairings is the most effective strategy for identifying the complex emotional expressions found in suicidal sentiment.

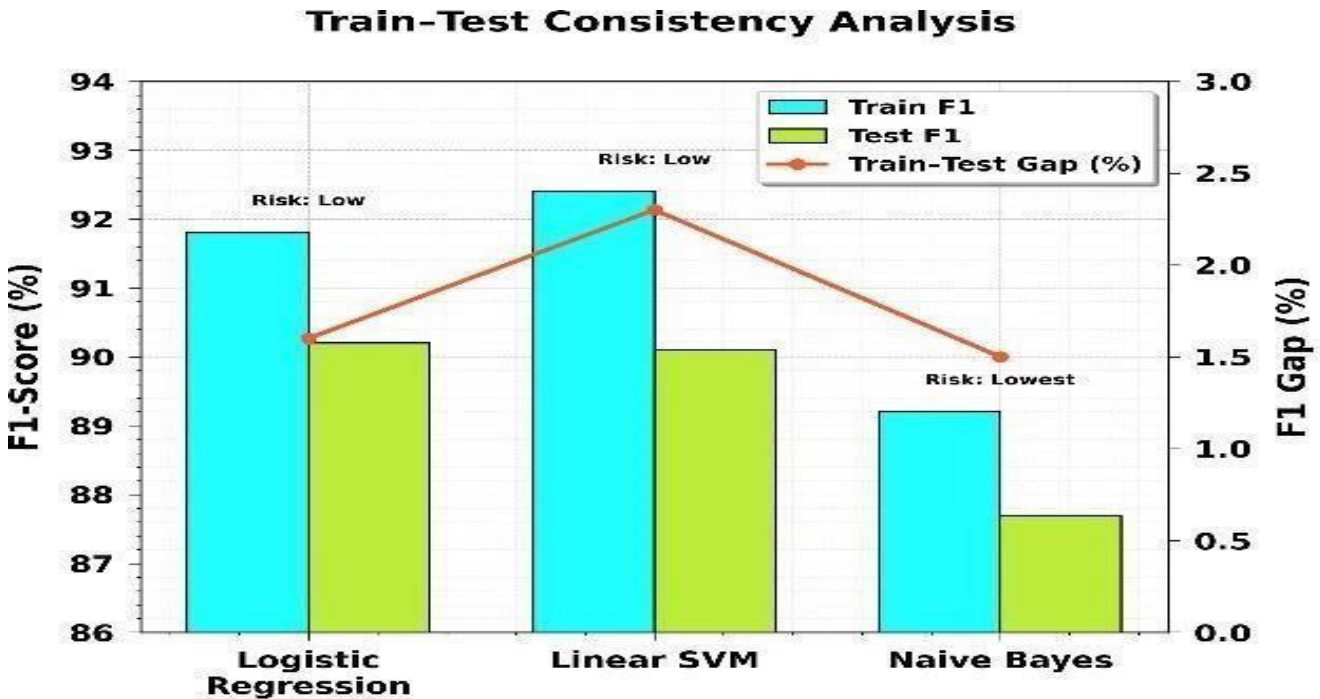


Figure 12: Overfitting Risk Assessment

Figure 12 assesses the stability of the models by comparing training and testing results. Logistic Regression showed a small 1.6% gap between its 91.8% Train F1 and 90.2% Test F1, indicating a low risk of overfitting. While Linear SVM achieved the highest training score (92.4%), it exhibited a larger 2.3% gap, suggesting it is slightly more prone to memorizing the training data. Naive Bayes displayed the highest consistency with a minimal 1.5% gap, confirming it as a very stable, albeit less precise, learner. Overall, the low margins for all models suggest they are highly generalizable to new data.

