A Mini Project Report

On

DEPRESSION ANALYSIS AND PREDICTION

Submitted in partial fulfillment of the requirement

of University of Mumbai for the Course

Natural Language Processing

In

Computer Engineering (VIII SEM)

Submitted By

Rishikesh Kadam (D17A-31)

Tanya Mohanani (D17A-44)

Nilesh Nenwani (D17A-50)

Subject Incharge

Mrs. Priya R. L.

Department of Computer Engineering

Vivekanand Education Society's Institute of Technology

Chembur - 400 074

UNIVERSITY OF MUMBAI

Academic Year 2019-20

Table of Contents

Abstract

Millennials often turn to social media forums for mental health support. Looking at comments and posts on such platforms can give insight into how people self-disclose and discuss mental health issues such as depression. Using a dataset of scraped Reddit comments, this project aims to classify depression in comments. Two types of comments were scraped- depressed and non-depressed. These comments were first analyzed to understand the basic differences between them and to understand trends. Focusing on the setting of social media, this project explores methods of machine learning and neural network architectures for identifying depression in digitally shared text entries. We developed machine learning models (logistic regression, support vector machines and multinomial naive bayes) and a BERT-based model for this classification task. It was found that the BERT-based model performed the best, with approximately 85.5% accuracy, followed by the Linear Support Vector Classifier with 84.8% accuracy.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Fundamentals

Depression detection is the interpretation and classification of textual data-points such as messages, posts, blogs and comments. Depression analysis is the study of these datasets to uncover hidden insights. Both of these tasks use various Natural Language Processing (NLP) methods and algorithms along with machine learning techniques. This project focuses on the website Reddit, which seems to fill a gap between other social media platforms such as Twitter or Facebook - which are often associated with permanent online identities - and health forums. Reddit is a unique platform in that users can choose to create "throwaway" accounts that are not associated with their main account in order to make posts or comments disclosing sensitive information.

The selected project topic comes under text classification as well as sentiment analysis. Following are some of the NLP techniques used in this area:

1. **Text Input:** Text input will be raw text or data set of topic which is going to be considered for analysis.
2. **Tokenization:** Tokenization is the process of tokenizing or splitting a string, text into a list of tokens.
3. **Stop Word Filtering**: These are the most frequently occurring words which slow down the processing of documents as these words are irrelevant. It includes articles, prepositions and other function words.
4. **Stemming**: It uses suffix list to remove or replace suffix list to remove suffixes from words. The corpus is used to remove suffixes from input documents.
5. **Lemmatization:** Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
6. **Classification:** The most common use of text classification is this of classifying a text to a class. Depending on the dataset and the reason, the classification can be a binary (positive or negative) or multi-class (3 or more classes) problem.

1.2 Objectives

- Develop a web scraper to extract forum posts and generate a dataset.
- Clean the dataset and analyze it to understand the trends in depressed people's posts.
- Use NLP techniques to clean the text in the dataset to make it useful for further analysis.
- Use machine learning models for depression detection.
- Use the BERT-based model for depression detection.

1.3 Scope

A computer has no concept of naturally spoken language, so, we need to break down this problem into mathematics which is the language of a computer. It cannot simply deduce whether something contains joy, frustration, anger, or otherwise, without any context of what those words mean. Text Classification and Sentiment Analysis solve this problem by using Natural Language Processing. Basically, it recognizes the necessary keywords and phrases within a document, which eventually help the algorithm to classify the pre-defined class of the document. We present a system that performs various tasks to analyze social media posts by depressed and non-depressed people. By doing this one can analyze or understand the trends in these posts. We also developed models to detect the depressed posts. It will help the practitioners to detect the depression easily among their patients and help them accordingly.

1.4 Organization of the Report

The report is organized as follows: The introduction is given in Chapter 1. It describes the fundamental terms used in this project. It gives a detail about text classification and depression detection. The Chapter 2 describes the review of the various papers which we surveyed. It includes different algorithms, techniques and models used for performing sentiment analysis. Chapter 3 presents the proposed work and methodology of the system. It describes the algorithms, models, dataset and modules which are used in the project. Chapter 4 includes the input and output details. It also shows the analysis and the inferences. In this chapter performance is also evaluated. The summary of the report and future scope is presented in Chapter 5.

