# Detection of Depression among Social Media Users with Machine Learning

**M. Senthil Raja**
Research Scholar, Department of CSE, B.S.A. Crescent Institute of Science and Technology, Tamilnadu, India.
Assistant Professor, Department of CSE, SRM Institute of Science and Technology, Tamilnadu, India.

**L. Arun Raj**
Associate Professor, Department of CSE, B.S.A. Crescent Institute of Science and Technology, Tamilnadu, India.

**A. Arun**
Assistant Professor, Department of CSE, SRM Institute of Science and Technology, Tamilnadu, India.

*Received August 05, 2021; Accepted November 20, 2021*
*ISSN: 1735-188X*
*DOI: 10.14704/WEB/V19I1/WEB19019*

## Abstract

Mental illnesses are a significant and growing public health concern. They have the potential to tremendously affect a person's life. Depression, in particular, is one of the major reasons for suicide. In recent times, the popularity of social media websites has burgeoned as they are platforms that facilitate discussion and free-flowing conversation about a plethora of topics. Information and dialogue about subjects like mental health, which are still considered as a taboo in various cultures, are becoming more and more accessible. The objective of this paper is to review and comprehensively compare various previously employed Natural Language Processing techniques for the purpose of classification of social media text posts as those written by depressed individuals. Furthermore, pros, cons, and evaluation metrics of these techniques, along with the challenges faced and future directions in this area of research are also summarized.

## Keywords

Tokenization, Stemming, TF-IDF.

## Introduction

According to the World Health Organization, about 260 million suffer from depression worldwide. It is a prevalent psychiatric disorder which is a leading cause of disability globally. If accurately diagnosed, it can be treated through pharmacologically as well psychologically. However, means for detection and the availability of treatment and other mental illnesses remain limited.

There is a strong relationship between the mental condition and language of an individual. Depression can be marked by linguistic predictors like high usage of words depicting greater negative emotions along with references to hopelessness and sadness. There has been no dearth of studies in the role that social media plays in the detection of mental illness, through the analysis of textual posts, comments, and other online activity. Since social media platforms like Facebook, Twitter, and Reddit have become an integral part of human lives, they also provide a platform where one can share their thoughts, emotions, and beliefs. Thus, social media can act as data banks that researches can employ to extract and analyze data. A platform of interest that has been widely utilized is Reddit, Inc since it contains groups called ‘subreddits’ that channel conversation focused on a particular subject. r/depression, r/anxiety, and r/mentalillness are examples of subreddits that may be useful for research purposes in this specific area.

With an explosion of availability of information in the modern age, it is impossible to manually label data. Text classification is the process of assigning labels for text and is an essential task in Natural Language Processing (NLP) that finds various applications in spam filtering, sentiment analysis, business intelligence, fraud detection etc. The major steps involved in classification of text data include– but are not limited to– text extraction, preprocessing, learning through a model, and evaluation of metrics. Preprocessing of raw text data is conclusive of many steps and can be done with the help of the NLTK (Natural Language Toolkit) package in python. Initially, commonly used words that do not contribute to the sentiment of the text (like ‘a’, ‘an’, ‘the’ etc.)– called stop words–are removed along with punctuations. Lemmatization is the process of reducing words to their root forms. For example, words like ‘played’, ‘playing’, and ‘plays’ will be reduced to their root word ‘play’. Stemming is another similar process in which the root word is not always a dictionary word. Tokenization is the process of separating text into a list of individual words. Before using models for learning, vectorization is used to convert text data into a vector of numbers called word embeddings. These may come from pre-trained models like Word2Vec or GloVe. They may also be computed using methods like N-Grams, Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words

(BOW) etc. with shallow learning models. BOW represents text with a vector that is the size of a dictionary.

The value of the vector is directly related to the word frequency corresponding to its position in the text. On the other hand, N-gram considers the information of N adjacent words and builds a dictionary with its help. TF-IDF uses the word frequency and inverses the document frequency to model the text using the following formulae:

TF = (Number of times a word appears) / (Total number of words in the document)
IDF = log_e(Total number of documents / Number of documents with the word in it)
The product of these quantities is termed as the TF-IDF weight.

The word2vec employs local context information to obtain word vectors. The GloVe – with both the local context and global statistical features – trains on the nonzero elements in a word-word co-occurrence matrix. Then according to the selected features, the classifier receives the input of the represented text.

Text Classification is used to demarcate the author of the text data into two categories: depressed and not depressed. Often, a third category of 'indeterminable' is also added for inconclusive text posts which do not lean towards either category.

These types of text classification problems can make use of shallow learning as well as deep learning models. In the former, classic machine learning algorithms are employed after feature engineering, which is a process in which useful linguistic features are extracted from the corpus and their most useful combinations are used. Linguistic Inquiry Word Count (LIWC) is one of the most widely used feature engineering methods across literature since it contains multiple categories of psychological constructs.

Since diagnosing mental illnesses is a complicated procedure, this is a very important step. Some of the classic algorithms are K-Nearest Neighbours (KNN), Support Vector Machine (SVM), etc. In deep learning, feature selection is integrated into the learning model itself. These deep neural network architectures may be based on Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) etc. with appropriate hyper parameter tuning.

