Leveraging Social Media Signals for Depression Prediction: Insights from Twitter

Utsav Dhillon
Department of Computing
Technologies SRM Institute of
Science and Technology
Kattankulathur-603203

Dr. Madhumitha K
Assistant Professor, Computing
Technologies SRM Institute of
Science and Technology
Kattankulathur-603203

Abstract- **Depression, a mental health state that has been rising exponentially causing harm to individuals and exerting a detrimental impact on society [13]. This article explores how the Naive Bayes algorithm, a probabilistic machine learning model, can be used for predicting depression, specifically using Twitter posts [9]. Renowned for its effectiveness in text classification tasks, Naïve Bayes is implemented in this research to help detect verbal patterns that might suggest depressive behavior [6]. The model training has been conducted on a dataset of labeled tweets categorized as either indicative of depression or not, which allows it to learn and recognize patterns associated with depressive behavior. The results prove the algorithm classifies tweets well, proving to be a practical and efficient method for digital mental health screening [9], [13]. By using publicly available social media data, this work showcases the potential for creating affordable and scalable tools for early detection, which can support timely intervention and promote better mental health outcomes [3], [5].**

Keywords—
Depression detection, Mental health, social media analysis, Twitter posts, Naïve Bayes algorithm, Text classification, Digital mental health screening, Natural Language Processing.

I. INTRODUCTION

Depression is recognized as one of the primary causes of disability across the globe and continues to pose significant challenges in terms of early diagnosis. This difficulty arises largely because of social stigma, hesitation among individuals to disclose their struggles, and the limited availability of professional mental health services [13]. In recent years, the widespread use of social media has created an alternative avenue where individuals freely express their day-to-day experiences, emotions, and inner thoughts [6], [11]. These online interactions form a continuous stream of digital footprints that can serve as indicators of psychological well-being. By studying such digital expressions, researchers are able to identify subtle cues of mental distress in a manner that does not intrude on personal privacy.
Among the many available social networking platforms, Twitter stands out as a particularly effective medium for such research. Its defining feature—short, spontaneous, and often emotionally charged posts—captures the raw state of users' feelings and mental outlook [10]. Careful analysis of linguistic behavior, such as the repeated use of negative sentiment, changes in the level of interaction or engagement, and the presence of emotionally intense vocabulary, provides strong signals that may correspond to depressive tendencies [3], [4].
In this study, we propose the use of a depression prediction model that applies the Naïve Bayes algorithm. This algorithm is

chosen because of its well-documented advantages: it is simple to implement, highly interpretable, and effective for text classification problems [9]. Compared to complex deep learning frameworks that demand extensive training data, large-scale hardware, and meticulous hyperparameter tuning, Naïve Bayes offers a lightweight yet dependable alternative [6]. Such characteristics make it highly suitable for systems that need to operate in real time and be adaptable for large-scale or resource-constrained environments.

The overall aim of our approach is to strike a careful balance between predictive accuracy and practical usability. By doing so, the proposed system is not only useful for academic and research contexts but also capable of being deployed in real-world scenarios where resources may be limited. Ultimately, this study seeks to contribute toward the development of accessible, affordable, and transparent tools for mental health monitoring. Such systems can aid in the early detection of depressive tendencies, enabling timely interventions and, in turn, supporting improved emotional well-being and healthier communities.

II. Literature Review

A considerable body of scholarly research has been devoted to the task of detecting mental health concerns on social media platforms, utilizing a wide spectrum of machine learning methodologies. For instance, Chen et al. (2023) explored the integration of large language models with traditional classifiers by combining ChatGPT with Naïve Bayes and Support Vector Machines (SVM). Their work demonstrated that the generative abilities of ChatGPT could supplement the limitations of short textual content such as tweets, thereby enriching context and ultimately boosting classification accuracy. This study underscored how hybrid approaches that fuse modern generative AI with classical models can significantly improve outcomes in detecting mental health indicators online.