Chapter 2

Literature Survey

2.1 Introduction

Natural language processing and machine learning have been used to perform sentiment analysis of social media posts. For example, previous work has involved building models predicting depression using the tweets of depressed Twitter users. It also has been found that Facebook status updates can reveal symptoms of major depressive episodes [1]. Yelp reviews, Amazon reviews and responses on Yahoo! answers can be classified as positive, negative and neutral. It was found that the deeper CNN worked well on user-generated data, such as Amazon reviews, and less well on text that is more carefully curated, such as responses on Yahoo! answers [2].

2.2 Literature Review

1. Yash Chetnani, Ankita Gosain, Divya Viswanath, Simran Wig, Manisha Gahirwal, “Deep Neural Network based mechanism to compute Depression in social media users” proposes the use of 3 layered Deep Neural Network architecture to analyze text and determine the emotions and habits displayed by the users. Data is collected from a micro-blogging website where users post their thoughts and day-to-day activities. The data is processed using NLTK in python. The use of regression is made to analyze the generated data efficiently [3].

2. Devakunchari Ramalingam, Vaibhav Sharma, Priyanka Zar, “Study of Depression Analysis using Machine Learning Techniques” in this paper the learning algorithm tries to find the possible and probabilistic meanings of the text via active and passive construction of all the grammatical and general features of text. They have also used semantic algorithms to find out the emotion of the user in case of depression analysis. SVM classification is used to differentiate between posts with and without depression stigma. The system using this achieved average detection accuracy of 82.2% in case of males and 70.5% in case of females [4].

3. Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu and Zhana Bao, “A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network” the sentiment analysis method proposed in the paper pays special attention to the characteristics of depression and Chinese micro-blog content, and ten features are applied in the depression

detection model. The precisions obtained from training dataset are all around 80%, and the significance of each feature in the model is also analyzed for model simplification [5].

4. Akshi Kumar, Aditi Sharma, Anshika Arora, “Anxious Depression Prediction in Real-time Social Data” uses the Twitter dataset and performs tokenization using TreebankWordTokenizer of NLTK toolkit. Three machine learning classifiers are used namely, Multinomial Naïve Bayes, Gradient Boosting and Random Forest. An ensemble vote classifier with majority voting mechanism is used to generate the final prediction. The data split for training and testing was 80 and 20.10 fold cross-validation was done [6].

5. Rafqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang and Anwaar Ulhaq, “Depression detection from social network data using machine learning techniques” The aim of this paper is to perform depression analysis on Facebook data collected from an online public source. The proposed method can significantly improve the accuracy and classification error rate. The result shows that in different experiments Decision Tree gives the highest accuracy than other ML approaches to find the depression [7].

6. Munmun De Choudhury, Scott Counts, Eric Horvitz, “Social Media as a Measurement Tool of Depression in Populations” uses a crowdsourcing methodology to build a large corpus of postings on Twitter that have been shared by individuals diagnosed with clinical depression. A probabilistic model is trained on this corpus to determine if posts could indicate depression [8]. The model leverages signals of social activity, emotion, and language manifested on Twitter.

7. Munmun De Choudhury, Scott Counts, Eric Horvitz, Aaron Hoff, “Characterizing and predicting postpartum depression from shared facebook data” has leveraged Facebook data shared by 165 new mothers as streams of evidence for characterizing their postnatal experiences. This study includes detecting and predicting onset of postpartum depression (PPD) [9].

8. Nicola J. Reavley and Pamela D. Pilkington, “Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study” reports on an exploratory study of the usefulness of Twitter for unobtrusive assessment of stigmatizing attitudes in the community. Tweets with the hashtags #depression or #schizophrenia posts on Twitter during a 7-day period were collected [10]. Statistical analysis was performed on this data.

2.3 Literature Summary

Table 2.1 Literature Survey Summary

Sr. No.	Paper	Dataset	Overview
1	Deep Neural Network based mechanism to compute Depression in social media users	Twitter API	● This paper proposes the use of 3 layered Deep Neural Network architecture to analyze text and determine the emotions and habits displayed by the users. ● Data is collected from a micro-blogging website where users post their thoughts and day-to-day activities. The data is processed using NLTK in python. ● The use of regression is made to analyze the generated data efficiently.
2	Study of Depression Analysis using Machine Learning Techniques	Twitter API, Weibo Data	● In this paper, the learning algorithm tries to find the possible and probabilistic meanings of the text via active and passive construction of all the grammatical and general features of text. ● They have also used semantic algorithms to find out the emotion of the user in case of depression analysis. ● SVM classification is used to differentiate between posts with and without depression stigma. ● The system using this achieved average detection accuracy of 82.2% in case of males and 70.5% in case of females.

