# Classify Depressive Tweet using Machine Learning Tools

Jingxuan Bao
University of Pennsylvania

## Abstract:

*This study aims to classify depressive tweets using various machine learning techniques, contributing to the early detection of depression and promoting mental health research. The methods examined include traditional classifiers (Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Random Forests), Long Short Term Memory (LSTM), and BERT (Bidirectional Encoder Representations from Transformers). Results indicate that BERT outperforms all other methods, achieving perfect accuracy and exceptional performance metrics. This highlights the potential of BERT in effectively detecting early signs of depression in social media data. Traditional classifiers like Support Vector Machines and Random Forests also perform competitively in this task. However, LSTM does not demonstrate a significant advantage over traditional classifiers, possibly due to the dataset's simplicity, collected using specific keywords related to depression.*

## Introduction:

Depression, a common and serious mental health condition, afflicts millions of people across the globe. An estimated 5% of the adult population grapples with this disorder, making it imperative to find effective ways to identify and address it (*Depressive Disorder (Depression)*, n.d.). In recent years, social media platforms like Twitter have emerged as a crucial medium for communication and self-expression, offering valuable insights into mental health patterns. In this context, the timely recognition and understanding of depressive symptoms pose a significant challenge for individuals seeking professional help. Machine learning tools present an accessible, non-invasive solution to this problem, detecting depressive symptoms through the analysis of text shared on social media platforms.

The integration of automated detection tools into social media platforms, such as Twitter, Facebook, Instagram, and TikTok, can raise users' awareness of their mental health and prompt them to seek help if their content indicates depressive symptoms. Furthermore, these tools can help promote mental health awareness, destigmatize conversations surrounding depression, and cultivate a supportive online community. This report delves into the application of machine learning techniques, such as supervised learning, deep learning, and pre-trained transformers, for classifying depressive tweets. It discusses the challenges and potential applications of this technology while exploring future research directions aimed at enhancing the efficacy and utility of these tools within the context of mental health monitoring and support.

## Literature Review:

The classification of depressive tweets has been explored using a variety of machine learning techniques, including traditional supervised learning methods, deep learning models, and pre-trained transformers. Early works like (Nadeem et al., n.d.) used statistical classifiers to estimate the risk of depression, achieving about 86% accuracy. (Mowery et al., 2016) developed classifiers to detect depressive symptoms from tweets, further highlighting the potential of machine learning in this domain. Deep learning models, such as Long Short-Term Memory (LSTM) networks, have been employed for detecting depressive symptoms in text data, reporting high accuracy and performance

(Apoorva et al., 2023). Despite their success, LSTMs can be computationally intensive and may require large training data. Recently, pre-trained transformers like BERT have emerged as powerful tools for natural language processing tasks, including detecting depression in social media posts (Lin et al., 2020; Rizwan et al., 2022). Fine-tuned models have also been employed by researchers like (Lin et al., 2020), who combined CNN and BERT to use both visual and textual content.

## Dataset Description:

Normal Tweet:

Normal tweets were obtained from the Sentiment140 dataset available on Kaggle (https://www.kaggle.com/datasets/kazanova/sentiment140). A random sample of 10,000 tweets was selected from this dataset to represent non-depressive content.

Depressive Tweet:

Depressive tweets were initially sourced from a separate dataset available on GitHub (https://github.com/miladrezazadeh/twitter_depression_detection/blob/main/data/processed/processed_data.csv). This dataset consists of tweets that were scraped using the public Twitter API, with keywords related to depression to identify potentially depressive content. However, concerns have been raised in the literature that relying solely on depressive keywords may not accurately identify depressive tweets, prompting the use of additional filtering methods to enhance the dataset's quality, additional filtering was applied to ensure the quality of the dataset.

First, TextBlob sentiment analysis was used to filter the depressive tweets, setting a threshold of -0.2 to retain only tweets with negative sentiment. This process resulted in retaining only about 19.8% of the original depressive tweets. To further refine the dataset, the remaining tweets were classified using the zero-shot-classification model 'facebook/bart-large-mnli.' BART (Bidirectional and Auto-Regressive Transformers) is a powerful pre-trained transformer model designed for a range of natural language processing tasks. The 'facebook/bart-large-mnli' model has been fine-tuned on the MultiNLI dataset, which enables it to perform zero-shot classification, i.e., classifying text into categories without being explicitly trained on those categories. After filtering with TextBlob, about 94.5% of the tweets were predicted as "depressive tweets" by the zero-shot classification model.

