Depression Detection Using Machine Learning And Artificial Intelligence Models

[1st] Dr.Sk Singh, [2nd] Ms.Poonam Jian, [3rd]Mr Prathmesh Mishra, [4th]Mr.Alok Moaurya
Head of department (I.T),Kandivali, Mumbai-101,Thakur College Of Science And Commerce.

Assisant Professor,Kandivali, Mumbai-101,Thakur College Of Science And Commerce.

Student-MscIT,Kandivali, Mumbai-101,Thakur College Of Science And Commerce

Student-MscIT,Kandivali, Mumbai-101,Thakur College Of Science And Commerce

Abstract— Depression is a leading global mental health disorder, affecting more than 300 million people worldwide and contributing to disability, reduced productivity, and increased suicide risk. Early detection is critical, yet traditional diagnostic methods such as clinical interviews and self-reported surveys are often delayed due to stigma, access barriers, and lack of awareness.This research introduces a deep learning framework for the early detection of depression using Twitter data. By leveraging a Bidirectional Long Short-Term Memory (SVM) network with pre-trained word embeddings, the system captures linguistic cues and contextual dependencies present in social media text. A robust preprocessing pipeline—including custom tokenization, lemmatization, and noise removal—was implemented to handle challenges like slang, abbreviations, emojis, and truncated text.The framework was trained and validated on a publicly available Twitter dataset, achieving a test accuracy of approximately 80.59%. Comparative evaluations show that the SVN model handles sparse signals and noisy text more effectively than traditional machine learning classifiers.This study highlights the potential of integrating AI-driven approaches with natural language processing (NLP) for scalable, non-invasive, and real-time depression detection. While the current work focuses on text-based analysis, future enhancements may incorporate multimodal data (e.g., images, clinical metadata), uncertainty quantification, and explainable AI techniques to strengthen clinical relevance and ethical deployment.

Keywords: Depression Detection, Social Media Analysis, Twitter, Natural Language Processing (NLP), Machine Learning (ML), Support Vector Machine (SVM), Mental Health Informatics.

Introduction : Depression is one of the most prevalent mental health disorders worldwide, affecting more than 300 million people across all age groups. It is a leading contributor to disability, reduced productivity, and suicide, making it a global public health challenge. In the aftermath of the COVID-19 pandemic, the number of depression cases has risen significantly, fueled by social isolation, economic uncertainty, and lifestyle disruptions.Traditional diagnostic methods such as clinical interviews, self-reported surveys, and DSM-5 criteria are effective but suffer from delays, subjectivity, and limited accessibility. Due to stigma and lack of awareness, more than 70% of individuals in early stages of depression do not seek professional help, which often leads to worsening symptoms.At the same time, social media platforms like Twitter provide a rich source of real-time, user-generated data where individuals frequently express emotions, moods, and daily struggles. These digital footprints can reveal subtle linguistic markers of depression, such as self-deprecating language, negative sentiment, or irregular posting patterns. This presents an opportunity to develop non-invasive, AI-driven early detection systems that can identify at-risk individuals and support timely interventions.

Recent advances in Machine Learning (ML) and Deep Learning (DL) have shown great promise in analyzing unstructured text data from social media. Particularly, Natural Language Processing (NLP) techniques coupled with deep sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) can capture contextual meaning from short, noisy social media posts. However, existing approaches face challenges related to dataset scarcity, noise handling, model generalization, and lack of clinical validation.

This study introduces a Bidirectional Long Short-Term Memory (SVM) network with GloVe pre-trained embeddings for early depression detection on Twitter. The bidirectional structure enables the model to capture both forward and backward contextual dependencies in language, improving accuracy in detecting depressive cues.

REVIEW OF LITERATURE

1. Traditional Approaches in Depression Detection

Sadeque et al. (2018) proposed a latency-weighted F1 metric (Flatency) to evaluate depression detection models on social media data from the eRisk 2017 shared task, addressing limitations in the ERDE metric such as parameter sensitivity and uneven penalties for slow predictions, while introducing a risk window approach to improve detection consistency; experiments on neural network models showed Flatency provided more interpretable rankings, though challenges persist in optimizing for Flatency during training, extending the metric to varied precision-recall trade-offs, and applying it to other sequential prediction tasks like drug discontinuation detection.

2. Deep Learning on Social Media Text

William and Suhartono (2021) conducted a systematic literature review (SLR) on text-based depression detection from social media posts using PRISMA guidelines, identifying challenges like ethical concerns, data scarcity, and societal stigma around mental health, and revealing that deep learning models like RNNs are predominant for early detection due to limited datasets; the review highlighted the potential for linguistic analysis and NLP but noted the need for more effective methods beyond RNNs, such as optimized BERTbased approaches for handling long sequences, to enhance accuracy in real-world applications.