3	A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network	Sina Micro-blog Open Platform API	• The sentiment analysis method proposed in the paper pays special attention to the characteristics of depression and Chinese micro-blog content, and ten features are applied in the depression detection model. • The precisions obtained from training dataset are all around 80%, and the significance of each feature in the model is also analyzed for model simplification.
4	Anxious Depression Prediction in Real-time Social Data	Twitter API	• The proposed system uses the Twitter dataset and performs tokenization using TreebankWordTokenizer of NLTK toolkit. • Three machine learning classifiers are used namely, Multinomial Naïve Bayes, Gradient Boosting and Random Forest. An ensemble vote classifier with majority voting mechanism is used to generate the final prediction. • The data split for training and testing was 80 and 20.10 fold cross-validation was done.
5	Depression detection from social network data using machine learning techniques	Facebook Data	• The aim of this paper is to perform depression analysis on Facebook data collected from an online public source. • The proposed method can significantly improve the accuracy and classification error rate. • The result shows that in different

			experiments Decision Tree gives the highest accuracy than other ML approaches to find the depression.
6	Social media as a measurement tool of depression in populations	Twitter API	• The work is done on using a crowdsourcing methodology to build a large corpus of postings on Twitter that have been shared by individuals diagnosed with clinical depression. • A probabilistic model is trained on this corpus to determine if posts could indicate depression. • The model leverages signals of social activity, emotion, and language manifested on Twitter.
7	Characterizing and predicting postpartum depression from shared facebook data	Facebook Data	• They leveraged Facebook data shared by 165 new mothers as streams of evidence for characterizing their postnatal experiences. • This study includes detecting and predicting onset of postpartum depression (PPD).
8	Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study	Twitter API	• The paper reports on an exploratory study of the usefulness of Twitter for unobtrusive assessment of stigmatizing attitudes in the community. • Tweets with the hashtags #depression or #schizophrenia posted on Twitter during a 7-day period were collected. • Statistical analysis was performed on this data.

Chapter 3

Implementation Details

3.1 Overview

The Depression has been shown to affect the language of individuals [1]. This project aims to use natural language processing, machine learning techniques, and neural network architectures to build, tune, and evaluate models that classify Reddit text comments as "depressed" or "non-depressed."

3.1.1 Existing Methodology and Systems

Most of the depression self-diagnostic tools available online are in the form of questionnaires. While these are designed by psychologists and can be highly accurate, it is often the case that users cheat the system. As has been reported by psychologists interviewed by the authors, patients tend to subconsciously hide their real intentions and feelings when answering these questionnaires. While systems have been developed to analyze text for emotions exhibited, the informal nature and imprecise use of grammar has been a hurdle in using traditional NLP and Sentiment Analysis tools like the Stanford CoreNL. Lack of proper grammar is the root of these troubles. These systems also only use crisp binary classification. They also fail to connect the emotions with their subjects.

In recent years, a lot efforts have been made to use machine learning techniques to analyze the social media content to detect depression among users. The most used datasets for this purpose come from FaceBook and Twitter. Although these datasets are custom made, the posts or comment may not always reflect depression-like characteristics. Moreover, these datasets, when downloaded from websites like Kaggle, tend to be older. These machine learning or deep learning models developed using these datasets may give poor performance when deployed and made to work on real world data.

3.1.2 Proposed Methodology and System

To tackle the problem with datasets mentioned in the previous section, we searched for online forums where we could find trustworthy datasets. One of the forums is BeyondBlue. It is a website for support of depressed individuals. People post their issues and thoughts on the forums. Other people post comments trying to solve the issue or to offer mental support and encouragement. These posts could have been scraped and labeled as “depressed”, but this gives rise to a major issue-absence of control group or non-depressed posts which we need to make the dataset exhaustive. The

solution to this problem is Reddit.

Reddit is a forum which basically consists of different topic-threads, called subreddits. People post their opinions, stories and questions under appropriate subreddit and other people comment on the posts. It has a subreddit “Depression” dedicated to depressed people who ask others for help and support. This is a good source for depressed-data. It also has another subreddit called “AskReddit”. This subreddit is used by the users to post the questions which are generally considered as funny, but also contains recent happenings around the world, trends and people’s own experiences. This thread is the source for non-depressed-data. Using BeyondBlue for depressed data and AskReddit for neutral data would not be ideal as these two forums are completely different and have different norms and trends followed by the users. So the two categories would have been distinguished easily by our models.