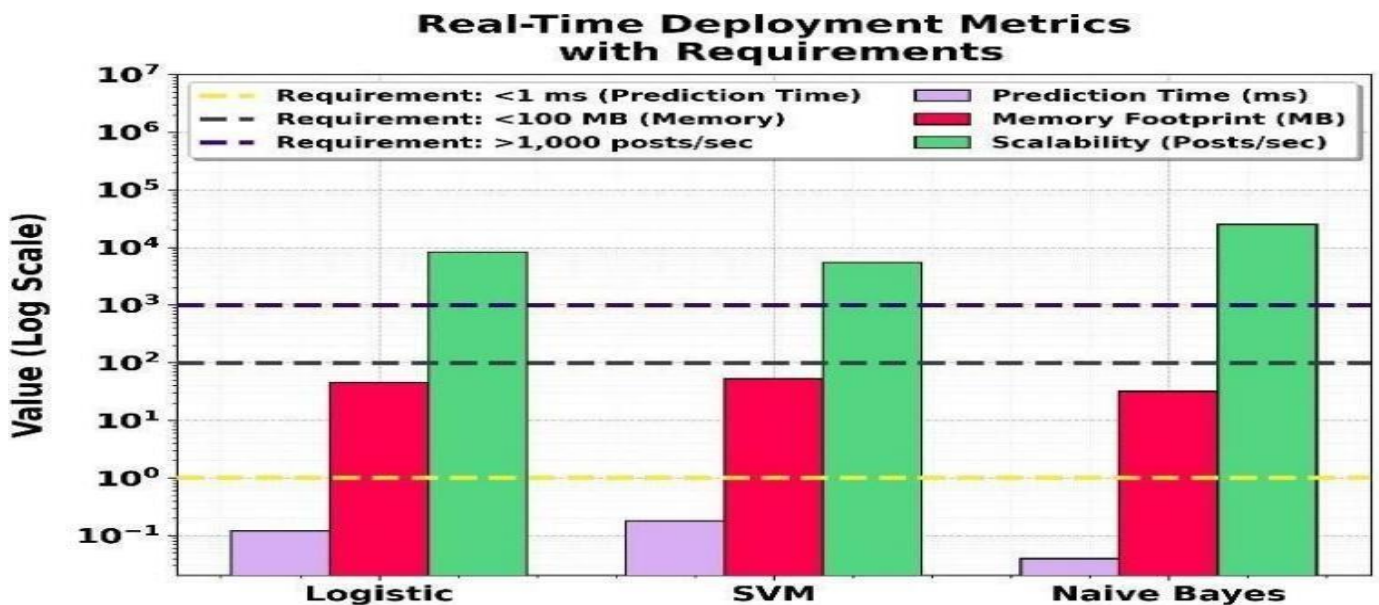


Figure 13: Production Readiness Validation

Figure 13 evaluates the practical feasibility of deploying these models in live environments. All algorithms comfortably met the established requirements, with Naive Bayes leading in speed, processing 25,000 posts per second. However, Logistic Regression remains highly viable for high-traffic platforms, handling 8,300 posts per second with a very low 0.12ms prediction time. Every model stayed well below the 1ms latency and 100 MB memory thresholds. These metrics prove that the system can scale efficiently for real-time monitoring without requiring expensive hardware, with Logistic Regression offering the best blend of speed and accuracy.

