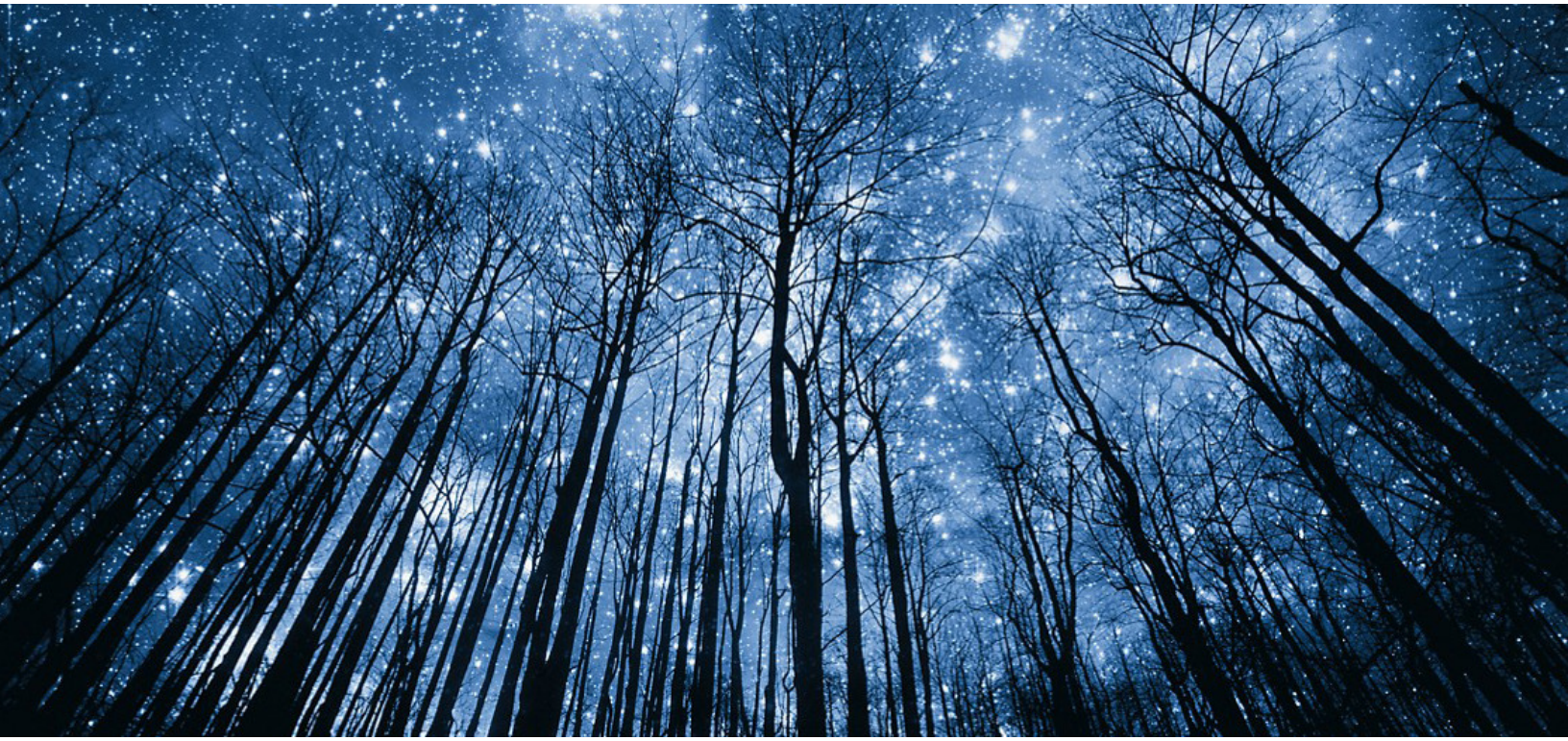
DELL Technologies

DEPRESSION DETECTION IN SOCIAL MEDIA USING MACHINE LEARNING



Shruthi Krishnakumar

DELL Technologies
Proven Professional

Knowledge Sharing Article

Table of Content

1. Introduction

Depression as a common mental health disorder has long been defined as a single disease with a set of diagnostic criteria. It often co-occurs with anxiety or other psychological and physical disorders; and has an impact on feelings and behavior of the affected individuals. According to the WHO study, there are 322 million people estimated to suffer from depression, equivalent to 4.4% of the global population. Nearly half of the in-risk individuals live in the South-East Asia (27%) and Western Pacific region (27%) including China and India. In many countries depression is still under-diagnosed and left without any adequate treatment which can lead into a serious self-perception and at its worst, to suicide [1]. In addition, the social stigma surrounding depression prevents many affected individuals from seeking appropriate professional assistance. As a result, they turn to less formal resources such as social media to express personal emotions and mental state. With the development of Internet usage, people have started to share their experiences and challenges with mental health through tweets or any other social media platform.