The system we propose uses the Reddit API and the PRAW library which are used to develop a scraper to generate the dataset. Then the dataset is cleaned to make it suitable for analysis. Initially, the basic analysis is performed to understand trends and to compare posts under depressed and non-depressed categories. In the last stage two types of models are developed- machine learning models and a BERT-based model to predict if an input post/comment is by a depressed user or not.

We have used following modules for the System:

- NLTK- The Natural Language Toolkit for Python
- Word Tokenizing techniques
- Preprocessing data for NLP
- Building and training models using Scikit-Learn
- Building a BERT-based model
- Training the model using TensorFlow

3.2 Implementation Details

3.2.1 Methodology

The developed system has following stages:

1. Building a scraper and data collection:

We developed a scraper using PRAW library which uses the Reddit API. Reddit API provides access to all posts and comments which can be scraped and stored into files. We scraped comments from two threads- Depression (label=1) and AskReddit (label=0). We used 5000 comments from each category to develop the dataset.

2. Basic data cleaning:

Initially, we removed the user tags, hashtags, numbers, URLs from the data. We also removed the comments which had no textual content. One of the important steps was to remove the comments with body- [‘removed’] or [‘deleted’].

3. Basic trend study and comparative analysis:

We added a few columns to the dataset such as lexical diversity, word count, character count, stopwords count, etc. These features were studied by generating histograms to find visible distinctions between depressed and non-depressed content. WordClouds were also generated for the same purpose.

4. Advanced data cleaning:

To make the data suitable for machine learning models, further cleaning was done. The stopwords were removed using NLTK library. The words which appeared too frequently or too rarely were removed. The techniques such as stemming, lemmatization and n-grams were also used.

5. Machine Learning Models:

We developed three machine learning models- Linear Regression, Multinomial Naive Bayes and Linear Support Vector Classifier. We trained these models using different values for parameters such as learning rate. The values which gave the best results were selected. The confusion matrices were also developed and studied.

6. BERT-based model:

We developed a Bidirectional Encoder Representations from Transformers (BERT)-based model, which is a new language representation model as described in [5]. As the name suggests, it was

designed to pre-train deep bidirectional representations that can be fine-tuned with an additional output layer. For this project, this output layer - a pooled output - was used for the binary classification of the comments. From the many pre-trained models available, we chose the English-language uncased (all lowercase before tokenization) model of BERT, as case information is not particularly important to the task of social media comment classification. For this model, we used a dropout probability of 0.2, learning rate of 2e-5, batch size of 32 and 3 epochs.

3.2.2 Details of packages

- PRAW library is specifically designed for scraping the Reddit website.
- The Reddit API was used to get the comments from selected subreddits.
- NLTK was used for NLP techniques and preprocessing.
- Scikit-Learn was used to develop machine learning models.
- Numpy, Pandas, MatPlotLib and Seaborn were used for data analysis and visualizations.
- TensorFlow and BERT were used to develop and train BERT-based model

Chapter 4

Project Inputs and Outputs

4.1 Inputs Details

The dataset generated using the scraper was used as input. It consists of the label i.e. 1 for depressed and l2 for non-depressed and username, date and the text of the comment or post.

We have used the label and the text columns to train our model. The text is first processed by removing the stop words and performing stemming to get the root form of the words. The data is then divided into train and test data and fed to the models.

4.2 Evaluation Parameters

The performance of the machine learning models is evaluated on the testing dataset. This provides an estimate of the performance of the network at making predictions for unseen data in the future. The parameters such as accuracy, precision, recall, F-1 score and the confusion matrix were considered to evaluate the performance. Although, the maximum importance was given to accuracy and confusion matrix. The accuracy is the number of correct predictions divided by the total number of predictions. The BERT-based model was evaluated using the loss, classification accuracy and the confusion matrix.