The filtered depressive tweets were then assigned a label of 1, while normal tweets received a label of 0. The two sets of tweets were merged and shuffled, ensuring a diverse and balanced representation of both normal and depressive content. This combined dataset serves as the basis for training and evaluating the various machine learning models discussed in this report for the task of classifying depressive tweets.

## Methodology and Results:

Text Processing:

Before training the machine learning models, it is essential to preprocess the tweet text to remove noise, normalize the text, and extract meaningful features. The preprocessing steps performed in this project are as follows:

1. The tweet text was first converted to a string and then transformed to lowercase to ensure uniformity across the dataset.

2. Contractions, such as "I'm" becoming "I am," were expanded using the contractions library to simplify the text and promote consistency.
3. URLs were removed using regular expressions to eliminate irrelevant information and potential noise in the data.
4. Emojis were removed with a custom function called 'remove_emojis' and a regular expression that matched various emoji patterns. This step was crucial in isolating the textual information from visual elements.
5. Punctuation was removed using the 'string.punctuation' method to further simplify the text.
6. Numbers were removed using regular expressions to concentrate on the textual aspects of the tweets.
7. Stop words, which are commonly used words that do not carry much meaning (e.g., 'and', 'is', 'in'), were removed using the NLTK library to retain only the most significant words.

The processed tweet text resulting from these preprocessing steps was then used as input for the various machine learning models to classify depressive tweets. By performing these text processing steps, the dataset was optimized for analysis, allowing for a more accurate and effective classification of depressive content. A table summarizing the top 5 most common words in each category is provided below:

| *Rank* | *Normal Tweets* | *Frequency* | *Depressive Tweets* | *Frequency* |
|---|---|---|---|---|
| *1* | work | 3942 | mentalhealth | 9259 |
| *2* | go | 3606 | anxiety | 8755 |
| *3* | going | 3511 | depression | 8042 |
| *4* | cannot | 3331 | sad | 4368 |
| *5* | get | 3320 | suicide | 2451 |

Table 1.

From Table 1, it is evident that normal tweets predominantly contain words related to daily activities and general sentiments, such as 'work,' 'go,' 'going,' and 'get.' On the other hand, the most common words in depressive tweets are associated with mental health issues, including 'mental health,' 'anxiety,' 'depression,' 'sad,' and 'suicide.' The prevalence of these words highlights the focus on mental health and negative emotions in depressive tweets. Additionally, word clouds were created for both normal and depressive tweets, visually representing the frequency of words in each category. These word clouds further emphasize the differences in the most common words between normal and depressive tweets, reinforcing the patterns observed in the table.


Fig 1.

Traditional Classifiers:

In this study, we employed the popular scikit-learn library to split the dataset into training and validation sets, with 80% of the data used for training and 20% for validation. This approach helps ensure the model's robustness and its ability to generalize well to unseen data. For the traditional classifiers, we utilized TF-IDF (Term Frequency-Inverse Document Frequency) to tokenize the text data. The TF-IDF technique assigns a weight to each word based on its frequency in the document and across the entire dataset.

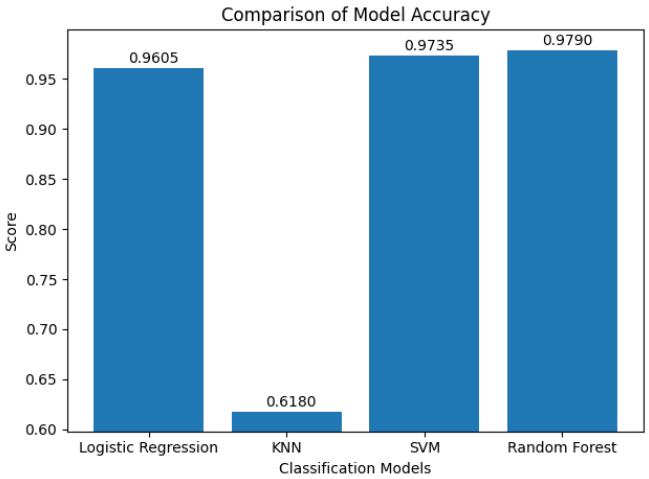


Fig 2.