Given the prevalence of depression, its consequences, and the possibility of detecting depression using social media texts, the main goal of this study is to analyze whether machine learning (ML) methods could be effectively used to detect depression in people by analyzing their social media texts, but without relying on specific keywords. Social media texts are often unstructured, and therefore ML, which is good at dealing with such data, can give better results than other traditional methods.

A humongous quantity of data is generated every second through various social media platforms like Facebook, Twitter, etc. These all data from different platforms have a lot of relevant information for doing behavior analysis. Twitter is one of the maximum visited social networking web sites wherein a median of fifty eight million tweets are generated according to day and there are over 271 million month-to-month lively customers in Twitter So, it becomes easy to get data for analyzing depression with tweets. In this study, the main goal is to find out which users are depressed.

2.Research Methodology

There are a growing number of methodologies for detecting depression from text messages. In our study, we combine a technical description of the methods applied to identify depression using NLP and a text classification technique. There are several steps in the research method, which are:

1. Data Collection and Preparation

2. Data Preprocessing

3. Word Analysis

4. Tokenization

5. Extract Features

6. Depression Detection

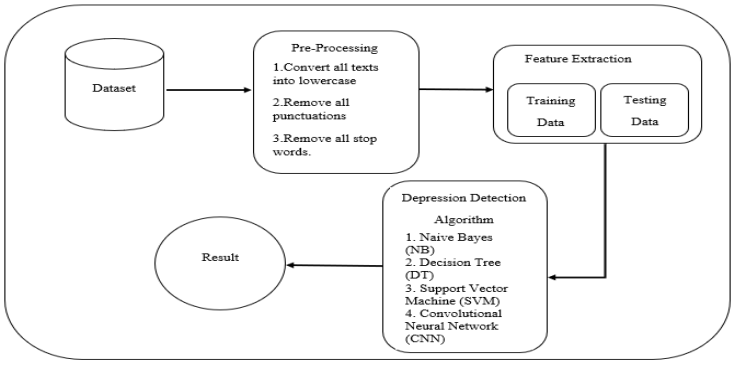
7. Output

2.1 Data Collection and Preparation

Here we have used the labeled dataset from Kaggle that collected and used around 5,000 plus tweets and all of these data were in English. And we have split the dataset into two parts in the ratio of 80:20.

Where 80 is for training and 20 is for testing data. The new dataset was prepared for testing the work with a new dataset for getting new or experimental output.

Fig 1 Workflow



Here the dataset is gone through a pre-processing phase where all the texts are converted into lowercase and removal of punctuations and stop words. Then we are vectorizing the data and tokenizing. After that the datasets are split into training and test dataset. Then, on the basis of test data, we are getting outputs for different machine learning algorithms.

2.2 Data Preprocessing

Process of removing noise from the data and cleaning and simplifying the data. However, three types of preprocessing is performed in this study's dataset named as:

1. Convert all texts into lowercase

2. Remove all punctuations

3. Remove all stop words.

We use the NLP tools to pre-process the dataset before it is proceeded to the feature selection and training stage. First, we use tokenization to divide the posts into individual tokens. Next, we remove all the URLs, punctuations and stop words which could lead to erratic results if stay ignored.

2.3 Word Analysis

The process of learning more about any word in a meaningful way is referred to as word analysis. It's possible that different results would have been obtained if word analysis had been undertaken prior to data preprocessing. The most often used terms were displayed in large letters, while the least frequently used words were displayed in small letters. As a result of word analysis, the idea of the most and least used words was clarified in this study. The use of a line kernel convolutional neural network (LKCNN) to handle baseband data directly to detect fall motions is proposed in this paper. The ability of a convolutional neural network (CNN) to learn to extract valuable features during the training process is used in this strategy.

2.4 Tokenization

Tokenization is the process of breaking down raw data into discrete tokens made up of words and sentences. The tokens make it simple to comprehend the context or model that is being developed. Tokenization can be divided into two categories. The first is known as word tokenization, and the second is known as sentence tokenization. Word tokenization is when text is broken into words using any tokenization technique, while sentence tokenization is when text is split into sentences using any tokenization technique. Tokenizing strings and assigning an integer id to each potential token, for example by utilizing white-spaces and punctuation as token separators, are the most prevalent methods for extracting numerical information from text material.

