Depression Detection of Tweets and A Comparative Test

P. V. Rajaraman [1]
[1]Professor,
Department of Computer science and Engineering,
Rajalakshmi Engineering College, India
Chennai, India

Asim Nath [2],Akshaya.P.R [3], Chatur Bhuja.G [4]
[2,3,4]Student,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai, India

*Abstract:***Detection of depression through messages sent by a user on social media can be a complex task due to the popularity and trends in them. In recent years, messages and social media has ended up being a very close representation of a person's life and his mental state. This is a huge stockpile of data about a person's behaviour and can be used for detection of various mental illnesses (depression in our case) using Natural Language Processing and Deep Learning. This project is about constructing a deep learning model using NLP to predict such mental disorders.**

***Keywords:-** NLP; Machine learning; Deep learning; Naive Bayes; LSTM; Twitter; Depression detection; Suicide.*

I. INTRODUCTION

In today's world, communication through social media is emerging as a big deal. They're willing to share their thoughts, stories and their personal feelings, mental states, desires on social network sites , blogging platforms etc.. Receivers use the manuscripts from emails and other types of social media comments to form proper reasoning and to correct the mistakes.When people write digitally on social media ,their texts are processed automatically. Natural language processing techniques are used to infer people's mentalbehaviour.

According to WHO, depression is a common worldwide folie that affects an enormous amount of individuals irrespective of their age. There are multiple factors that interfere the depression detection and treatment like lack of professional specialists, social shaming, improper diagnosis and so on.The ever-lasting depression disorder could lead to suicide if the depressed individuals are not supplied with proper consultance ,instant help and can also suffer from anxiety.This work is targeted on the detection of depression and anxiety from tweets.The experiment conducted during this work requires the text data so the chosen data source is Twitter where peopletweet about their feelings,hopes,desires,thoughts,stories and mental states.

The goals of our research are: collect the publicly available media messages of healthy and self-diagnosed individuals which contains mixed emotions so evaluate the extracted Twitter data and apply machine learning classifiers such as Naive Bayes,SVM and deep learning classifiers such as LSTM-RNN to predict depressive and anxiety tweets.

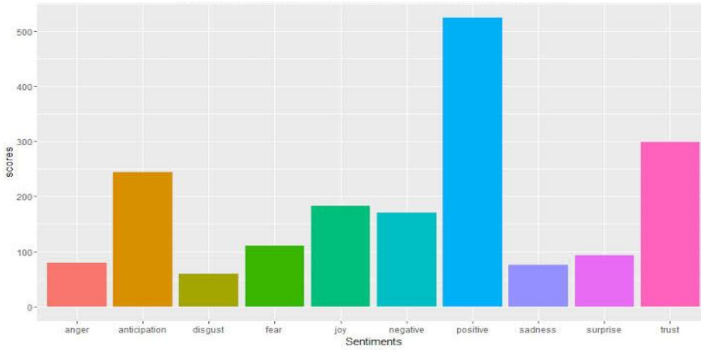


Fig 1. Various sentiments and its score

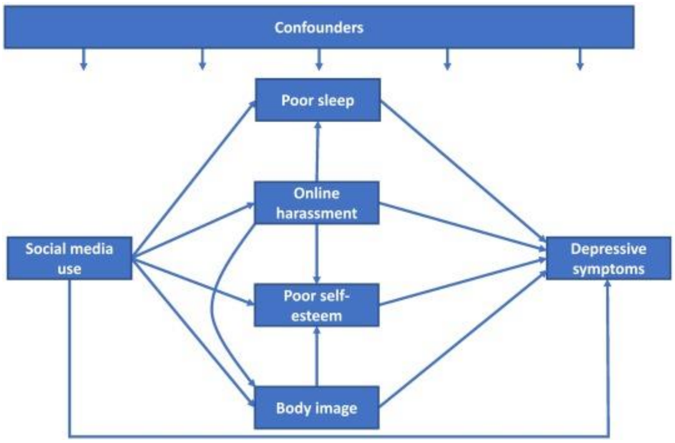


Fig 2 .Social media usage and it's drawbacks

II. RELATED WORK

There are various kinds of relationship between psychotic behaviour and the language being used in social media that replaced depression detecting dimension.