4.3 Basic Analysis and Inferences

4.3.1 Lexical Diversity, Word Count and Stopword Count

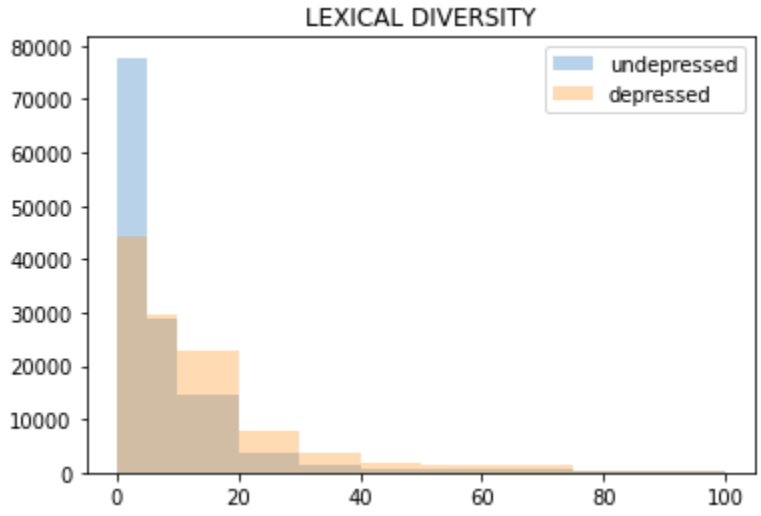

Fig 4.1 Lexical Diversity

As seen Fig 4.1, depressed people tend to use a broader vocabulary of words. Nearly 50% of the non-depressed comments had lexical diversity ratio less than 5. This was not the case for depressed people.

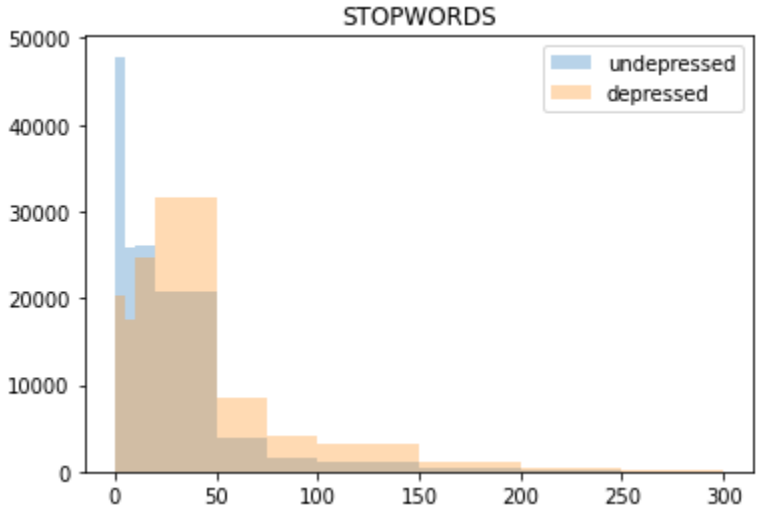

Fig 4.2 Stop Words

As seen in Fig 4.2, depressed people use a lot more stopwords in their comments.

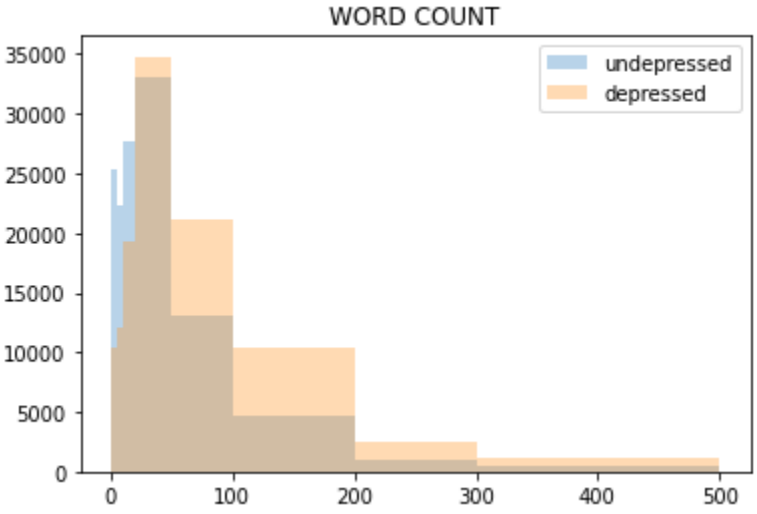

Fig 4.3 Word Count

Fig 4.3 clearly illustrates that depressed people tend to use more words as compared to their counterparts. This tendency arises from their fear that they might be making themselves and also from their need to be understood and well-perceived.