2.5 Extract Features

Feature extraction is a general term for a method for developing blends of the factors to get around these issues while depicting the information with adequate exactness. The feature extraction method extracts the aspect (adjective) from the improved dataset and the adjective is used to determine the polarity of a sentence. The Unigram model extracts the adjective and segregates it. Here we do vectorization, which is the general process of turning a collection of

text documents into numerical feature vectors. This specific strategy (tokenization, counting and normalization) is called the Bag of Words.

In our study we have used TfidfVectorizer, which helps to convert a collection of raw documents to a matrix of TF-IDF features. TfidfVectorizer integrates all the parameters of CountVectorizer and TfidfTransformer into a single model because Tf-idf is frequently used for text features. This method is used in Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF).

2.6 Depression Detection:

After splitting the data into testing and training, now it is the time to work on those data. The 20% testing data was utilized to obtain accuracy, precision, recall, and F1-score from multiple machine learning algorithms for the purpose of detecting depression. This study generated varied outcomes for several machine learning algorithms based on test data. The following algorithms were used in the study:

1. Naive Bayes (NB)
2. Decision Tree (DT)
3. Vector Support Machine (SVM)
4. Forest of Chance (RF).
5.Convolutional Neural Networks (CNNs)

Classification problems are significantly solved using Supervised learning at runtime. CSV file of tweets split in training and test sets are read using pandas library. Model is trained and tested using text_emotion.csv that contains labeled tweets and also we calculate accuracy using F1 score.

In our study we have also included emotion prediction using tensorflow.Here we are giving an image to predict. so by using the haar cascade frontal face algorithm we can identify the faces in an image then crop them. After that we have rescale it into 48 x 48. We have predicted the emotion using a pretrained model with CNN algorithm.

2.7 Output

Machine learning algorithms were applied for depression detection from the tweets in order to find out how many users are in depression and how many users are not. After applying all those machine learning algorithms, this study concludes that different algorithms gave

different accuracy. However, in this study, it is shown which algorithm is best in terms of providing accuracy, precision, recall and F1-score for getting the best machine learning algorithm for depression detection.

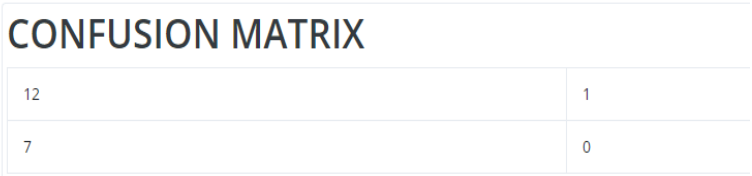
3 Experimental Results and Discussions

The effectiveness of the proposed research methodology had been evaluated in the study. Several machine learning algorithms had been applied in this research work for getting a better result of which algorithm gave better results for detecting depression from Twitter's tweets. In our section, all the performance of different applied algorithms are analyzed and predict the accurate algorithm. Performance metrics used in this study will also be discussed in our section.

3.1 Performance Metrics

There are several metrics which can be used in order to evaluate the machine learning algorithms. However, for choosing any performance metric, the basic things needed to be remembered are the performance of the machine learning algorithms, dependency, and importance of various characteristics which will influence the result of the algorithms. Performance metrics used for this study were Confusion matrix, Accuracy, Precision, Recall and F1-score.

1. Naive Bayes (NB)



Here we are showing the confusion matrix for Naive Bayes (NB). In this matrix the actual depressed data to predicted depressed data is 12 which is truly negative and in predicted time actual depressed data to predicted normal is 1 which is false positive and actual normal to predicted normal is 0 which is truly positive and in actual normal to predicted depressed is 7. So for calculating the accuracy score = (12+0)/ (12+1+0+7).

2. Decision Tree (DT)

Confusion Matrix

	DEPRESSED	NORMAL
DEPRESSED	12	2
NORMAL	1	5

Here we are showing the confusion matrix for Decision Tree (DT). In this matrix the actual depressed data to predicted depressed data is 12 which is truly negative and in predicted time actual depressed data to predicted normal is 2 which is false positive and actual normal to predicted normal is 5 which is truly positive and in actual normal to predicted depressed is 1.

3. Support Vector Machine (SVM)

Here we are showing the confusion matrix for Support Vector Machine (SVM). In this matrix the actual depressed data to predicted depressed data is 13 which is truly negative and in predicted time actual depressed data to predicted normal is 5 which is false positive and actual normal to predicted normal is 2 which is truly positive and in actual normal to predicted depressed is 0. So for calculating the accuracy score = (13+2)/ (13+5+0+2).