1. In the past few years a significant portion of our daily routine has been consumed by social media . We should keep updating our social media sites regularly, as they contribute the first source of communication.[1](AmnaNoureen and UsmanQamar in 2017) They wrote-up on user behavior and associated psychotic problems that was equipped and a comparative technique for psychotic behavior classification was also provided effectively.

2. [4](DQingCong,ZhiyongFeng,Guozheng Rao) The writers focused on solving the matter caused by data instability within the physical world.. The X-A-BiLSTM model consisted of two essential elements,where the primary element acquired balanced data by means of an end to finish boosting system, and also the second component, BiLSTM by using attention mechanism, which resulted in good classification performance. Reddit dataset was used to detect the depression emotion.Linguistic traces of communication was being used to find reasons for suicidal thoughts.

3. Emotion AI is a popular field for emotional detection research using text mining . The emergence of web based social forum sources have paved way for notable data that's present for sentiment analysis of text and images.[3] (Mandar Deshpande and Vignesh Rao in 2017) .The main aim of the writers were to detect the depression on twitter feeds using natural language processing. Individual tweets are classified as depressive or happy tweets, supported a collected wordlist to detect depression propensity.For class prediction Support vector machine and Naive bayes models are want to predict depression.The outcome was presented using different mining measurements such as precision,f1-score,accuracy etc.

4. [2](Rafael.A, Calvo, David N. Milne, M. Sazzadhussain, Helen Christensen in 2017).during the observation the authors provide a glossary of knowledgeable sources and techniques that involve prediction of mental illness . Specifically, they experimented how social media data lead the individuals to get into mental breakdown. The computational techniques utilized in labeling and diagnosis and eventually there are some ways to generate and personalize psychological behaviours.

5. [5](Jane H. K. Seah and Kyong Jin Shim in 2018).Their study demonstrated that a data mining approach will be useful for detecting depression in social media.. In Singapo a 24-hour suicide helplines are available.These services make sure that the people are safe and at the same time, tapping into digital traces such as public forums can help authorities take incharge to reach out people who need help.

III. METHODOLOGY

The Implementation of the project is carried out in the python 3.1 and the following libraries are used:

- Numpy
- Scikit Learn
- Matplot
- NLTK
- WordCloud
- Keras

Since this is a comparative analysis we are going to be using several models from different families of Algorithms. The various Models to be used are:

- TF-IDF Classifier
- LSTM
- Naive Bayes
- Linear Support Vector
- Logistic Regression.

The three datasets we are using for the Model are:

- Sentiment 140
- tweet Scrap from TWINT
- google word2vec

The implementation is broken down into several parts and those are as follows:

1. **Data Retrieval:**
   Getting the datasets and loading them into pandas dataframe to be used for processing.
2. **Data Preprocessing:**
   a. Tokenization:
   i. This is splitting each tweet into individual tokens or words.
   b .Stop Words Removal:
   ii. The stop words which hold no value in an emotional context are removed from the data.
   c .Stemming:
   iii. The various forms of the words are converted into a single word.
   d . Vectorization:
   iv. The Words or tokens are converted in m*n matrices to make it easier to process for the ML models.
3. **Data Splitting:**
   The data we have is split into two categories i.e test data and training data
4. **Training:**
   The Model is then trained with the training data set.
5. **Testing:**
   The test data is applied to the model and the accuracy of the model is verified.