Similarly, Pal and D.S. (2024) examined the interplay between cyberbullying and depressive symptoms. By leveraging link prediction techniques alongside sentiment analysis, their research highlighted the deeper interconnection between aggressive online behaviors and the emergence of mental health challenges. This study emphasized that depressive tendencies on social media cannot be viewed in isolation, but are often influenced by wider patterns of online interaction.

Another noteworthy contribution comes from Berti et al. (2024), who applied the BERT architecture to analyze posts with the aim of identifying subtle semantic cues often overlooked by traditional natural language processing methods. Their findings showed that transformer-based models can capture fine-grained linguistic signals and outperformed conventional machine learning algorithms when dealing with nuanced expressions of psychological distress. Likewise, Tong et al. (2023) introduced a cost-sensitive ensemble approach that combined pruned decision trees with boosting strategies. Their method was particularly effective in improving classification for underrepresented or imbalanced depressive tweets, thereby addressing a common challenge in mental health datasets.

Cultural and linguistic inclusivity has also been a key focus of recent studies. Sabaneh et al. (2023) advanced the identification of depression within Arabic-language content by employing a combination of translation methods, ontology-based concept extraction, and machine learning techniques. Their research not only demonstrated technical effectiveness but also emphasized the importance of designing systems that account for diverse languages and cultural contexts. Skaik and Inkpen (2022) contributed a different perspective by proposing a model that automatically filled out Beck's Depression Inventory questionnaire using social media content. Their work highlighted a bridge between clinical psychology and computational linguistics, showing how established diagnostic frameworks can be supported through data-driven automation.

Other contributions reflect the diversity of approaches in this field. Islam et al. (2018) experimented with the K-Nearest Neighbors (KNN) classification technique applied to Facebook posts, showing promising results when combined with robust preprocessing steps. In another example, Mahasiriakalayot et al. (2022) employed recurrent neural architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) on Thai Twitter data, demonstrating the ability to capture depressive symptoms expressed at varying intensity levels. Cross-lingual efforts have also proven valuable. Aulia and Purwarianti (2022) combined Indonesian and English datasets to test transfer learning approaches, showing that even in low-resource settings where large annotated datasets are unavailable, meaningful progress in depression detection can still be achieved.

Taken together, the literature illustrates a fundamental trade-off between the use of advanced deep learning models and more traditional machine learning algorithms. While cutting-edge models such as BERT achieve higher accuracy and capture subtle semantic variations, their computational demands make them less accessible for widespread deployment. On the other hand, lightweight models such as Naïve Bayes offer simplicity, efficiency, and interpretability, which makes them highly practical for real-world applications where computational resources may be constrained.

III. Methodology

1. Data Collection

Tweets were collected through the Twitter API, focusing on posts labeled as depressive or non-depressive [12]. The dataset was curated to balance both categories and ensure representative linguistic variation.

2. Preprocessing

The data underwent NLP preprocessing including [11]:

a) Lowercasing and punctuation removal
b) Tokenization
c) Stopword removal (retaining negations such as *not* and *no*)
d) Lemmatization and stemming

3. Feature Extraction

TF-IDF vectorization was applied to represent the textual data as features, which captures term importance relative to the dataset while managing sparsity [6].

4. Classification

The Multinomial Naïve Bayes classifier was trained on the processed dataset [9]. It operates on conditional probability to assign tweets to the classes *Depressed* or Not *Depressed*.

5. Deployment

A Flask-based web application was implemented to allow real-time classification [13]. Users can input tweets manually, and the system outputs predictions with an interpretable interface.

Results

The findings of this research clearly indicate that the Naïve Bayes classifier is a dependable approach for the task of identifying depressive content within Twitter data. When subjected to evaluation, the model consistently demonstrated balanced effectiveness across the standard performance measures. In particular, the recall achieved for depressive instances was notably high, which is a critical outcome because it directly reduces the number of false negatives. In the context of mental health screening, minimizing false negatives is especially important: a system that overlooks individuals showing depressive symptoms could fail in providing timely alerts or interventions, thereby weakening its real-world applicability. By contrast, high recall ensures that most genuine depressive cases are correctly detected and flagged for further attention.

