Depression Detection on Twitter using Machine Learning Approaches

Priyanshu[1], Sarthak Mittal[2], Sanchit Vasdev[3]
[1,2,3] *Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India*

***Abstract*— Social media networks like Facebook, Twitter, and Instagram have reshaped how individuals interact and shape their digital identities. While these platforms offer significant benefits for communication and fostering communities, a correlation has been observed between their extensive use and a rise in depression cases. The early identification of such mental health challenges is critical because timely intervention can dramatically enhance a person's well-being. This research investigates the application of machine learning methods to detect indicators of depression among Twitter users through an analysis of their tweets and network activity. To capture significant linguistic cues, the data was pre-processed using a variety of natural language processing techniques, such as tokenization, the removal of stop-words, and word embedding. On a dataset comprising both random and depressive tweets, classifiers including TF-IDF and Long Short-Term Memory (LSTM) were trained and assessed, showing proficient performance in differentiating between users. The results indicate that more comprehensive feature sets enhance detection precision and that specific language patterns—especially those conveying sadness, anxiety, or anger—are powerful predictors of depression. Beyond the technical outcomes, this research emphasizes how AI-powered tools can supplement conventional mental health services by providing scalable, readily available, and proactive monitoring solutions. With further development, these methods could address existing gaps in mental health care, facilitate timely support for individuals, and foster healthier online environments.**

I. INTRODUCTION

Mental health is a fundamental component of overall wellness, and the prompt identification of conditions like depression can significantly enhance recovery outcomes and life quality. Traditionally, mental health services have relied on face-to-face therapy and consultations, which, despite their effectiveness, remain inaccessible for many people. Recent technological advancements present new avenues for making mental healthcare more widely available, particularly through tools capable of identifying early warning signs within daily digital communications.

Over the last few years, the application of artificial intelligence for detecting emotions and depression has garnered significant interest. Researchers have demonstrated that by applying machine learning to analyze text, audio, or video, it is possible to identify subtle emotional signals that might be missed by conventional self-reporting or clinical assessments. Such technologies can facilitate the earlier identification of at-risk individuals, thereby improving their access to timely assistance. While this study is primarily centered on mental health, emotion detection is also valuable in fields like customer experience analysis and product feedback evaluation.

Nevertheless, accurately detecting emotions from language is a complex task, as it involves differentiating between a broad spectrum of expressions—like joy, fear, anger, and sadness—and interpreting their contextual meaning. This project concentrates on the short-form text from Twitter, a leading social media platform. The objective is to analyze user posts to classify emotional states and recognize signs associated with depression. In contrast to basic sentiment analysis, which only categorizes text as positive or negative, emotional analysis delves deeper into the subtleties of human expression. This approach enables a more thorough understanding of a user's mental state and lays the groundwork for creating tools that can aid in the early detection and management of depression.

II. METHODOLOGY

Natural Language Processing (NLP) offers a promising avenue for identifying emotions and mental health issues from text-based communications. In this research, we engineered a machine learning model designed to recognize emotions and detect indications of depression from user tweets. The model was developed using two distinct datasets from Kaggle: one composed of general, random tweets and another containing tweets identified as depressive. Subsequent training and evaluation confirmed the model's high potential for supporting mental health surveillance by providing an accessible means to monitor emotional states.

To achieve this goal, two separate models for emotion detection in text were built and evaluated. Both were trained using labeled data and assessed on their capacity to differentiate between depressive and non-depressive content. The outcomes confirmed their effectiveness in identifying expressions linked to depression, underscoring their potential for facilitating early intervention and offering valuable insights for both healthcare professionals and individuals

Data Preprocessing

To prepare the raw tweet data for analysis, a series of pre-processing steps were undertaken:

a. Removing HTML tags and URLs
   Tweets frequently include links and tags that contain minimal sentimental information. These elements were stripped out using regular expressions to clean the text.

b. Conversion to lowercase
   To ensure dataset consistency, all text was converted to lowercase with Python's lower() function, which helps standardize the vocabulary.

c. Tokenization
   This process involves breaking down text into smaller, individual units like words or phrases. It facilitates the analysis of linguistic structure and the extraction of valuable information. For instance, the phrase "The quick brown fox jumped over the lazy dog" is tokenized into individual words.