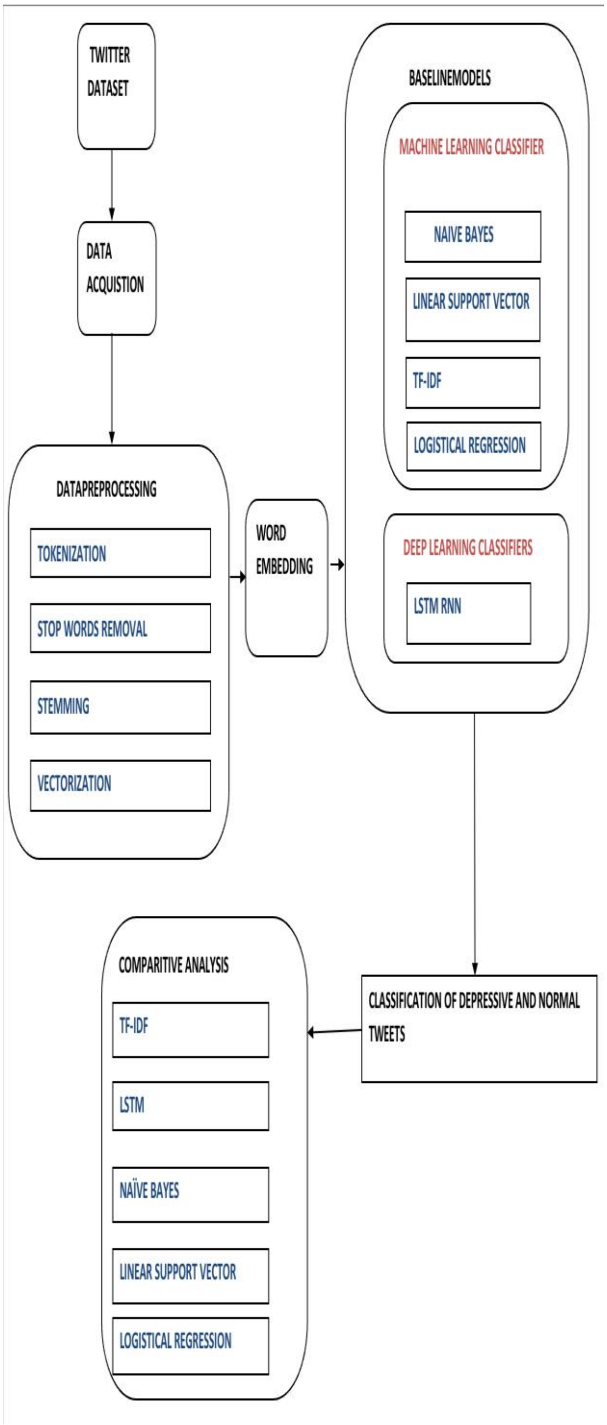
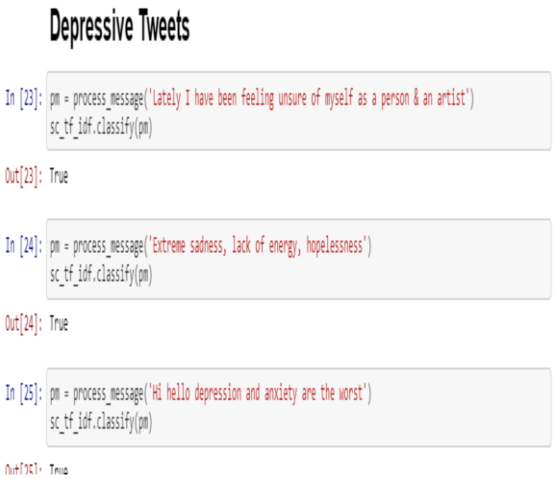


Fig 3.System Design

IV.RESULTS

1. *DETECTION OF DEPRESSIVE TWEETS:*

Using TF-IDF predictions depressive tweets have been detected.



Depressive Tweets

```
In [23]: pm = process_message('Lately I have been feeling unsure of myself as a person & an artist')
         sc_tf_idf.classify(pm)

Out[23]: True

In [24]: pm = process_message('Extreme sadness, lack of energy, hopelessness')
         sc_tf_idf.classify(pm)

Out[24]: True

In [25]: pm = process_message('Hi hello depression and anxiety are the worst')
         sc_tf_idf.classify(pm)

Out[25]: True
```