4. Convolutional Neural Network (CNN)

CONFUSION MATRIX

15	12	133
1	0	2
9	8	72

Here we are showing the confusion matrix for Convolutional Neural Network (CNN). In this matrix the actual depressed data to predicted depressed data is 15 which is truly negative and in predicted time actual depressed data to predicted normal is 5 which is false positive and actual normal to predicted normal is 2 which is truly positive and in actual normal to predicted depressed is 0. So for calculating the accuracy score = (13+2)/ (13+5+0+2).

5. Random Forest (RF).

Confusion Matrix

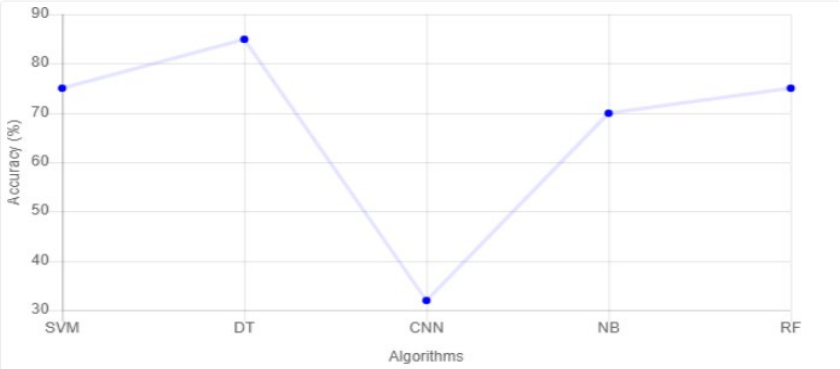
	DEPRESSED	NORMAL
DEPRESSED	12	3
NORMAL	1	4

Here we are showing the confusion matrix for Random Forest (RF). In this matrix the actual depressed data to predicted depressed data is 12 which is truly negative and in predicted time actual depressed data to predicted normal is 3 which is false positive and actual normal to predicted normal is 4 which is truly positive and in actual normal to predicted depressed is 1. So for calculating the accuracy score = (12+4)/ (12+4+3+1).

3.2 Depression Prediction Performance

Five types of machine learning algorithms were used in this research work for finding out which algorithm provides the best performance for detecting depression with tweets. The finalized and preprocessed dataset was used in those algorithms after splitting that dataset into 20% in testing and 80% in training data. On the basis of test data, the predictions were performed and found out the accuracy, precision, recall and f1-score for all of the machine learning algorithms. All the results of accuracy, precision, recall and f1-score for all applied machine learning algorithms were judged and compared among each other on account of declaring the best machine learning algorithm. Here goes the accuracy, precision, recall and f1-score of all applied machine learning techniques:

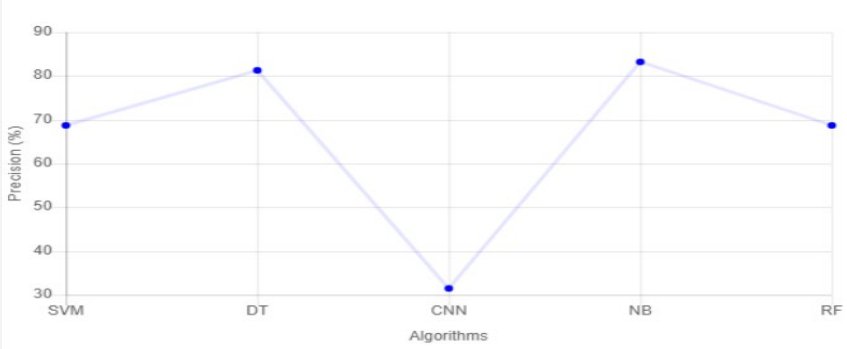
Fig.2



In Fig.2 ., there is a comparison graph for accuracy among all the applied algorithms. From the graph, it's clear that Decision Tree had the best accuracy among all the used machine learning algorithms which was 85.00% . So, Fig. 2. indicated that among all used machine

learning algorithms in this research, Decision Tree provided the highest accuracy where CNN provided the lowest accuracy.

Fig.3.



According to Fig.3., the highest precision provided machine learning algorithm was Naive Bayes for this research work. Naive Bayes gave 83.33% precision which was quite effective for depression detection. On one side, Naive Bayes had the highest precision, while on the other side, CNN had the lowest precision. Rest of the applied machine learning algorithms provided between 68–82% precision which was also standard.