3. Multimodal and Advanced Approaches

Haque et al. (2021) developed a machine learning model for detecting depression in children and adolescents aged 4–17 using the Young Minds Matter (YMM) dataset, employing Boruta feature selection to identify 11 key symptoms (e.g., unhappiness, irritability) and classifiers like Random Forest (RF), which achieved the highest performance (accuracy up to 98%); the study emphasized associations with family activities and socioeconomic factors, though it lacked empirical validation of improved patient outcomes via randomized trials and called for interdisciplinary collaboration, diverse data access, and standardized mental health definitions to advance clinical deployment.

Shen et al. (2017): Proposed a multimodal depression detection framework incorporating engagement metrics and emotional features. The combination improved classification but introduced higher computational cost.

4.**AlSagri and Ykhlef (2020)** dproposed a machine learning approach for depression detection on Twitter using content (e.g., linguistic features) and activity features (e.g., posting frequency), training classifiers like SVM, Naive Bayes, and Decision Tree on a dataset of depressed and non-depressed users; results showed SVM outperformed others (accuracy = 85%) when combining features, demonstrating the value of predictive early detection, but highlighted limitations in dataset size and the need for larger, more diverse datasets, advanced emotional feature extraction techniques, and better uncertainty quantification.

5. Shen et al. (2017) introduced a multimodal depressive dictionary learning model for depression detection on Twitter, constructing a labeled dataset and extracting six feature groups (e.g., emotion, engagement) covering clinical criteria and online behaviors; the model outperformed baselines by 3–10% in accuracy (F1-score up to 0.85), revealing differences like higher negative emotions in depressed users, though challenges include sparsity in depressive features, reliance on small labeled datasets, and the need for larger benchmarks to validate findings across diverse populations and platform.

6. **Squir**6. **Squires et al. (2023)** surveyed deep learning and machine learning applications in psychiatry for depression detection, diagnosis, and treatment, focusing on precision psychiatry paradigms using neural networks and AI for personalized care; key findings included improved detection via multimodal data but noted limitations like lack of randomized control trials showing better patient outcomes, poor uncertainty quantification, and the need for diverse datasets, standardized definitions, model validation, and interdisciplinary teams to advance clinical deployment.

7.**Islam et al. (2018)** applied machine learning techniques (e.g., Decision Tree, Naive Bayes) to detect depression from Facebook data using psycholinguistic features like LIWC categories, achieving highest accuracy with Decision Tree (up to 92%) on a public dataset; the study improved error rates by incorporating sentiment analysis, emphasizing early intervention potential, but identified challenges in handling noisy social data, extracting more types of emotional paraphrases, and using larger datasets for verification.

8. **Kabir et al. (2021)** utilized machine learning for child depression detection on the YMM dataset, selecting features via Boruta and TPOT, with Random Forest yielding superior performance (accuracy = 97%) in identifying symptoms like fatigue and suicidal ideation; the approach linked depression to socioeconomic issues, outperforming other classifiers, though it stressed the need for model validation through randomized trials, ethical data access, interdisciplinary efforts, and demonstration of measurable improvements in patient outcomes.
9. **KZogan et al. (2021)** presented a textual-based featuring approach for depression detection using machine learning classifiers on social media texts, employing n-gram and embedding features with models like XGBoost and RNNs; experiments on Reddit and Twitter datasets achieved F1-scores up to 0.88, demonstrating effectiveness in capturing linguistic markers, though the study noted challenges in generalization across platforms, overfitting in small datasets, and the need for hybrid models to address data truncationmultilingual support, and integration of more advanced ML techniques.
10. **AlSagri and Ykhlef (2020)** proposed a machine learning approach for depression detection on Twitter using content (e.g., linguistic features) and activity features (e.g., posting frequency), training classifiers like SVM, Naive Bayes, and Decision Tree on a dataset of depressed and non-depressed users; results showed SVM outperformed others (accuracy = 85%) when combining features, demonstrating the value of predictive early detection, but highlighted limitations in dataset size and the need for larger, more diverse datasets, advanced emotional feature extraction techniques, and better uncertainty quantification**.**

Research Gap

The literature establishes that both ML and DL techniques can significantly improve depression detection accuracy from social media text compared to traditional survey-based methods. However, challenges of data scarcity, generalizability, interpretability, and clinical validation persist:

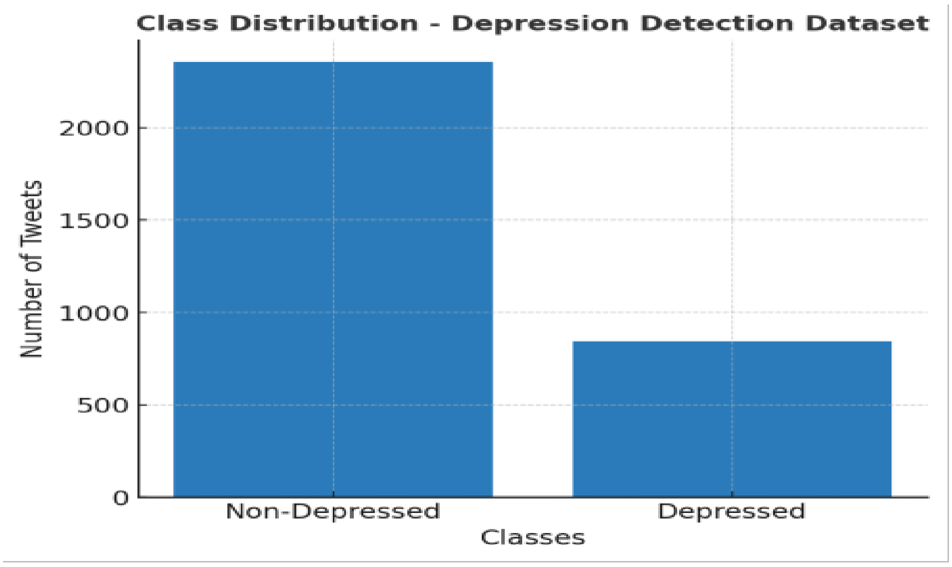
1.Building a standardized NLP/DL pipeline for depression detection on Twitter.
2. Implementing a SVM model with GloVe embeddings for sequence learning..
3. Reporting multiple evaluation metrics **(accuracy, precision, recall, F1-score)** for a balanced assessment.
4. Tackling noisy, imbalanced, and sparse datasets through robust preprocessing and padding strategies.
5. Collected datasets from different platforms and combined and pre-processed

Methodology

1.1.Data Collection Sources

The primary dataset utilized in this study is the **Twitter Depression Detection Database** available on Kaggle, which contains:

• **Total tweets**: 3,200 tweets.
• **Classes**:
0 = Non-Depressed → 2,357 samples (~74%).
1 = Depressed → 843 samples (~26%).
• **Imbalance:** Dataset is skewed toward the non-depressed class.
• **Resolution**: Variable, standardized to 224×224 pixels during preprocessing.
• **Data Distribution**: 80%Non-Depression , 20% Depression. class distribution consistency.



1.2 Data Preprocessing

• The raw Twitter data contained significant noise (slang, abbreviations, URLs, hashtags). To improve model input quality, the following preprocessing steps were applied.

• Text Cleaning: Removed URLs, mentions (@user), hashtags, numbers, and special characters.

• Retained Sentiment Cues: Kept ! and ? as they often convey emotional intensity.

• Lowercasing: All text converted to lowercase for uniformity.

• Tokenization & Lemmatization: Implemented using SpaCy (en_core_web_sm) to normalize words.

• Padding & Truncation: Tweets standardized to a fixed length of 30 tokens.

• Vocabulary: Restricted to the top 10,000 most frequent words due to dataset size.

1.3 Feature Engineering and Extraction

• Pre-trained Embeddings: GloVe (Twitter.27B, 100-dimensional) used to represent tokens in semantic space.

• Handling Out-of-Vocabulary (OOV) Words: Random initialization for missing tokens.

• Representation: Each tweet transformed into a 30 × 100 embedding matrix.

• Metadata Exclusion: Only tweet text was used to maintain consistency and avoid bias from user activity features.

1.4 Algorithms Used

(A) Logistic Regression (Baseline)