The structure of the paper is as follows. Firstly, an examination of techniques that have already been employed in regards to detection of depression with the help of social media activity are presented along with the various datasets used. Their architectures, performance metrics, advantages, and disadvantages are compared. Then, a proposed

model for the same problem is delineated. And finally, further directions of research in this area of research are summarized with a conclusion of key implications stemming from this study.

## Literature Review

### Datasets

Due to the correlation between the mental health and social media activity of individuals, there is a need for labelled data with large. However, due to privacy concerns of patients, only public posts of self-diagnosed users are extracted and labelled.

The Self-Reported Mental Health Diagnosis (SMHD) dataset has been widely used for research purposes.[Arman Cohan,2018] It is a large-scale labelled dataset containing posts from users diagnosed with one or more mental conditions like autism, eating disorders, bipolar disorder, depression etc. Another important dataset is the Reddit Self-Diagnosed Depression dataset (RSDD). It contains data from 28 subreddits which is labelled according to the precision of diagnosis patterns. [Andrew Yates, 2017] Across literature, the Application Programming Interfaces (APIs) of Twitter and Reddit have been used to extract data containing keywords related to depression in the former or from relevant subreddits from the latter.

### Classification Techniques

Reference [Michael M. Tadesse, 2019] focused on determining the connection between text language and depression by checking the performance of feature sets by using several classification methods. Similarly, reference [Michael M. Tadesse, 2019] also proposed methods for identifying posts in support communities that may indicate a risk of self-harm. Reference [Subhan Tariq, 2019] proposed a methodology to classify the patients with chronic mental illness diseases based on the data extracted from Reddit, a well-known network community platform.

Reference [G. Rao, Y. Zhang, 2020] emphasised on early detection of depression and risk of self harm and suicide by comparing models against several baselines with the help of Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) classifiers.

In reference [X. Fu, J. Yang, 2018] two inputs: word embedding and its sentiment embedding were fed into a Long Short Term Memory (LSTM) model to increase the

accuracy of word representation Reference [A. Abdul Aziz and A. Starkey, 2020] focused on various methods through which the decline in the performance of sentiment analysis models could be measured. Reference [A. Feizollah, S. Ainin, 2019] depicted the process of data extraction and sentiment analysis of tweets.

For pre-processing of data, reference [Michael M. Tadesse, 2019] and [Subhan Tariq, 2019] used NLP tools like tokenization, stemming to obtain clean data. To study the features from the posts, N-gram and Latent Dirichlet Allocation (LDA) models were employed. For example, Multilayer Perceptron (MLP) classifier which consisted of combined features showed the best results by reaching 91% accuracy and .93 F1 score.

For feature extraction [Subhan Tariq, 2019] utilises TF-IDF words factorization, whose output consisted of vectors containing the weight of the words. It was followed by the usage of clustering and pruning techniques for feature selection. Namely three classifiers were considered: SVM, Random Forest (RF) and Naive Bayes (NB).

Three models were built and using the method of cross validation the models were trained. With co-training, SVM, NB and RF performed better as compared to the usual recall, precision and F-measure terms. In [G. Rao, Y. Zhang, 2020] and [X. Fu, J. Yang, 2018], the models were compared against MNB and SVM classifiers. The model proposed in [G. Rao, Y. Zhang, 2020] consisted of an architecture based on a CNN, and a combination of output and merge layer. The suggested model performed well in terms of F1 and recall but not in terms of precision. Whereas in the suggested Single-Gated ReLu CNN (SGL-CNN) and Multi-Gated ReLu CNN (MGL-CNN) models consisted of two convolutional layers along with global average pooling. This model performed well in terms of precision as well. In [A. Abdul Aziz and A. Starkey, 2020] the word sentiment classifier was trained by using sentiment lexicon to produce sentiment embeddings of words. In [A. Feizollah, S. Ainin, 2019] for sentiment analysis of tweets, feature extraction was performed by using Word2Seq and Word2Vec techniques. For training of the neural network algorithm, different datasets like IMDB (movie reviews), Amazon (product reviews) etc. were used. For the implementation of neural network architecture, a Python library called Keras was used.

From the Python library three functions were utilised, Hyperbolic tangent (Tanh), Rectifier linear unit (ReLU) and Softmax function (Softmax). It was concluded from the experiment that CNN and LSTM algorithms showed an accuracy of 93.78%.

## Proposed Model

### Data

Our research intend to extract text posts (not comments) from the subreddit r/depression–a very popular subreddit where users seek help and community among other people suffering from depression– using the Reddit Application Programming Interface and the Python Reddit API Wrapper (PRAW). These posts will be labelled as 'depressed'. The data corresponding to the label 'not depressed' will be extracted from various other subreddits that channel conversations on lighter notes like r/AskReddit and r/TodayILearned. These posts will be split into training and test data.