Fig.4.

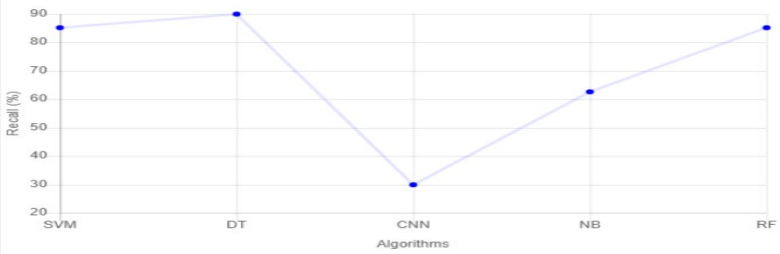
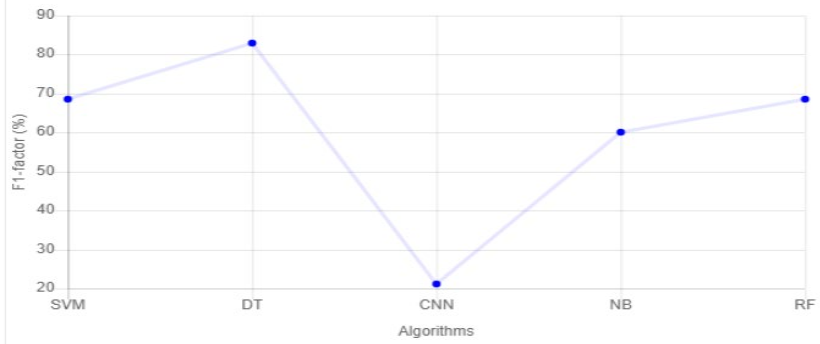


Fig.4. indicates the recall rate of all used machine learning algorithms through a graph with data level. Through the graph, it shows that Decision Tree resulted in the highest recall rate and CNN resulted in the lowest recall rate. The recall rates of rest algorithms are also shown in Fig.4.

Fig.5.



F1-score for all the machine learning algorithms which were used in this study had been picturized beautifully in Fig.5. In the Fig.5 ., 82.90% is the highest F1-score which is Decision Tree and 29.81% is the lowest F1-score rate which is CNN.

3.3 Comparison the Algorithms' Results:

In the previous component, accuracy, precision, recall and f1-factor of all applied algorithms had been described elaborately. Though that description, it had been stated Decision Tree (DT) gave the highest accuracy, Naive Bayes(NB) provided the best precision and Random Forest (RF) and Support Vector Machine (SVM) provided the highest recall and Decision Tree (DT) provided the highest f1-factor.

Table 2 Comparison of all applied algorithms' results:

Algorithms' name	Accuracy (%)	Precision (%)	Recall (%)	F1-factor (%)
Support Vector Machine (SVM)	75.00	68.75	85.29	68.65
Decision Tree (DT)	85.00	81.25	90.00	82.90
Convolutional Neural Network (CNN)	32.00	31.46	29.81	21.09
Naive Bayes (NB)	70.00	83.33	62.50	60.00
Random Forest (RF)	75.00	68.75	85.29	68.65

Fig.6 Comparison of all applied algorithms' results:

4 Limitations and Future Research Direction

Our study also mentioned that the work was only for the English language. So in future we can train the data of different languages. And also we can include training using the tweets with emojis.

5 Conclusions

This study presented a generic depression detection model with the use of five machine learning algorithms. The goal of this study is to detect depression from any tweets that contain any depressive word, and to reduce depression and suicides as a result of depression. In this work, a methodology has been followed and through following this methodology, any tool can be built for detecting depression easily in future. Any other tweet which is pre-processed, it can be told if the data indicates depression or not by following the methodology of this study. This study shows that while comparing CNN, SVN and NB, we get NB with highest accuracy. And when combined with all five algorithms that are CNN, SVM, NB, RF and DT, we get DT with highest accuracy. So, this study also resulted that among all the applied machine learning algorithms Decision Tree (DT) had provided the best result for depression detection which means if any other study wants to work on depression detection based on machine learning algorithms, then that study can use Decision Tree (DT) without any doubt for getting the best result for depression detection.

6. Bibliography

1.M. J. Friedrich, ‘‘Depression is the leading cause of disability around the world,’’ JAMA, vol. 317, no. 15, p. 1517, Apr. 2017.

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes, or methodologies.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.