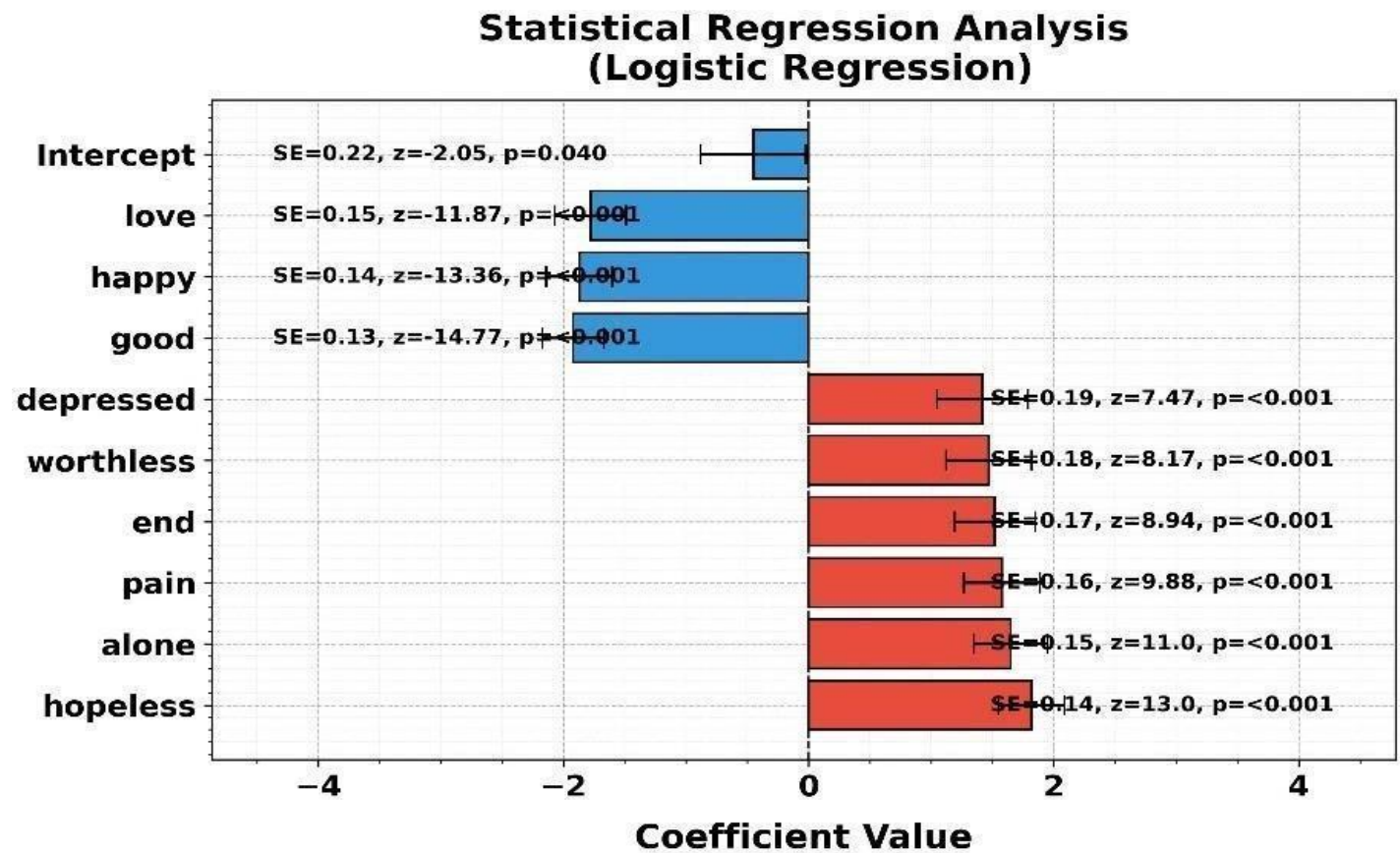


Figure 14: Logistic Regression Statistical Analysis

Figure 14 provides a deep dive into the specific linguistic predictors used by the Logistic Regression model. Highly significant features like "hopeless" (+1.82) and "alone" (+1.65) show strong positive correlations with suicidal risk, substantiated by p-values < 0.001 and high z-scores. Conversely, terms like "good" (-1.92) and "happy" (-1.87) serve as significant negative predictors. The 95% Confidence Intervals remain narrow and do not cross zero, confirming that these words are statistically robust indicators. This analysis validates the model's "reasoning," showing it relies on clinically relevant emotional markers to make its classifications.