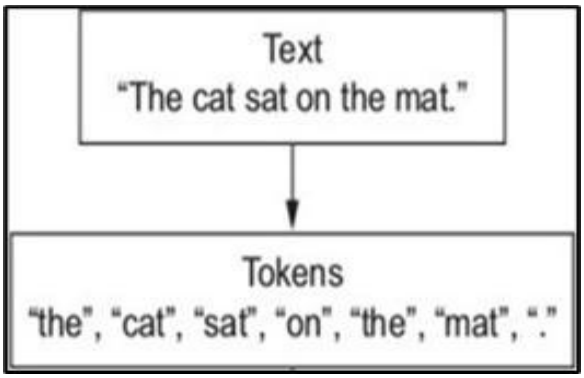

Figure 1 Visual Representation of Tokenization

d. Removing Punctuation and Special Symbols

   Punctuation marks and special characters (e.g., &, „ ;) were eliminated, with the exception of certain emoticons. These symbols typically do not add meaningful content and can create noise in the data, so their removal enhances the text's clarity for analysis.

e. Stop Word Removal

   Common words like "is," "the," and "a" provide little semantic meaning and are known as stop words. Since they lack sentimental weight, they were removed using the NLTK library. This filtration reduces textual redundancy and helps the model concentrate on more emotionally relevant words.

f. Word2Vec

   The Word2Vec embedding technique was used to create numerical representations of words. This method employs neural networks to generate vector representations that capture semantic and syntactic word relationships. As a result, words with similar meanings, like "sad" and "unhappy," are positioned closer in the vector space, making this technique highly suitable for this study.

g. Stemming and Lemmatization

Tweets often contain various forms of the same word. To address this, stemming and lemmatization were used to reduce words to their root forms. For example, "cats" is reduced to "cat". While stemming removes suffixes via simple rules, lemmatization is a more sophisticated process that converts a word to its base dictionary form. These techniques standardize the vocabulary, which improves the model's ability to generalize from textual patterns.

Through these steps, the data was cleaned and transformed into a format suitable for further analysis. Natural Language Processing (NLP) techniques have become a promising way to detect emotions and mental health conditions from text-based interactions. In this study, we designed a machine learning model to identify emotions and detect signs of depression in users' tweets. Two datasets from Kaggle were used: one containing random tweets and another containing only depressive tweets. Training and evaluation on these datasets showed that the model has strong potential to support mental health monitoring by offering an accessible method to track emotional well-being.

To meet this objective, we developed and tested two models for emotion detection in text. Both models were trained on labelled data and evaluated for their ability to distinguish between depressive and non-depressive content. The results demonstrated their effectiveness in identifying depression-related expressions, highlighting their usefulness for early intervention and for providing insights to both professionals and affected individuals.

Dataset Description

This research utilized two datasets sourced from Kaggle. The first was a collection of random tweets, while the second was comprised of tweets specifically identified as depressive. Merging these datasets created a balanced basis for model training and validation. Following the combination, the data was cleaned to remove duplicate entries and irrelevant content, ensuring its quality. The final dataset was partitioned, with 80% allocated for training the models and the remaining 20% reserved for testing and evaluation. This division allows for robust training while ensuring an objective assessment on previously unseen data.

Feature Extraction

Once the data was pre-processed, features were extracted to enable the models to learn from the text. Two main approaches were used:

- **TF-IDF Representation**: Term Frequency–Inverse Document Frequency (TF-IDF) was used to transform text into numerical vectors. This method highlights informative words by diminishing the significance of common terms across the dataset
- **Word2Vec Embedding**: Tweets were also converted into dense vector representations using Word2Vec. This technique excels at capturing the semantic and syntactic connections between words, allowing the model to better grasp the contextual meaning of expressions associated with depression.

Using both sparse (TF-IDF) and dense (Word2Vec) representations enabled a broader comparison of feature extraction methods.

III. ALGORITHMS

Machine learning methods can be categorized based on the task and data type, primarily as supervised or unsupervised learning. In supervised learning, a model learns from data that is already labeled, enabling it to identify patterns between input features and corresponding outputs. Conversely, unsupervised learning involves identifying hidden patterns in data without predefined labels. This study employed a **supervised learning** approach, as our dataset consisted of tweets pre-classified as either depressive or non-depressive. The following models and algorithms were utilized:

1. **TF-IDF with Classifier**

For this method, tweets were initially transformed into numerical vectors via Term Frequency–Inverse Document Frequency (TF-IDF). Subsequently, a classifier was trained on these vectors to differentiate

between depressive and non-depressive content. The TF-IDF technique is effective at accentuating key terms by assigning lower weights to frequent words and higher weights to more distinctive ones.

2. **Long Short-Term Memory (LSTM)**

As a specialized type of recurrent neural network (RNN), LSTM is particularly effective for handling sequential data like text. Unlike standard feed-forward networks, LSTMs are capable of recognizing long-range dependencies and contextual links between words in a sentence. By utilizing Word2Vec embeddings as its input, the LSTM model could effectively learn linguistic patterns and identify subtle textual indicators of depression.

Employing these two distinct methods allowed for a comparative analysis between a traditional feature-based model (TF-IDF) and a deep learning model (LSTM), providing a fuller understanding of their respective effectiveness in depression detection.