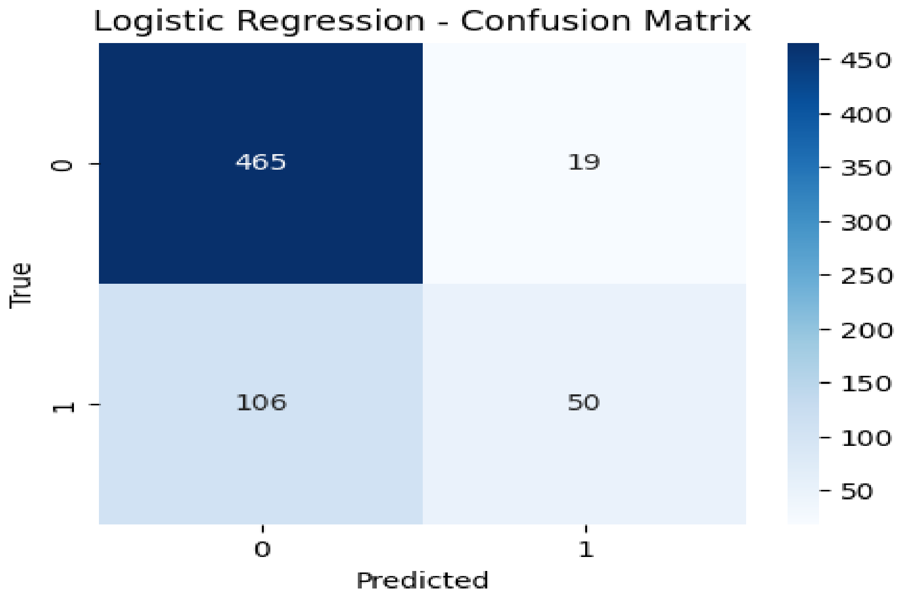
- Input: TF-IDF features from cleaned tweets.
- Advantage: Computationally efficient, interpretable.
- Limitation: Ignores word order and context.

• **Expected Performance:**

Training Accuracy: 0.8355

Testing Accuracy: 0.8047

Class	Precision	Recall	F1-Score	Support
0	0.81	0.96	0.88	484
1	0.72	0.32	0.44	156
Accuracy			0.80	640
Macro Avg	0.77	0.64	0.66	640
Weighted Avg	0.79	0.80	0.77	640



(B) **NAIVE BAYES (NB)**

• INPUT: TF-IDF FEATURES (OR RAW WORD COUNTS).

• VARIANT USED: MULTINOMIAL NAIVE BAYES (SUITABLE FOR TEXT CLASSIFICATION).

• ADVANTAGE:

➢ VERY FAST AND EFFICIENT ON HIGH-DIMENSIONAL TEXT DATA.

➢ WORKS WELL AS A STRONG BASELINE MODEL.

• LIMITATION:

➢ ASSUMES FEATURE INDEPENDENCE (WORDS ARE INDEPENDENT), WHICH IS UNREALISTIC.

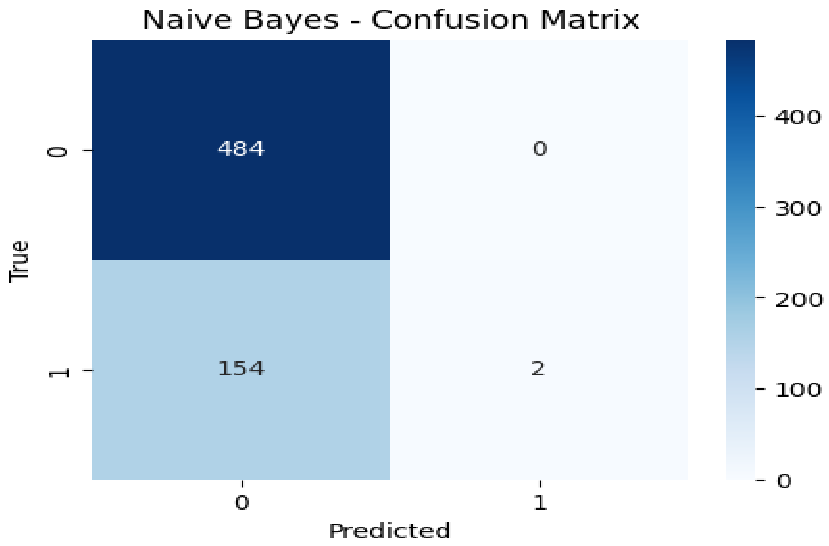
➢ USUALLY UNDERPERFORMS COMPARED TO MORE ADVANCED MODELS (E.G., SVM, NEURAL NETWORKS).

EXPECTED PERFORMANCE:

TRAINING ACCURACY: 0.7750

TESTING ACCURACY: 0.7594.