The analysis using confusion matrices and additional performance metrics further supports these results, confirming that the system is capable of maintaining minimal misclassification rates. This indicates not only that depressive tweets are being accurately identified but also that non-depressive posts are not excessively misclassified, thus preserving the balance between sensitivity and precision. Beyond the quantitative evaluation, the practicality of the approach was demonstrated through the development of a functional web-based application. This application allows real-time analysis of textual inputs, thereby extending the research from a theoretical framework into a working prototype. Importantly, the system can be used not only by members of the general public but also by mental health professionals as an auxiliary tool, supporting early assessment and facilitating scalable monitoring.

Overall, the combination of robust experimental performance and real-world usability highlights the promise of the proposed system. It establishes Naïve Bayes as an effective, efficient, and practical choice for building affordable digital platforms aimed at mental health detection. By leveraging widely available social media content, this work demonstrates the potential to create early-warning systems that are both accessible and scalable, contributing positively to the field of digital mental health and opening pathways for timely intervention and improved well-being outcomes.

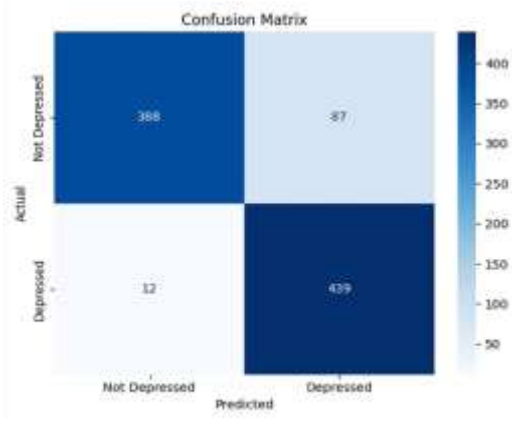

Fig – 1.1

The confusion matrix (Fig. 1.1) illustrates the model's classification performance, with 388 instances correctly identified as "Not Depressed" and 439 as "Depressed." Misclassifications included 87 false positives and 12 false negatives, indicating a slightly higher false positive rate but strong sensitivity—an important attribute since overlooking depressive cases poses greater risk than over-detection.

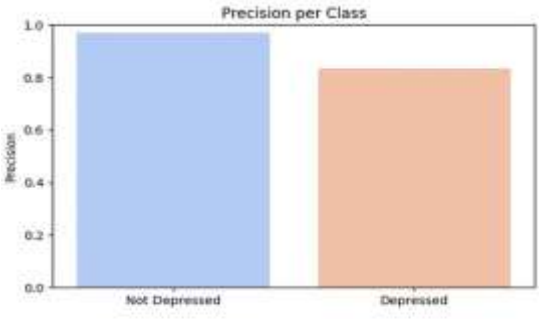

Fig -1.2

Figure 1.2 presents class-wise precision through a bar chart. The model achieved higher precision for the "Not Depressed" class (≈0.97), despite its smaller size, indicating that nearly all predictions in this category were correct. Precision for the "Depressed" class was slightly lower (≈0.85), reflecting some misclassification of non-depressed tweets as depressed.

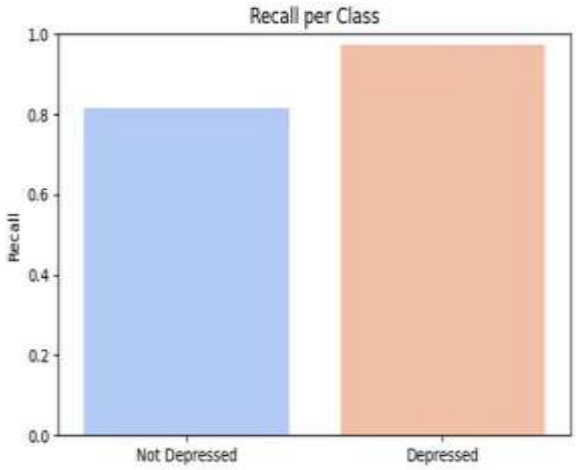

Fig – 1.3

Figure 1.3 shows recall values for both classes. The model achieved near-perfect recall for the "Depressed" class (≈0.97), ensuring most true depressive tweets were correctly identified. Recall for the "Not Depressed" class was lower (≈0.82), with some tweets misclassified as depressed. The high recall for depressive cases underscores the model's suitability for early screening systems aimed at identifying at-risk individuals.