4.3.2 WordClouds

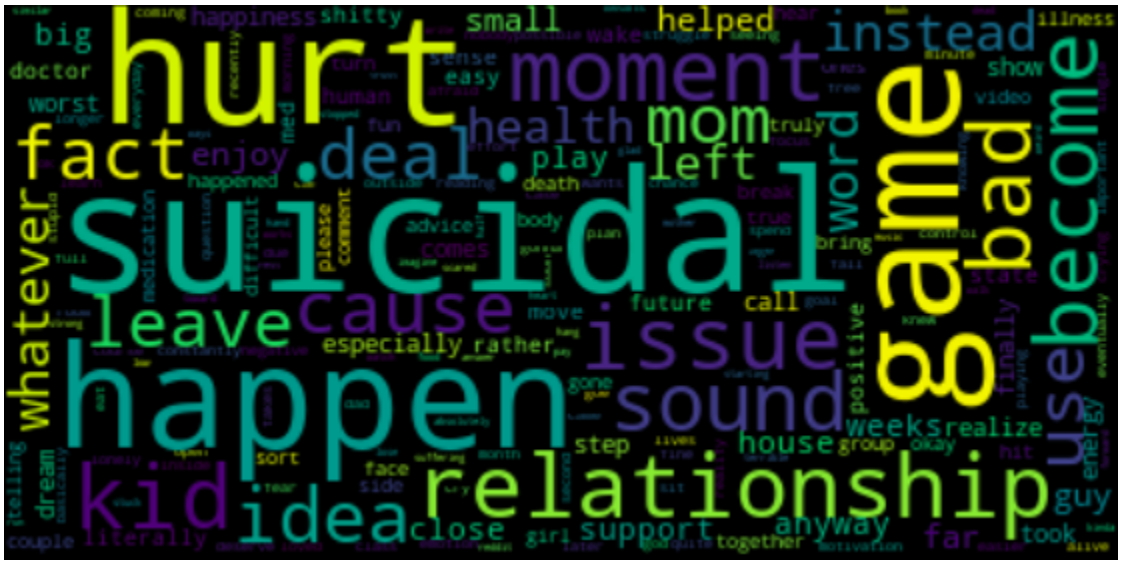

Fig 4.4 Wordcloud of depressed comments

Fig 4.5 Wordcloud of non depressed comments

Fig 4.4 and Fig 4.5 show word clouds of depressed and non-depressed comments, respectively. The difference in dominant words can be seen clearly. The first word cloud has words such as hurt, suicidal, bad, mom, kid, left, etc whereas the second one has lighter words such as food, movie, weel, show, cool, etc.

4.4 Machine Learning Models

4.4.1 Linear Regression

Fig 4.6 shows the performance details of the linear regression model with learning rate 1. Fig 4.7 shows the confusion matrix visualization of the same.

```
Acc  0.8470514560066799
Precision  0.8534710431993964
Recall  0.8109064838463951
F1 score  0.831644493462926
              precision    recall  f1-score   support

           0       0.84      0.88      0.86     25588
           1       0.85      0.81      0.83     22317

    accuracy                           0.85     47905
   macro avg       0.85      0.84      0.85     47905
weighted avg       0.85      0.85      0.85     47905
```

Fig 4.6 Performance of linear regression

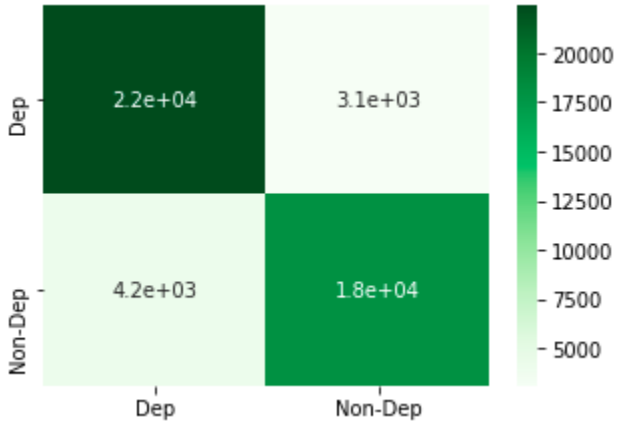

Fig 4.7 Confusion matrix of linear regression

4.4.2 Linear SVC

Fig 4.8 shows the performance details of the linear support vector classifier with regularization parameter 0.25. Fig 4.9 shows the confusion matrix visualization of the same.

```
Acc  0.8486170545872038
Precision  0.8514768326256358
Recall  0.8176726262490478
F1 score  0.8342324220535796
              precision    recall  f1-score   support

           0       0.85      0.88      0.86     25588
           1       0.85      0.82      0.83     22317

    accuracy                           0.85     47905
   macro avg       0.85      0.85      0.85     47905
weighted avg       0.85      0.85      0.85     47905
```