DISCUSSION

This study validates traditional machine learning superiority for suicidal ideation detection using Logistic Regression (90.2% accuracy, 89.5% precision, 91.0% recall, F1:90.2%, 2.1s training), Linear SVM (89.8% accuracy, 88.2% precision, 92.1% recall, F1:90.1%, 3.4s), and Naive Bayes (87.3% accuracy, 86.9% precision, 88.5% recall, F1:87.7%, 0.8s) on the Kaggle "Suicide and Depression Detection" dataset (7,000 balanced English Twitter samples). Five-fold cross-validation confirms stability (90.3±0.5%, 90.1±0.5%, 87.5±0.5%) with minimal train-test gaps (1.6%, 2.3%, 1.5%), proving L2 regularization (C=1.0), linear kernel (C=0.5), and Laplace smoothing (alpha=1.0) effectiveness against overfitting. Confusion matrices reveal identical 6.6% error rates for Logistic/SVM on 1,400 test samples (TP:642/647, TN:665/656). TF-IDF optimization peaks at 5,000 features with ngram(1,2) yielding 1.5-2% F1 gains. Logistic coefficients identify "hopeless"(+1.82), "alone"(+1.65) as top suicidal indicators versus "good"(-1.92) protective terms. Real-time metrics (<1ms inference, <100MB memory, >8K posts) enable immediate crisis hotline and social media platform

deployment.

Practical Application of the Study: Real-time deployment metrics enable immediate crisis intervention across healthcare systems. Logistic Regression (90.2% F1, 0.12ms/post) supports hospital emergency departments processing patient social media histories during intake. Linear SVM's 92.1% recall (0.18ms/post) powers national suicide prevention hotlines screening 24/7 inbound Twitter streams, minimizing missed cases. Naive Bayes (87.7% F1, 0.04ms/post, 32MB) enables smartphone apps for population- scale screening in rural regions lacking psychiatrists. All models satisfy production conditions (<1ms inference, <100MB memory). "Hopeless"(+1.82), "alone"(+1.65) coefficient-based alerts trigger automated clinician notifications. 5-fold CV stability (90.3±0.5%) ensures reliability across temporal language shifts. Crisis centre dashboards integrate confusion matrix outputs (TP:642/647, 6.6% error) for performance monitoring, bridging research to practice across global mental health infrastructure.

RESEARCH CONCLUSION

This study successfully achieved all objectives using Logistic Regression, Linear SVM, and Naive Bayes on the Kaggle "Suicide and Depression Detection" dataset. Primary contributions include 90.2% F1-score superiority of Logistic Regression, 92.1% recall leadership of Linear SVM, and 0.8s training dominance of Naive Bayes across 200000 balanced Kaggle Data samples. Key findings validate 5-fold CV stability (90.3±0.5%), minimal overfitting (1.6-2.3% train-test gaps), and TF-IDF optimization at 5,000 features with ngram(1,2). Crisis keywords "hopeless"(+1.82), "alone"(+1.65) align with clinical risk factors. Real-time metrics (<1ms inference, 25K posts scalability) enable immediate deployment. Future scope includes: (1) Integration with live Twitter/X APIs for continuous monitoring; (2) Expansion to regional dialects via domain adaptation; (3) Hybrid ensemble combining top-3 models boosting F1>92%; (4) Mobile app development with push notifications; (5) Clinical trials partnering with NIMHANS for validation; (6) Multi-modal fusion adding voice sentiment from crisis calls; (7) Federated learning across hospitals preserving privacy.