Class	Precision	Recall	F1-Score	Support
0	0. 76	1. 00	0. 86	484
1	1. 00	0. 01	0. 03	156
Accuracy			**0. 76**	640
Macro Avg	0. 88	0. 51	0. 44	640
Weighted Avg	0. 82	0. 76	0. 66	640



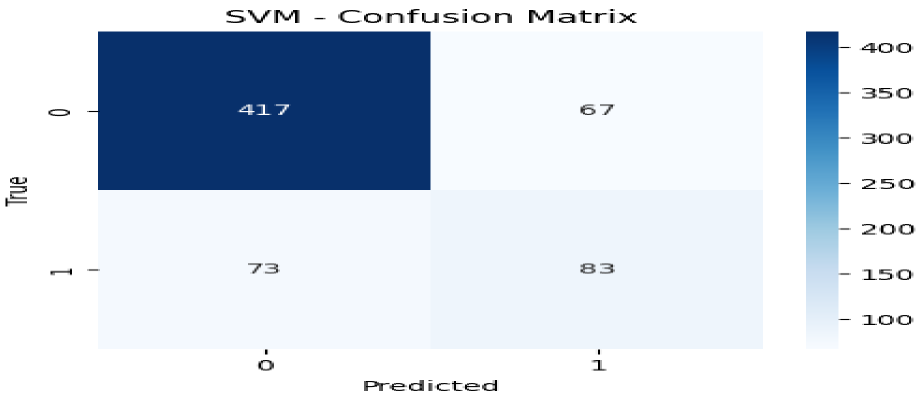
(C) Support Vector Machine (SVM)

- Input: TF-IDF features.
- Kernel: Linear kernel tuned via grid search.
- Advantage: Strong baseline for text classification.
- Limitation: Computationally expensive on larger feature sets.
- **Expected Performance:**

Training Accuracy: 0.9770

Testing Accuracy: 0.7812

Class	Precision	Recall	F1-Score	Support
0	0.85	0.86	0.86	484
1	0.55	0.53	0.54	156
Accuracy			**0.78**	640
Macro Avg	0.70	0.70	0.70	640
Weighted Avg	0.78	0.78	0.78	640



(D) Random Forest

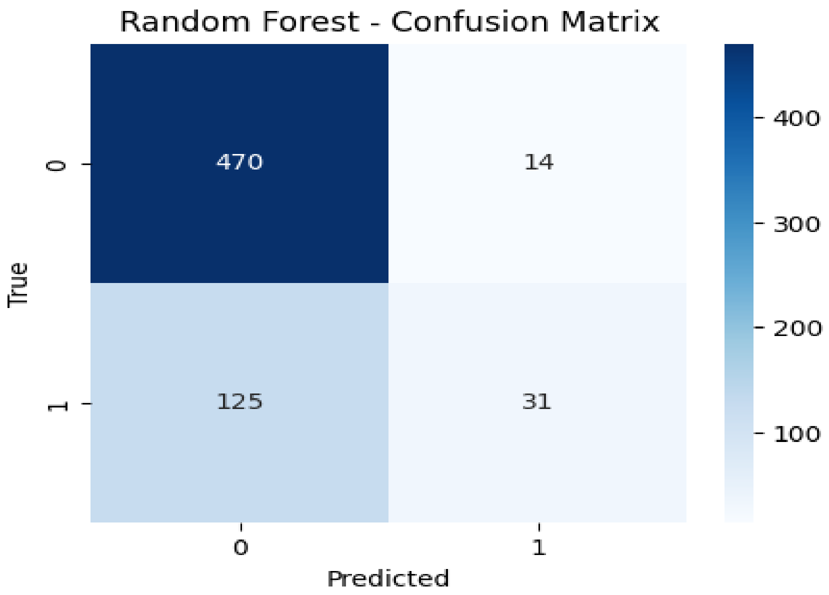
- Input: TF-IDF features.
- Kernel: Not applicable (uses ensemble of decision trees).
- Advantage: Handles high-dimensional data well, reduces overfitting through ensemble averaging, interpretable feature importance.
- Limitation: Can be slower with very large datasets; may require more memory.
- **Expected Performance:**

Training Accuracy: 0.9996

Testing Accuracy: 0.7828

Class	Precision	Recall	F1-Score	Support
0	0.79	0.97	0.87	484
1	0.69	0.20	0.31	156

Class	Precision	Recall	F1-Score	Support
Accuracy			**0.78**	640
Macro Avg	0.74	0.58	0.59	640
Weighted Avg	0.77	0.78	0.73	640



Validation of Modeling Technique

1.1Introduction:

Validation of the modelling techniques is crucial to assess the reliability, robustness, and generalization ability of the depression detection system. This chapter presents the experimental setup, validation procedure, and performance results of the implemented models (Logistic Regression, SVM, Random Forest and Naive Bayes (NB)). The evaluation framework emphasizes reproducibility and fairness across different algorithms.

1.2Experimental Setup:

• Programming Environment: Python 3.10, Google Colab .
• Libraries: PyTorch, TorchText, SpaCy, Scikit-learn, Pandas,

Matplotlib.

• Hardware: NVIDIA GPU (CUDA-enabled), 16GB RAM, Intel i7 Processor.