Based on the model performance in Figure 2, Logistic Regression demonstrated a remarkable accuracy of 96.05% in classifying depressive tweets, but still a bit lower than the results from Support Vector Machines (97.35%) and Random Forest (97.90%). These models showcased high accuracy, precision, and recall rates, underlining their potential for detecting early signs of depression in social media data. The superior performance of Random Forests can be explained by its inherent capacity to reduce overfitting and increase model stability through averaging predictions across multiple decision trees. Logistic Regression also performed well but was slightly less effective compared to the top-performing classifiers. The K-Nearest Neighbors method, however, demonstrated weaker performance with about 61.80% accuracy. A potential explanation for this performance could be the presence of noisy data points or an imbalanced dataset, leading to misclassification of the minority class. Moreover, the choice of an appropriate distance metric and the number of neighbors (k) can greatly impact the model's performance.

Long Short Term Memory and Bidirectional Encoder Representations from Transformers:

In addition to the traditional classifiers, we have also implemented Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). Table 2 and Figure 3 summarized the performance of each model. In general, LSTM achieved an accuracy of about 97.05% and BERT achieved a perfect accuracy of 100%.

| *Model* | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|
| *Logistic Regression* | 0.9605 | 0.9593 | 0.9583 | 0.9588 |
| *KNN* | 0.6180 | 0.7921 | 0.5221 | 0.4223 |
| *SVM* | 0.9735 | 0.9700 | 0.9756 | 0.9726 |
| *Random Forests* | 0.9790 | 0.9754 | 0.9819 | 0.9783 |
| *LSTM* | 0.9705 | 0.9692 | 0.9694 | 0.9693 |
| *BERT* | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Table 2.

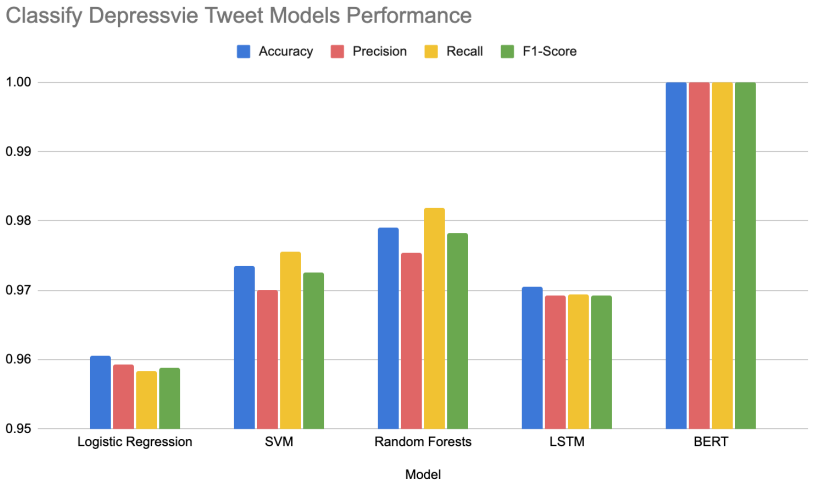


Fig 3.

The performance of LSTM in our analysis, although competitive, did not surpass that of traditional classifiers like SVM and Random Forest. A possible reason for this outcome could be the nature of our dataset, which was collected using specific keywords, potentially simplifying the classification task. This might have allowed traditional classifiers to achieve high accuracy without considering the context provided by word order. We hypothesize that LSTM might demonstrate superior performance on a more diverse and less obvious dataset, such as tweets from users diagnosed with depression by medical professionals. In such cases, the emotional content of the tweets may be more subtle, and the context derived from word order might become crucial for accurate classification. This would likely emphasize the advantages of LSTM over traditional classifiers in handling complex text data, highlighting its potential for detecting early signs of depression in social media.

Additionally, It can be observed that BERT achieved a perfect accuracy of 100% in classifying depressive tweets. In our analysis, BERT, a state-of-the-art transformer-based model, demonstrated exceptional performance in classifying depressive tweets, achieving a perfect accuracy of 100%. The model's outstanding precision, recall, and F1-score metrics underline its ability to accurately and consistently distinguish between depressive and non-depressive tweets. The success of BERT can be attributed to its innovative architecture, which leverages bidirectional context from the input text and utilizes self-attention mechanisms to capture long-range dependencies.

Web Application:

In addition to the research on classifying depressive tweets, we have also developed a web application to make this study more practical and accessible to the public, for example as Figure 4. This user-friendly application enables individuals to perform self-detection or assess the mental well-being of people they care about by analyzing their social media posts for signs of depression. By offering this resource, we aim to raise awareness about mental health and facilitate early detection and intervention. Furthermore, we encourage social media platforms to consider integrating this depression detection functionality into their systems. By automating the identification of early signs of depression in users' posts, platforms can play a proactive role in supporting mental health and well-being among their user base. Such a feature could alert users to potential mental health concerns and connect them with appropriate resources or support networks, ultimately contributing to a healthier online community.