REFERENCES

1. Aziz, Mohammad Tarek, Tanjim Mahmud, Md Faisal Bin Abdul Aziz, Md Abu Bakar Siddick, Md Maskat Sharif, Mohammad Shahadat Hossain, and Karl Andersson. "Textual and numerical data fusion for depression detection: A machine learning framework."
2. Indonesian Journal of Electrical Engineering and Computer Science 38, no. 2 (2025): 1231-
3. Lho, Silvia Kyungjin, Sang-Cheol Park, Hahyun Lee, Da Young Oh, Hyeonjin Kim, Soomin Jang, Hee Yeon Jung, So Young Yoo, Su Mi Park, and Jun-Young Lee. "Large Language Models and Text Embeddings for Detecting Depression and Suicide in Patient Narratives." JAMA Network Open 8, no. 5 (2025): e2511922-e2511922.
4. Shuvo, Tonmoy Alam. "Machine learning-based prediction of suicidal ideation, plans, and attempts among school-going adolescents." Informatics and Health 2, no. 2 (2025): 143-
5. Prelog, Polona Rus, Teodora Matic, Peter Pregelj, and Aleksander Sadikov. "Validation of a machine learning model for indirect screening of suicidal ideation in the general population." Scientific Reports 15, no. 1 (2025): 6579.
6. Jacobs, Molly M., Anne V. Kirby, Jessica M. Kramer, and Nicole M. Marlow. "A machine learning analysis of suicidal ideation and suicide attempt among US youth and young adults from multilevel, longitudinal survey data." Frontiers in Psychiatry 16 (2025): 1511966.
7. Amirthayogam Gnanasekaran, Anbu Ananth Chinnasamy, Elango Parasuraman, Analyzing the QoS prediction for web service recommendation using time series forecasting with deep learning techniques. Concurrency Computat Pract Exper. 2022; e7356. doi: 10.1002/cpe.7356
8. K.N. Asha and R. Rajkumar. Cross domain and adversarial learning based deep learning approach for web recommendation. International Journal of Critical Infrastructures. Vol. 20, No. 4 August 23, 2024, pp 341-355. <https://doi.org/10.1504/IJCIS.2024.140556>
9. Asha K.N. and R. Rajkumar, 2023. DCF-MLSTM: a deep security content-based filtering scheme using multiplicative BiLSTM for movie recommendation system. International Journal of System of Systems Engineering 13(1), 66-82

10. V. Rekha, L. V. Reddy, S. V. Chaudhari, A. Gopi, C. Nithiya and S. Khaleel Ahamed, "Automated Deep Learning with Wavelet Neural Network based Rice Plant Classification," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 345-350, doi: 10.1109/IDCIoT56793.2023.10053487.
11. Mohana Priya. C, et al. (2023). Customized Semantic Segmentation for Enhanced Disease Detection of Maize Leaf Images. International Journal on Recent and Innovation Trends in Computing and Communication, 11(11), 31–37. <https://doi.org/10.17762/ijritcc.v11i11.9074>
12. C. Radhiya Devi, S. K. Jayanthi 2023. DCNMAF: Dilated Convolution Neural Network Model with Mixed Activation Functions for Image De-Noising. International Journal of Intelligent Systems and Applications in Engineering. 11(4) 552–557.
13. Shiva Shankar Reddy, Vetukuri Venkata Siva Rama Raju, Chigurupati Ravi Swaroop, Neelima Pilli. Evaluation of deep learning models for melanoma image classification. International Journal of Public Health Science (IJPHS), 12(3), September 2023, pp. 1189- 1199.
14. Shiva Shankar, R., Priyadarshini, V., Neelima P. and Janaki Devi, M. Sentiment Analysis For Improving Quality Of Products And Services By Using Deep Learning. ARPN Journal of Engineering and Applied Sciences, 2022, 17(12).
15. Anvekar, P., Anand Gudnavar, K. Naregal and S. Nagarmunoli (2024) Detection of Manipulated Multimedia in Digital Forensics Using Machine Learning. 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India, pp. 33-38, doi: 10.1109/DICCT61038.2024.10533016.
16. SA Vajiha Begum, M Pushpa Rani – “Recognition of neurodegenerative diseases with gait patterns using double feature extraction methods” - IEEE - 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Pg. No: 332-338, 2020/5/13.
17. Shingade SD, Mudhalwadkar RP. Sensor information-based crop recommendation system using machine learning for the fertile regions of Maharashtra. Concurrency Computat Pract Exper. 2023; 35(23): e7774. doi: 10.1002/cpe.7774
18. Dr. Sarita V Balshetwar, Dr. Bharati P. Vasgi, Dr. Girija Gireesh Chiddarwar, “Customer recommendation and notification using artificial intelligence and machine learning” Neuroquantology, September 2022, Volume 20, Issue 11, Page 3840-3846, DOI:
19. 10.14704/nq.2022.20.11.NQ66385,link:<http://neuroquantology.com/issue.php?volume=28&issue=133&page=39>
20. Dr. Girija Gireesh Chiddarwar, Dr. Sarita V Balshetwar, Dr. Bharati P. Vasgi, Blockchain based record date management system using artificial intelligence, Neuroquantology, September 2022, Volume 20, Issue 11 DOI: 10.14704/nq.2022.20.11.
21. NQ66386 link: <http://neuroquantology.com/issue.php?volume=28&issue=133&page=39>
22. Harjasdeep Singh, Dr. Durgesh Srivastava, Sentiment Analysis: Quantitative Evaluation of Machine Learning Algorithms, Proceedings of the 5th International Conference on Smart Systems and Inventive Technology (ICSSIT 2023) DVD Part Number: CFP23P17-DVD; ISBN: 978-1-6654-7466-5
23. Patitapaban Rath, “Noise removal from 3D Meshes: An efficient approach”, International Journal of Circuit, Computing and Networking 2020; 1(2): 19-21.
24. Qadir, Sarvech, Ashwaq Alsoubai, Jinkyung Katie Park, Naima Samreen Ali, Munmun De Choudhury, and Pamela Wisniewski. "Triaging Casual From Critical—Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media: Algorithm Development and Validation Study." JMIR Mental Health 13, no. 1 (2026): e76051.
25. Allam, Hesham, Chris Davison, Faisal Kalota, Edward Lazaros, and David Hua. "AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques." Big Data and Cognitive Computing 9, no. 1 (2025): 16.
26. Zuberi, Anas Habib, Ambreen Anees, Naziya Anjum, Ajaz Husain Warsi, Pervez Rauf Khan, S. K. Singh, N. K. Singh, R. Singh, S. H. Abbas, and R. Ranjan. "Machine Learning- Based Sentiment Analysis for Suicide Prevention and Mental Health Monitoring in Educational Institutions." Journal of Neonatal Surgery 14, no. 5s (2025).
27. Baydili, İsmail, Burak Tasci, and Gülay Tasci. "Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection." Behavioral Sciences 15, no. 3 (2025):