### Methodology

In this research work propose to analyse various deep learning models and compare their performance metrics with a baseline model that combines the TF-IDF vectorization method with an SVM linear classifier. Shallow models– unlike deep learning models– require feature engineering before fitting the data in the model. However, in deep learning, this step is integrated into the model by learning a set of nonlinear transformations that serve to map features directly to outputs. Due to this added advantage, employing RNNs as well as CNNs based architectures prove to be beneficial. We will apply CNNs with a single dimension as well as multiple convolutions with filters of varied lengths and max-pooling layers. RNNs are very widely used in NLP since they have 'memory' over various time durations. They also allow variable length processing while maintaining the sequence order. The Long Short Term Memory (LSTM) model is a commonly used RNN. We will compare the following models: LSTM, bi-directional LSTM, and bi-directional LSTM with attention. These trained modes will be used for evaluation on unseen data. These architectures will be evaluated alongside various commonly used word embeddings like skip-gram, CBOW etc.

The metrics that will be used for evaluation are F1-score, ROC area-under-the-curve (AUC), precision, recall, and accuracy. The area under the curve ROC (Receiver Operating Characteristics) is called AUC. The ROC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) to depict the performance of a classification model at different classification thresholds.

$$TPR = TP/(TP+FN)$$
$$FPR = FP/(FP+TN)$$

The formulae for the other aforementioned metrics are mentioned below.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1 score = (2*precision*recall)/ (precision+recall)

where,

TP = Number of true positives

FP = Number of false positives

TN = Number of true negatives

FN = Number of negatives

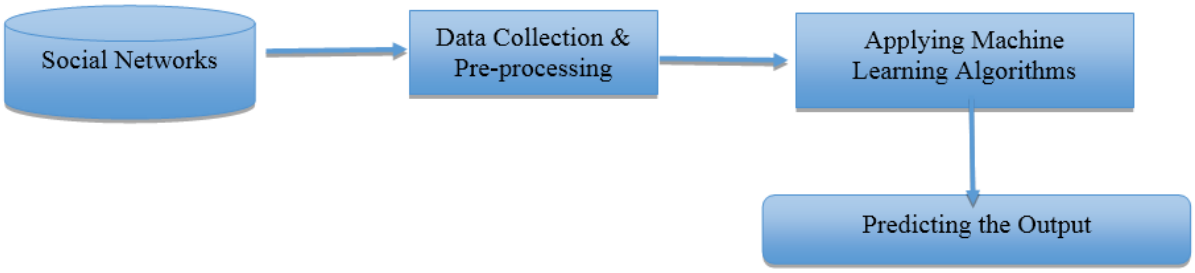


**Figure 1 Architecture Diagram**

## Future Research Directions

Text classification is the technique that has been primarily used to detect depression among social media users. It takes input text data which is represented as a vector by the pre-training model. This vector is then fed into multiple types of architectures for training and finally, the performance of these models is verified. This research depicts the possibility of creating applications that can be used by clinicians and other mental health professionals to notify them if their patients suggest that they are at the risk of self-harm or suicide. For text classification existing models have proved useful, but there is still a long way to go and many domains remain unexplored. Moreover, different combinations of these models can be studied with new machine learning algorithms. These results can also underline the infrastructure for new mechanisms applied in other areas of healthcare as well. For example, other mental disorders like Borderline Personality Disorder, Bipolar Disorder, and Generalized Anxiety Disorder can also be detected using NLP techniques.

## Conclusion

A survey of previously employed techniques to detect depression and other mental illnesses was performed. Various datasets and training techniques were evaluated based on many performance metrics. In this work, we argued for the close connection between

social media and mental health. For future work, we also proposed to perform a comparative evaluation on some of the widely used deep learning models for depression detection on data extracted from the Reddit API. We believe that these methods will help in the development of new tools to identify mentally ill individuals and it will enable those at-risk individuals to be detected so that they are able to seek treatment. Finally, we summarize the potential future research and challenges of depression detection using social media posts.

## References

Cohan, A., Desmet, B., Yates, A., Soldaini, L., Mac Avaney, S., & Goharian, N. (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Tadesse, M.M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access, 7*, 44883-44893.

Tariq, S., Akhtar, N., Afzal, H., Khalid, S., Mufti, M.R., Hussain, S., & Ahmad, G. (2019). A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access, 7*, 166165-166172. http://doi.org/10.1109/ACCESS.2019.2953087

Rao, G., Zhang, Y., Zhang, L., Cong, Q., & Feng, Z. (2020). MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access, 8*, 32395-32403. http://doi.org/10.1109/ACCESS.2020.2973737

Fu, X., Yang, J., Li, J., Fang, M., & Wang, H. (2018). Lexicon-enhanced LSTM with attention for general sentiment analysis. *IEEE Access, 6*, 71884-71891. http://doi.org/10.1109/ACCESS.2018.2878425.

Aziz, A.A., & Starkey, A. (2019). Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. *IEEE Access, 8*, 17722-17733. http://doi.org/10.1109/ACCESS.2019.2958702

Feizollah, A., Ainin, S., Anuar, N.B., Abdullah, N.A.B., & Hazim, M. (2019). Halal products on Twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. *IEEE Access, 7*, 83354-83362. http://doi.org/10.1109/ACCESS.2019.2923275