IV. EVALUATION MATRIX

There are several metrics which are considered to evaluate the effectiveness of our proposed approach: true positive rate, false positive rate, true negative rate, false negative rate, F1 score, accuracy, precision, and recall. These metrics are commonly used to assess the performance of spam detection systems. Here is a brief explanation of each metric.

a. True Positives (TP): It represents the number of spam messages correctly classified by the machine learning algorithm.
b. True Negatives (TN): It indicates the number of non-spam (ham) messages accurately classified as ham by the machine learning algorithm.
c. False Positives (FP): It denotes the number of ham messages wrongly classified as spam by the machine learning algorithm (Type 1 Error)
d. False Negatives (FN): It represents the number of spam messages incorrectly classified as ham by the machine learning algorithm (Type 2 Error)
e. Accuracy: This is the proportion of total correct predictions (both true positives and true negatives) among the total number of cases examined. It can be calculated as

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Population}$$

f. Precision: It indicates the percentage of messages classified as spam that are actually spam. Precision measures the exact correctness and can be calculated as

$$Precision = \frac{True\ Positives}{Predicted\ Positive\ (TP + FP)}$$

g. Recall (Sensitivity): It represents the percentage of actual spam messages that are correctly classified as spam. Recall measures the completeness and can be calculated as:

$$Recall = \frac{True\ Positives}{Actual\ Positive(TP + FN)}$$

h. F1-Score: It is defined as the harmonic mean of Precision and Recall, providing a balanced measure of the algorithm's performance. It can be calculated as.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \pm Recall}$$

i. Receiver Operating Characteristics (ROC) Area: It refers to the area under the curve plotted between True Positive Rate and False Positive Rate for different threshold values

These metrics collectively provide valuable insights into the performance and accuracy of the Depression Detection System

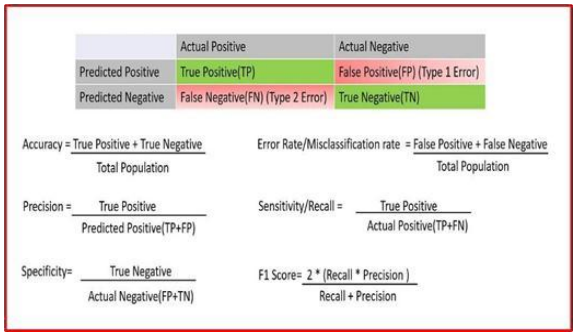

Figure 2: Confusion Matrix and Evaluation Metrics

V. RESULTS AND DISCUSSION

The experiments were conducted using the dataset described in Section II. The dataset was split into

training and testing subsets, with 80% used for training and 20% for testing.

Two approaches were implemented for comparison. The first used **TF-IDF features** in combination with a machine learning classifier, while the second applied **Word2Vec embedding with an LSTM network**. Each model was evaluated using standard classification metrics, namely **Accuracy, Precision, Recall, and F1-score**. These metrics provide a comprehensive understanding of performance by considering not only overall correctness but also the balance between correctly identified depressive tweets and false classifications.

The results are summarized in Table 1. As shown, the TF-IDF based model achieved higher accuracy and F1-score compared to the LSTM model. While LSTM was able to capture sequential dependencies in text, the TF-IDF approach proved more effective for this dataset, likely due to its ability to emphasize discriminative words.

Model	Precision	Recall	F1-Score	Accuracy
TF-IDF Classifier	0.97	0.99	0.98	0.99
LSTM	0.83	0.84	0.87	0.89

Table 1 Accuracy comparison between the models

The study's findings underscore the potential of machine learning techniques for emotion detection based on textual data. The TF-IDF classifier exhibited exceptional accuracy in predicting the emotions present in the text, displaying high precision and recall. This implies that the model was able to learn the underlying patterns and relationships in the data efficiently. However, it is important to acknowledge that the performance of classifier models can fluctuate based on the specific dataset and evaluation metrics used. Furthermore, the use of advanced machine learning techniques, such as deep learning, may lead to even better outcomes. Overall, this research highlights the potential of machine learning for emotion detection based on textual data and emphasizes the importance of further investigation in this field. The application of such techniques could have practical implications in numerous settings, including social media analysis, chat applications, and customer service.