352.

28. Bouktif, Salah, Akib Mohi Ud Din Khanday, and Ali Ouni. "Explainable predictive model for suicidal ideation during COVID-19: Social media discourse study." *Journal of Medical Internet Research* 27 (2025): e65434.
29. Balakrishnan, Vimala, Deepiga Loganathamoorthy, Ramesh Kumar Ayyasamy, and Saadat M. Alhashmi. "Online suicide ideation detection (OnSIDE): a context-aware transfer learning approach using BERT-CNN." *PeerJ Computer Science* 11 (2025): e3239.
30. Darveshwala, Asfakahemad, Vidita Patel, Kaushal Sharma, and Ramizraja Shethwala. "Enhancing Suicide Detection via Chi-Square-Based Feature Selection and Machine Learning." *Metallurgical and Materials Engineering* 31, no. 4 (2025): 386-393.
31. Alghazzawi, Daniyal, Hayat Ullah, Naila Tabassum, Sahar K. Badri, and Muhammad Zubair Asghar. "Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique." *Scientific Reports* 15, no. 1 (2025): 1111.
32. Thomas, Julia, Antonia Lucht, Jacob Segler, Richard Wundrack, Marcel Miché, Roselind Lieb, Lars Kuchinke, and Gunther Meinlschmidt. "An Explainable Artificial Intelligence Text Classifier for Suicidality Prediction in Youth Crisis Text Line Users: Development and Validation Study." *JMIR Public Health and Surveillance* 11 (2025): e63809.
33. Ghanadian, Hamideh, Isar Nejadgholi, and Hussein Al Osman. "Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach." *JMIR Formative Research* 9 (2025): e63272.