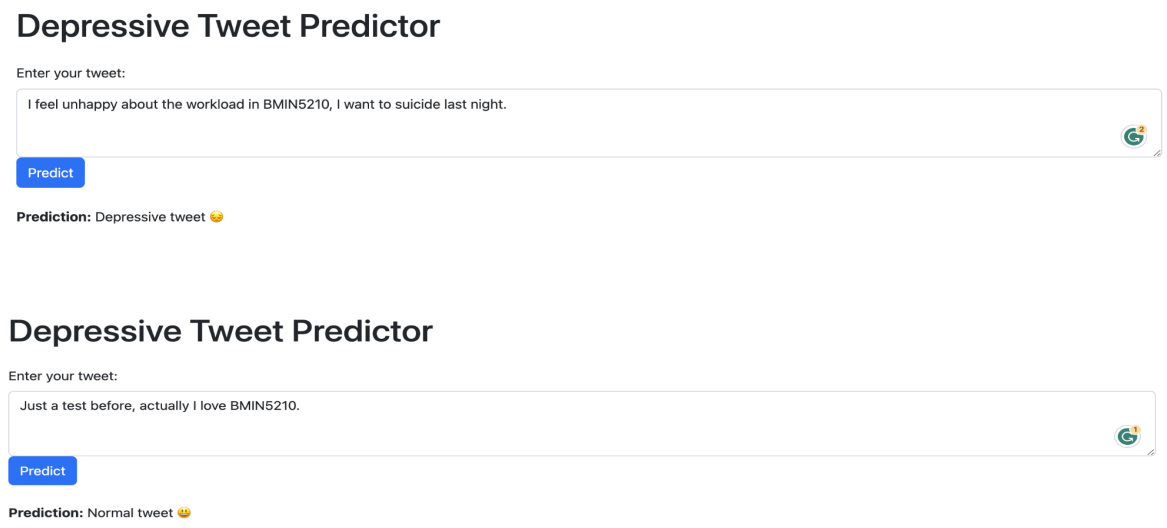


Fig 4.

## Conclusion:

In conclusion, this study has explored various machine learning methods, including traditional classifiers, LSTM, and BERT, to classify depressive tweets. The results demonstrate that BERT, a state-of-the-art transformer-based model, outperforms other methods by achieving perfect accuracy and exceptional performance metrics. This highlights the potential of BERT in detecting early signs of depression in social media data and advancing mental health research. The findings also reveal that traditional classifiers like SVM and Random Forest perform competitively in this classification task, while LSTM does not show a significant advantage over these methods. It is hypothesized that the dataset's nature, collected using specific keywords, simplifies the classification task and may not fully exploit LSTM's ability to capture context and retain information from the past.

Future work could focus on improving the dataset's diversity and complexity by incorporating tweets from users diagnosed with depression by medical professionals. This would likely provide a more nuanced representation of depressive language patterns and better leverage the contextual understanding of advanced models like LSTM and BERT. Additionally, exploring other transformer-based models or fine-tuning strategies could further enhance classification performance. Another potential direction for future research is to investigate multimodal approaches that combine text analysis with other data sources, such as user interactions, social networks, or multimedia content. This could provide a more comprehensive understanding of users' mental health and enable the development of more effective early detection and intervention strategies.

References:


1. Apoorva, A., Goyal, V., Kumar, A., Singh, R., & Sharma, S. (2023). Depression Detection on Twitter Using RNN and LSTM Models. *Communications in Computer and Information Science*, *1798 CCIS*, 305–319. https://doi.org/10.1007/978-3-031-28183-9_22/COVER
2. *Depressive disorder (depression)*. (n.d.). Retrieved April 3, 2023, from https://www.who.int/news-room/fact-sheets/detail/depression
3. Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020). SenseMood: Depression detection on social media. *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval*, 407–411. https://doi.org/10.1145/3372278.3391932
4. Mowery, D., Park, A., & Bryan, C. (2016). *Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health*. http://liwc.wpengine.com/
5. Nadeem, M., Horn, M., Coppersmith, G., Hopkins University, J., & Sen, S. (n.d.). *Identifying Depression on Twitter*.
6. Rizwan, M., Mushtaq, M. F., Akram, U., Mehmood, A., Ashraf, I., & Sahelices, B. (2022). Depression Classification From Tweets Using Small Deep Transfer Learning Language Models. *IEEE Access*, *10*, 129176–129189. https://doi.org/10.1109/ACCESS.2022.3223049