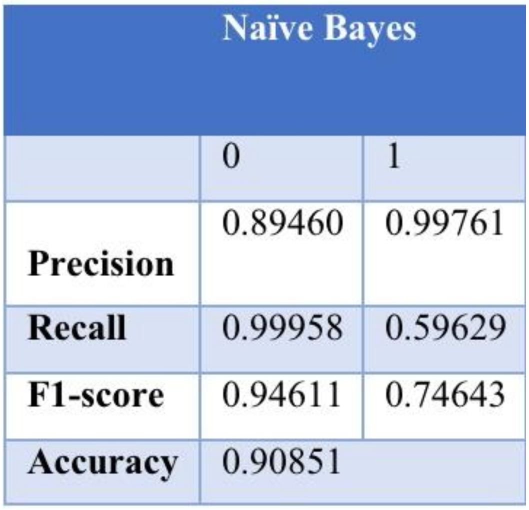
2. *COMPARATIVE STUDY:*

After running various ML models on the same datasets, with the same data preprocessing, the following results have been achieved for each Model.

The Results consists of 5 major:

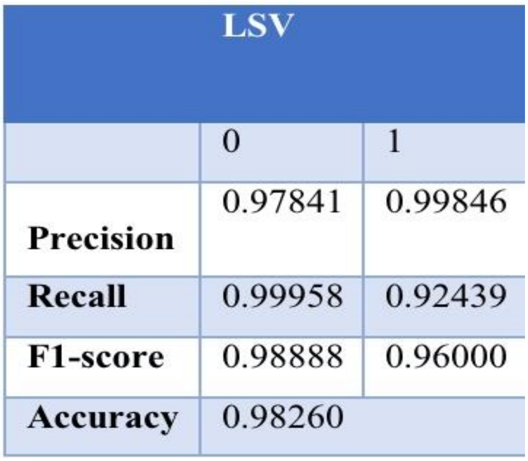
1. Precision
2. Recall
3. F1-Score
4. Support
5. Accuracy

Each of these metrics are proof to how optimised or accurate a ML model is when it comes to classification or prediction. Among all 5 methods, Long Short-Term Memory(LSTM) has the highest accuracy to detect the depressive tweets. Whereas TF-IDF has the second highest accuracy, Linear Support vector (LSV) has third highest accuracy to detect the depressive tweets.In the following tables , 0 column refers to NON-DEPRESSIVE tweets value and 1 column refers DEPRESSIVE tweets value.



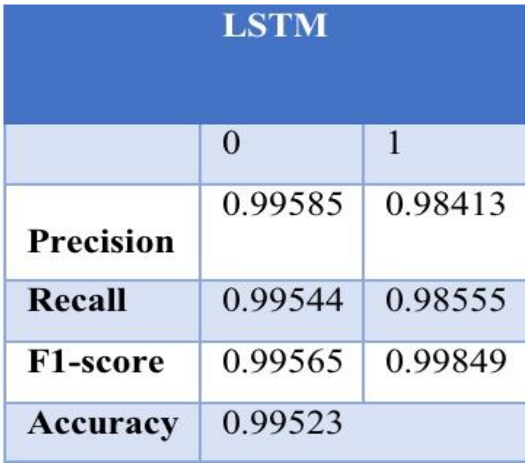
Naïve Bayes		
	0	1
Precision	0.89460	0.99761
Recall	0.99958	0.59629
F1-score	0.94611	0.74643
Accuracy	0.90851	

TABLE 1: Naïve bayes result



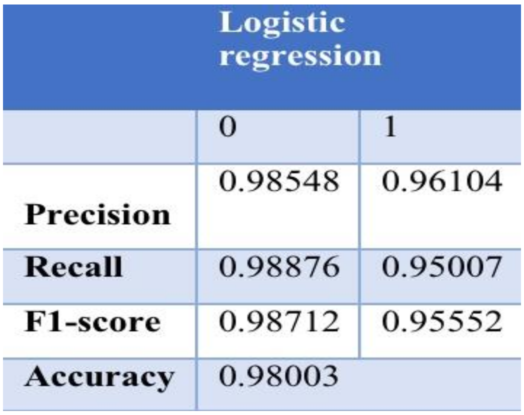
LSV		
	0	1
Precision	0.97841	0.99846
Recall	0.99958	0.92439
F1-score	0.98888	0.96000
Accuracy	0.98260	