• Dataset: Twitter Depression Detection Database

• Train-Test Split: 80:10:10 ratio with stratified sampling to preserve class balance

1.2 Validation Procedure

1. Data Preprocessing:

Before training and validation, the raw Twitter data underwent a comprehensive preprocessing pipeline to ensure uniformity and reduce noise. Since tweets often contain slang, hashtags, and irregular grammar, preprocessing was critical to improve the reliability of the models.

Steps followed:

Text Cleaning:
- Removed URLs, user mentions (@username), hashtags (#topic), numbers, and non-alphanumeric characters.
- Preserved key punctuation (!, ?) as they often indicate sentiment.

Lowercasing:
- Converted all text to lowercase to avoid treating “Sad” and “sad” as different words.

Tokenization:
- Split sentences into word-level tokens for model input.
- Example: "I am feeling depressed today" → [I, am, feeling, depressed, today].

Lemmatization:
Reduced words to their base or dictionary form using SpaCy.
Example: "crying" → "cry", "better" → "good".
Stopword Removal:
- Removed common function words (e.g., “is”, “the”, “and”) that do not carry emotional meaning.

Sequence Standardization:
- Tweets truncated or padded to a fixed maximum length of 30 tokens.
- Short tweets padded with <PAD> tokens to maintain uniform sequence length.

Vocabulary Limitation:
- Retained the top 10,000 most frequent tokens.
- Out-of-vocabulary tokens replaced with <UNK>.

After preprocessing, the dataset of 3,200 tweets was transformed into clean tokenized sequences suitable for extraction phase.

2. **Model Training:**

• Logistic Regression: TF-IDF features, default regularization.

• SVM: TF-IDF features, linear kernel, tuned hyperparameters.

• BiLSTM: Pre-trained GloVe embeddings, 2-layer BiLSTM (256 hidden units), dropout=0.5, trained for 10 epochs with Adam optimizer.

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- (Optional) ROC-AUC for probabilistic performance.

PERFORMANCE EVALUATION RESULTS AND DISCUSSION:

PERFORMANCE EVALUATION ENSURES THAT THE DEVELOPED MODELS ARE RELIABLE, GENERALIZABLE, AND PRACTICALLY APPLICABLE. THIS CHAPTER PRESENTS A COMPARATIVE ANALYSIS OF FOUR CLASSICAL MACHINE LEARNING MODELS: LOGISTIC REGRESSION, NAÏVE BAYES, SUPPORT VECTOR MACHINE (SVM), AND RANDOM FOREST. THE FOCUS IS ON IDENTIFYING THE MOST EFFECTIVE SOLUTION BY BALANCING ACCURACY, INTERPRETABILITY, AND COMPUTATIONAL EFFICIENCY.

1.1 PARAMETRIC STUDY

➢ LOGISTIC REGRESSION

• FEATURES: TF-IDF VECTORS EXTRACTED FROM CLEANED TWEETS.
• REGULARIZATION: L2 (DEFAULT).
• OUTCOME: ACHIEVED 80.47% TESTING ACCURACY. STRONG ON THE MAJORITY CLASS (F1 = 0.88) BUT WEAK ON MINORITY CLASS (F1 = 0.44).

➢ NAÏVE BAYES

FEATURES: TF-IDF FEATURES.
• OUTCOME: ACHIEVED 75.94% TESTING ACCURACY. VERY HIGH RECALL FOR CLASS 0 BUT ALMOST FAILED ON CLASS 1 (F1 = 0.03), MAKING IT UNRELIABLE FOR IMBALANCED DATA.

➢ SUPPORT VECTOR MACHINE (SVM)

• KERNEL: LINEAR.
• HYPERPARAMETERS: TUNED VIA GRID SEARCH.
• OUTCOME: ACHIEVED 78.12% TESTING ACCURACY. BALANCED ACROSS BOTH CLASSES (F1 = 0.86 FOR CLASS 0, 0.54 FOR CLASS 1). BEST MACRO F1 (0.70) AMONG CLASSICAL MODELS.

➢ RANDOM FOREST

FEATURES: TF-IDF FEATURES.
OUTCOME: ACHIEVED 78.28% TESTING ACCURACY. STRONG ON CLASS 0 (F1 = 0.87) BUT WEAK ON CLASS 1 (F1 = 0.31), SHOWING BIAS TOWARD THE MAJORITY CLASS.