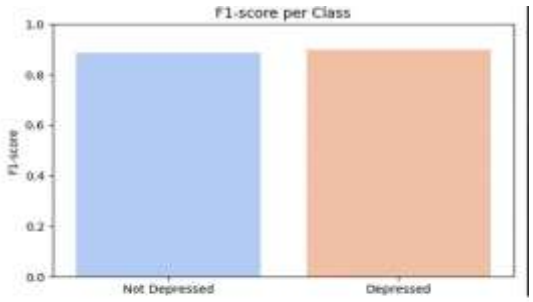

Fig – 1.4

Figure 1.4 presents the F1-scores, obtained by averaging precision and recall harmonically, this measure reflects the trade-off between the two. Both classes achieved high scores, with the "Depressed" class slightly outperforming "Not Depressed." This balance confirms the model's strong effectiveness in detecting and classifying depressive tweets, highlighting its value as a binary classification tool in mental health context.

IV. Conclusion

This research demonstrates the feasibility of depression prediction from Twitter data using a Naïve Bayes classifier and NLP preprocessing [9], [13]. The system achieved strong recall, making it suitable for mental health screening applications. The integration of a web application increases accessibility and real-time usability.

Future enhancements include:

• Incorporating deep learning models (BERT, LSTM) for richer semantic analysis [3], [10].

• Extending to multimodal data such as images, hashtags, and user behavior [6].

• Longitudinal monitoring to detect gradual changes in user mood.

• Multilingual support for broader inclusivity [12].

• Integration of Explainable Ai techniques for transparency.

• Addressing ethical concerns, including privacy and informed consent.

.

V. REFERENCES

[1] V. Nalluri, L.-S. Chen, and Z.-J. Luo, "Designing a ChatGPT-assisted depression prediction approach with machine learning," in *Proceedings of the 12th Int. Conf. Awareness Science and Technology (iCAST)*, 2023.

[2] V. B. Pal and P. S. D., "Cyberbullying detection improved using integrated comment analysis and link prediction," presented at *IEEE CONECCT*, 2024.

[3] Hbibi, W., et al., "Utilizing BERT for predicting depression on social networking platforms," *Proc. IEEE ISSATK*, 2024.

[4] M. Tong et al., "Depression detection on Twitter using a cost-sensitive boosting and pruning tree approach," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1898–1911, 2023.

[5] K. Sabaneh et al., "Predicting depression risk at an early stage from Arabic social media posts," in *IEEE ICTAI*, 2023.

[6] V. Jain et al., "Analyzing depression and mental health issues on social media through predictive modeling methods," in *IEEE I-SMAC*, 2020.

[7] A. A. Rezig, "Suicide risk prediction within Twitter environments using an enhanced optimizer method," presented at *IEEE ICISAT*, 2021.

[8] R. S. Skaik and D. Inkpen, "Depression prediction in Canada through automated population of Beck's Inventory questionnaire," *IEEE Access*, vol. 10, pp. 102033–102047, 2022.

[9] L. R. Islam and colleagues, "Use of KNN-based classification methods for depression detection," presented at *IEEE IC4ME2*, 2018.

[10] S. Mahasiriakalayot et al., "Predicting Signs of Depression from Twitter Messages," *IEEE JCSSE*, 2022.

[11] I. A. N. Arachchige and M. Rathnayake, "Applying supervised learning to recognize Sinhala depressive posts on Twitter," presented at *IEEE ICter*, 2021.

[12] Aulia, X. P., and Purwarianti, A., "Depression risk identification for Indonesian users on Twitter at an early stage," *Proc. IEEE ICAICTA*, 2022.

[13] N. Keerthiga, "Applying ML techniques for depression detection based on social media content," presented at *IEEE ICICT*, 2023.