TABLE 2: LSV result



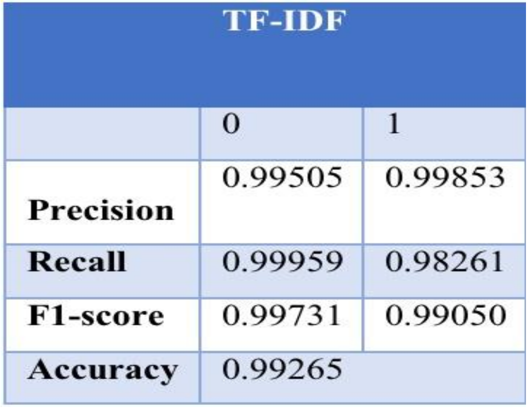
LSTM		
	0	1
Precision	0.99585	0.98413
Recall	0.99544	0.98555
F1-score	0.99565	0.99849
Accuracy	0.99523	

TABLE 3: LSTM result



Logistic regression		
	0	1
Precision	0.98548	0.96104
Recall	0.98876	0.95007
F1-score	0.98712	0.95552
Accuracy	0.98003	

TABLE 4: Logistic regression result



TF-IDF		
	0	1
Precision	0.99505	0.99853
Recall	0.99959	0.98261
F1-score	0.99731	0.99050
Accuracy	0.99265	

TABLE 5: Tf-IDF result

V. FUTURE WORK

In the future, we will be able to use more models to do analysis of tweets and more social media outlets along with emails to determine various mental health issues other than depression such as PTSD, stress and anxiety.

VI.CONCLUSION

In conclusion ,we presented a novel approach word embedding for classification tasks to detect the depressive tweets from Twitter. Also in this paper we have done a comparative analysis among five approaches say TF-IDF, Naive bayes, LSTM, Logistic Regression, Linear support vector.Among all 5 methods we have found that Long Short-Term Memory(LSTM)-RNN has the highest accuracy to detect the depressive tweets from twitter.

REFERENCES

[1] AmnaNoureen,UsmanQamar, “Semantic Analysis of Social Media and Associated Psychotic Behavior ”,Department of Computer Engineering, College of EME, National University of Sciences and Technology (NUST) H-12,Islamabad, Pakistan.

[2] RAFAEL A. CALVO1 DAVID N. MILNE1 SAZZAD M. HUSSAIN, and HELEN CHRISTENSEN, “Natural Language Processing in Mental Health applications using non-clinical texts”,School of Electrical and Information Engineering, The University of Sydney;Commonwealth Scientific and Industrial Research Organisation, CSIRO; Black Dog Institute, University of New South Wales, Australia.

[3] Mandar Deshpande and Vignesh Rao, “Depression Detection using Emotion Artificial Intelligence”,Electrical and Electronics Engineering Department Visvesvaraya National Institute of Technology, Nagpur, India.

[4] DQingCong,ZhiyongFeng,Guozheng Rao, “X-A-BiLSTM: a Deep Learning Approach for Depression Detection in Imbalanced Data”,College of Intelligence and Computing Tianjin University ,Tianjin, China.

[5] Jane H. K. Seah,Kyong Jin Shim, “Data Mining Approach to the Detection of Suicide in Social Media: A Case Study of Singapore”,School of Information Systems Singapore Management University Singapore.

[6] YevhenTyshchenko, “Depression and anxiety detection from blog posts data”, UNIVERSITY OF TARTU, Institute of Computer Science ,Computer Science Curriculum.

[7] MICHAEL M. TADESSE, HONGFEI LIN, BO XU, AND LIANG YANG, “Detection of Depression-Related Posts in Reddit Social Media Forum”,Dalian University of Technology, Ganjingzi District, Dalian, 116024, P.R. China.

[8] JT Wolohan, MisatoHiraga, “Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP”,Department of Information and Library Science Indiana University - Bloomington.

[9] Tuan D. Pham, TruongCong Thang, “Toward the Development of a Cost-Effective e-Depression Detection System”, Aizu Research Cluster for Medical Engineering and Informatics Center for Advanced Information Science and Technology,The University of Aizu, Aizuwakamatsu, Fukushima, Japan