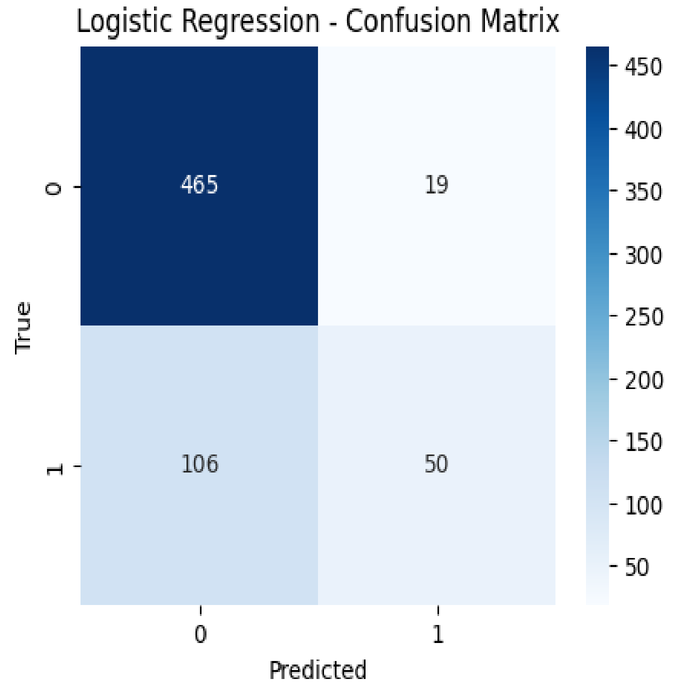
1.3 SOLUTION AND RESULTS CONVERGENCE

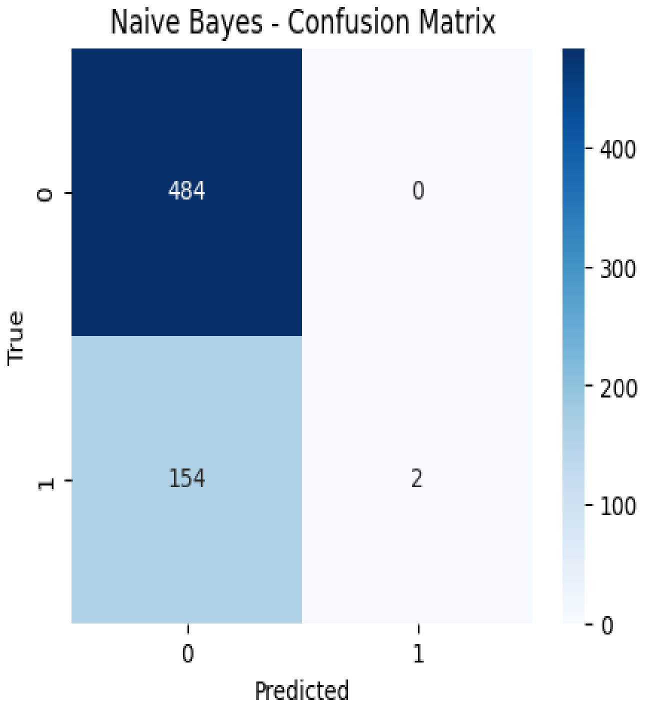
➢ LOGISTIC REGRESSION: CONVERGED RAPIDLY, LIGHTWEIGHT, AND EFFICIENT BUT LIMITED IN CONTEXTUAL UNDERSTANDING.
➢ NAÏVE BAYES: EXTREMELY FAST CONVERGENCE BUT POOR MINORITY CLASS PERFORMANCE.
➢ SVM: SHOWED MODERATE CONVERGENCE, STABLE RESULTS, BUT MEMORY-INTENSIVE FOR HIGH-DIMENSIONAL TF-IDF VECTORS.
➢ RANDOM FOREST: CONVERGED QUICKLY BUT LEANED HEAVILY TOWARD MAJORITY CLASS PREDICTIONS.