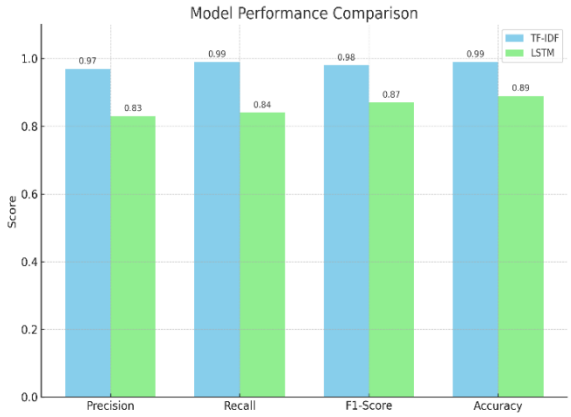

Figure 3 Performance comparison of TF-IDF and LSTM models across different evaluation metrics.

VI. CONCLUSION

This study explored the use of deep learning and machine learning techniques to detect emotional states and signs of depression in Twitter users. By examining patterns in language, we found that individuals experiencing stress or depression often use words that reflect anxiety, sadness, anger, or even suicidal thoughts. These linguistic indicators, when analysed effectively, can provide valuable insight into a user's mental state.

Two models were developed and evaluated in this work: a TF-IDF-based classifier and an LSTM deep learning model. Both models were assessed using standard evaluation metrics, including precision, recall, F1-score, and accuracy. The experimental results showed that the TF-IDF-based classifier performed better overall, achieving higher scores across most metrics compared to the LSTM model. This suggests that for short and direct text such as tweets, traditional feature-based methods can sometimes be more effective than complex deep learning models.

Beyond technical results, this research highlights the potential of automated text analysis as an early detection tool for mental health challenges. With further refinement, such approaches can be integrated into digital health platforms to provide scalable, accessible, and timely support to individuals who may be at risk. Future work can extend this study by incorporating multilingual data, larger datasets, or additional deep learning architectures to improve generalizability and accuracy.

VII. FUTURE SCOPE

While the proposed approach has shown promising results for detecting depression and emotional states from Twitter data, there are several opportunities to enhance and extend this work. One important direction is to incorporate **larger and more diverse datasets**, including multilingual tweets, to improve the model's ability to generalize across different languages, cultures, and communication styles.

Another promising area is the integration of **multimodal data**, such as images, audio, or user interaction patterns, alongside text. This can provide a more complete view of users' mental states and lead to more accurate predictions. Exploring **advanced deep learning architectures**, such as transformer-based models (e.g., BERT, RoBERTa), can also help capture deeper contextual relationships within language, potentially improving detection performance.

In addition, future research can focus on **real-time monitoring systems** that can analyze tweets continuously to identify early warning signs of stress or depression. Such systems, when developed responsibly with proper ethical guidelines and privacy safeguards, could be valuable tools for mental health organizations and support services.

Finally, incorporating this technology into **mental health support platforms** could make early detection more accessible, scalable, and proactive. By combining computational methods with professional mental health care, future systems could assist individuals more effectively and contribute to building healthier digital communities.

REFERENCE

[1] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., & Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution in IJCAI 3838-3844 (2017).

[2] Mowery, D., Bryan, C., & Conway, M. Feature studies to inform the classification of depressive symptoms from Twitter data for population health. Preprint at arXiv:1701.08229 (2017).

[3] Coppersmith, G., Dredze, M., & Harman, C. Quantifying mental health signals in Twitter in Book Quantifying mental health signals in Twitter 51-60 (2014).

[4] Stack, S. J. Mental illness and suicide. The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, andSociety. 1618-1623 (2014).

[5] Reece, A.G., et al. Forecasting the onset and course of mental illness with Twitter data. Sci. Rep. 7(1),13006(2017).

[6] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media in Discovering shifts to suicidal ideation from mental health content in social media 2098-2110 (ACM, 2016).

[7] Organization, W. H. Depression and other common mental disorders: global health estimates', 2017

[8] Yazdavar, A.H., Al-Olimat, H.S., Banerjee, T., Thirunarayan, K., & Sheth, A.P. Analyzing clinical depressive symptoms in twitter (2016).

[9] Nadeem, M. Identifying depression on Twitter. Preprint at arXiv:1607.07384 (2016).

[10] Harman, G., & Dredze, M. H. Measuring post traumatic stress disorder in Twitter. Proceedings of ICWSM,(2014).

[11] Benton, A., Mitchell, M., & Hovy, D. Multi-task learning for mental health using social media text. Preprintat arXiv:1712.03538 (2017).

[12] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing depression fromtwitter activity in Recognizing depression from twitter activity 3187-3196 (ACM, 2015).

[13] Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. Monitoring tweets for depression to detect at-risk users.CLPsych, 32, (2017).

[14] Park, G., et al. Automatic personality assessment through social media language. J. of Pers. and Soc. Psychol. 108 (6), 934 (2015).

[15] Coppersmith, G., Harman, C., & Dredze, M. Measuring post traumatic stress disorder in Twitter in Eighthinternational AAAI conference on weblogs and social media. (2014).