Fig 4.8 Performance of SVC

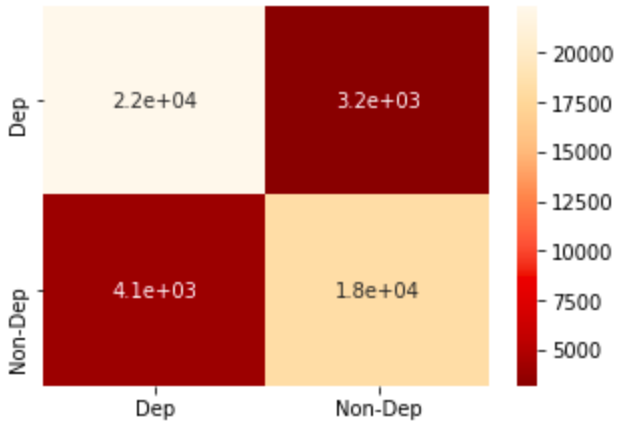

Fig 4.9 Confusion matrix of SVC

4.4.3 Multinomial Naive Bayes

Fig 4.10 shows the performance details of the multinomial naive bayes model. Fig 4.11 shows the confusion matrix visualization of the same.

```
Acc  0.8271579167101555
Precision  0.7794251373897917
Recall  0.8795003345973679
F1 score  0.826444202230234
              precision    recall  f1-score   support

           0       0.88      0.78      0.83     25490
           1       0.78      0.88      0.83     22415

    accuracy                           0.83     47905
   macro avg       0.83      0.83      0.83     47905
weighted avg       0.83      0.83      0.83     47905
```

Fig 4.10 Performance of multinomial naive bayes

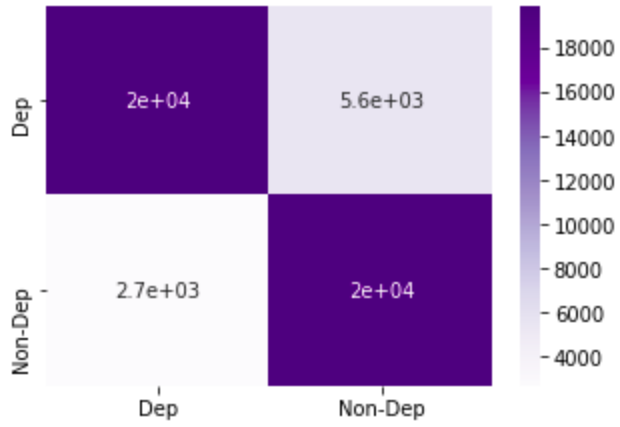

Fig 4.11 Confusion matrix of multinomial naive bayes

As seen in the above figures, the best performance was given by linear support vector classifier with testing accuracy of 84.8%.

4.5 BERT-Based Model

As seen in Fig 4.12, the training took much less time because of the use of GPU implementation of TensorFlow library.

```
INFO:tensorflow:Loss for final step: 0.021613998.
INFO:tensorflow:Loss for final step: 0.021613998.
Training took time  0:07:44.644841
```

Fig 4.12 Training completion

```
{'auc': 0.8545642,
 'eval_accuracy': 0.8552,
 'f1_score': 0.8456948,
 'false_negatives': 367.0,
 'false_positives': 357.0,
 'global_step': 468,
 'loss': 0.5396407,
 'precision': 0.84750104,
 'recall': 0.8438962,
 'true_negatives': 2292.0,
 'true_positives': 1984.0}
```

Fig 4.13 Performance of BERT model

The evaluation accuracy achieved was 85.5% as seen in Fig 4.13. This was better than the performance of machine learning techniques. Fig 4.14 shows the visualization of the confusion matrix.