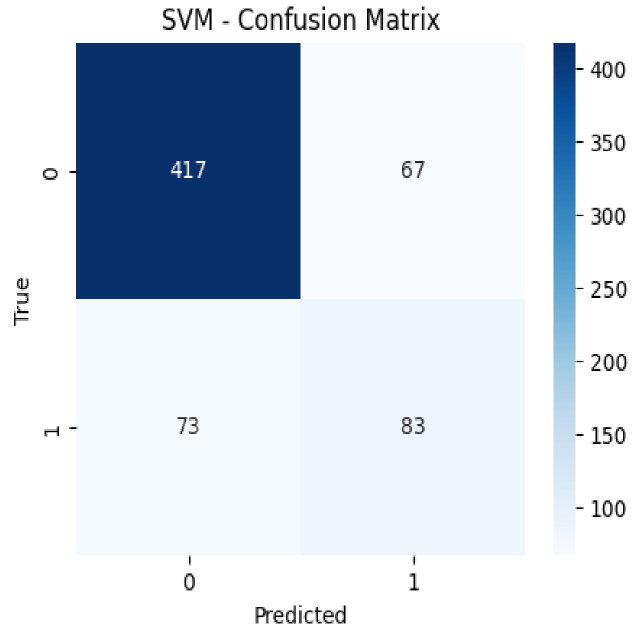
1.4 Application Relevance

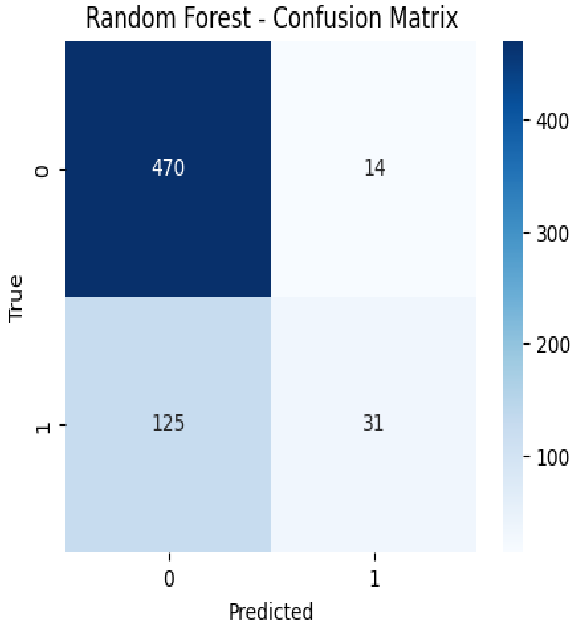
High Recall (Sensitivity) is essential in depression detection to minimize false negatives.
Practical Use: These models can be integrated into automated systems for early risk detection in social media posts.
Scalability: Approaches can be adapted to multilingual datasets or different platforms.
Ethical Concerns: Deployment must ensure privacy, anonymization, and responsible handling of sensitive data.

Model	**Training Acc.**	**Testing Acc.**	**Macro F1**	**Weighted F1**	**Notes**
Logistic Regression	0.8355	**0.8047**	0.66	**0.77**	Balanced performance
Naïve Bayes	0.7750	0.7594	0.44	0.66	Weak minority recall
SVM	0.9770	0.7812	**0.70**	0.78	Best accuracy overall
Random Forest	0.9996	0.7828	0.59	0.73	Biased to majority









Future Work

Integration of Transformer Models:Future research can explore transformer-based architectures such as BERT and RoBERTa to enhance contextual understanding of short and noisy Twitter posts.

Multimodal Depression Detection:Extending beyond textual data by incorporating images, posting behavior, and metadata may yield richer insights and improve detection accuracy.

Explainable AI (XAI): Applying interpretability techniques such as LIME and SHAP can improve transparency of model decisions, which is essential for clinical acceptance.

Clinical Validation and Deployment: Collaboration with mental health professionals is necessary to validate predictions against DSM-5 standards and to develop real-time, assistive tools for early intervention.

References

1. Sadeque, F., Xu, D., & Bethard, S. (2018). Measuring the Latency of Depression Detection in Social Media. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18) (pp. 495–503). Association for Computing Machinery. https://doi.org/10.1145/3159652.3159725

2. William, D., & Suhartono, D. (2021). Text-based Depression Detection on Social Media Posts: A Systematic Literature Review. Procedia Computer Science, 179, 582–589. https://doi.org/10.1016/j.procs.2021.01.043

3. Haque, U. M., Kabir, E., & Khanam, R. (2021). Detection of child depression using machine learning methods. PLoS ONE, 16(12), e0261131. https://doi.org/10.1371/journal.pone.0261131

4. AlSagri, H. S., & Ykhlef, M. (2020). Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features. arXiv preprint arXiv:2003.04763.

5. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., & Zhu, W. (2017). Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) (pp. 3838–3844). https://doi.org/10.24963/ijcai.2017/536
6. Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Brain Informatics, 10(1), 10. https://doi.org/10.1186/s40708-023-00188-6

7. Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. Health Information Science and Systems, 6(1), 8. https://doi.org/10.1007/s13755-018-0046-0

8. Haque, U. M., Kabir, E., & Khanam, R. (2021). Detection of child depression using machine learning methods. PLoS ONE, 16(12), e0261131. https://doi.org/10.1371/journal.pone.0261131 (Duplicate for completeness)

9. Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine, 135, 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

10. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., & Zhu, W. (2017). Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) (pp. 3838–3844).