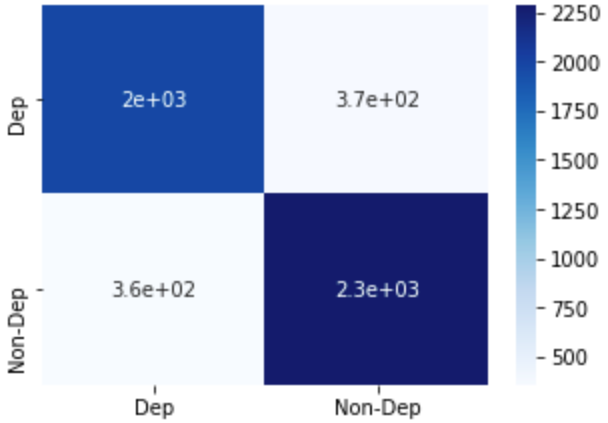

Fig 4.14 Confusion matrix of BERT model

```
pred_sentences = [
  "I had a black coffee last night to complete the assignments!",
  "I'm don't feel good... What should I do?"
]

predictions = getPrediction(pred_sentences)

for i in predictions :
    print(i[0],":",i[2])

I had a black coffee last night to complete the assignments! : Non-depressed
I'm don't feel good... What should I do? : Depressed
```

Fig 4.15 Prediction using BERT model

The Fig 4.15 shows two sample inputs given to the trained model. The model predicts the depression correctly

Chapter 5

Summary and Future Scope

5.1 Summary

We generated our own dataset by scraping comments from two subreddits- Depression and AskReddit. Basic analysis inferred that depressed people tend to use more words, stopwords and alss tend to use a broader vocabulary as compared to normal people. We used three machine learning models- linear regression, multinomial naive bayes and linear support vector classifier. Linear SVC performed the best with accuracy of 84.8%. Our BERT-based model gave accuracy of 85.5%. These results are acceptable.

5.2 Future Scope

A more robust dataset can be generated by scraping comments from more than one normal subreddits as well as by including depressed comments from more than one forum. Our models were limited to Reddit forums. LSTM and CNNs with word-embeddings can also be used to test if they perform better than the BERT model.

References

[1] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences”. In: CoRRabs/1804.07000 (2018).

[2] Xiang Zhang, Junbo Zhao, and Yann LeCun. “Character-level Convolutional Networks for Text Classification”. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. NIPS’15. Montreal, Canada: MIT Press, 2015, pp. 649–657.

[3] Yash Chetnani, Ankita Gosain, Divya Viswanath, Simran Wig, Manisha Gahirwal, “Deep Neural Network based mechanism to compute Depression in social media users”. International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 03, March 2018.

[4] Devakunchari Ramalingam, Vaibhav Sharma, Priyanka Zar, “Study of Depression Analysis using Machine Learning Techniques.” International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8, Issue-7C2, May 2019.

[5] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu and Zhana Bao. "A depression detection model based on sentiment analysis in micro-blog social network." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2013.

[6] Akshi Kumar, Aditi Sharma, Anshika Arora, “Anxious Depression Prediction in Real-time Social Data.” Proceedings of International Conference on Advanced Engineering, Science, Management and Technology (ICAESMT19), 2019.

[7] Rafqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang and Anwaar Ulhaq. "Depression detection from social network data using machine learning techniques." Health information science and systems 6.1, 8, 2018.

[8] Munmun De Choudhury, Scott Counts, and Eric Horvitz. "Social media as a measurement tool of depression in populations." Proceedings of the 5th Annual ACM Web Science Conference, 2013.

[9] Munmun De Choudhury, Scott Counts, and Eric Horvitz, Aaron Hoff. "Characterizing and predicting postpartum depression from shared facebook data." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014.

[10] Nicola J. Reavley and Pamela D. Pilkington. "Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study." PeerJ 2: e647, 2014.

Acknowledgement

This project was possible due to complete guidance and support from our NLP subject guide and teacher Prof. Priya RL. The project has been thoroughly administered by our guide to make sure it is feasible, completed in time and successfully achieving desired goals.

The Head of Department for Computer Engineering of V.E.S.I.T. -- Dr. Nupur Giri has been very supportive of research and development in this field and we thank the department for the same.

We are deeply indebted to our principal Dr. (Mrs.) J.M. Nair for giving this valuable opportunity to us to do this project.

Rishikesh Kadam (D17A/31)

Tanya Mohanani (D17A/44)

Nilesh Nenwani (D17A/50)