A Thesis/Project/Dissertation Report

on

PREDECTING-DEPRESSION-FROM-SOCIAL-MEDIA-NETWORKING-DATA-USING-MACHINE-LEARNING

Submitted in partial fulfillment of the requirement for the award of the degree of

Master of Computer Applications

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Name of Supervisor:
Designation

Submitted By

Kushagra Singh (18SCSE1010071)
Anushka Singh (18SCSE10101556)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING / DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitle **"PREDECTING DEPRESSION FROM SOCIAL MEDIA NETWORKING DATA USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Name… Designation, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Kushagra Singh (18SCSE1010071)
Anushka Singh (18SCSE10101556)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Ms. Vaishali Gupta

Associate Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Kushagra Singh 18SCSE1010071 Anushka Singh 18SCSE1010071 has been held on PREDECTING DEPRESSION FROM SOCIAL MEDIA NETWORKING DATA USING MACHINE LEARNING and his/her work is recommended for the award of Bachelor of Technology

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

Acknowledgment

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

We are highly indebted to our guide Ms. Vaishali Gupta, Associate Professor for his guidance and constant supervision as well as for providing necessary information regarding the project & also for motivating and encouraging every time. Without his encouragement and guidance this project would not have materialized.

The guidance and support received from all the members who contributed and who are contributing to this project, was vital for the success of the project.

Kushagra Singh (18SCSE1010071)
Anushka Singh (18SCSE10101556)

Abstract

In this era of COVID-19 pandemic, as more people self-isolate themselves, psychological health issues like depression, anxiety, and stress is an increasing concern all over the world. The purpose of this study is to investigate the data from social forums, where we found communities of depressed people sharing their thoughts and emotions in the forums, these forums also receive advices and support. In this paper, we will analyze the "depressed" text; by manipulating the data, extracting features, categorizing, and try to understand what are the attributes of "depressed" text, and how we can "predict" whether a text should be marked as depressed or not. Using text analysis and text data mining techniques, the text obtained from the social forums was analyzed and three different machine learning algorithms were used to predict depression. After cross validation overall accuracy of 99.69% was obtained as the best score using the proposed system. This study definitively answers the question regarding using human basic language and communication of personal experiences, for the prediction of depression and can be reached easily. Furthermore, not only actions, habits and behavior of a person, text too can be used for accurate diagnosis of depression.

KEYWORDS: NLP, Machine learning, text mining, artificial intelligence, social forums

Table of Contents

List of Table

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

CHAPTER-1

Introduction

1.1 INTRODUCTION

Depressive disorders are one of the top three causes for YLDs (years lived with disability) for both sexes combined [1]. Depressive disorders are a type of mental illness. Early diagnosis of mental illness has been given importance all over the world. Diagnosis of depression can be done with help of psychological treatments like CBT (cognitive behavioral therapy), behavioral activation and interpersonal psychotherapy [2]. In this research paper, we have introduced a way to diagnose depression with the help of efficient text mining, and machine learning techniques for accurate prediction of depression. Machine learning can be applied to recognize different illnesses including depression [3].

Mental illness also refers to emotional check, emotionally unstable people tend to have negative thoughts and they get discouraged or bored easily with their work and their decisions in life. In many cases lack of support from family and peers leads to loneliness and self-isolation. There are a lot of social, psychological, emotional, cognitive reasons for depression in different age groups of people. Depression is not an age specific illness. Students, adults, and older people all around the world have been diagnosed with depression.

We live in a general public where because of social disgrace and humiliation individuals are still awkward discussing their emotional well-being issues. In addition, the majority of the time individuals endure such ailment yet don't think about it since they can't perceive the manifestations. A huge extent of individuals doesn’t approach to emotional wellness care, and other people who approach frequently wonder whether or not to take mental wellbeing tests. Because of their dithering of taking psychological wellness exams, they never know whether they are having such issues and thus are left untreated.

Social media can be used as a tool to gather data from and use it for prediction and detection of mental illness [4]. Analyzing the features required for creating machine learning models for depression detection is made efficient with the help of text mining. NLP (natural language processing) is a machine learning technique which is used to discover patterns in languages. Over the years with several in depth researches on the topic of depression, it has been discovered that patterns can be found in languages related to people's emotions and psychology [5, 4]. The basic idea of finding patterns can be extended to different linguistic scenarios also [6]. There have also been other advances in research in this field which include data other than language data, researchers have used sensors to get brain waves data [7], even imaging [8] has been employed as a method for analysis and prediction of sentiments.

For treatment of depression self-help groups were formed to find ways to counter depression by informal association of people who choose to come together and they vent their problems to the group with the expectation of finding solution of those problems from the self-help group. Now due to the social distancing norms as a result of covid crisis, and due to increasing use of social media, social discussion forums have been created to address problems with the expectation to finding solutions from the social forums. On these social forums people write up about their bad experiences, or their situations in life which is troubling them. These social forums encourage people to seek support, and raise awareness about depression, anxiety, mental illness and suicide prevention. To build a depression detection system, language data is required, we have collected data from various suicidal/ depression forums. For this research, we used the data from Beyond Blue social forums website. In categorical approach, sentiment is described as negative or positive [9]. The dimensions which we can find in languages with the help of NLP affect underlying stimuli that influence sentiment responses.

1.2 FORMULATION OF PROBLEM

In the modern-day scenario with the expanded minds of youngsters with more and more minds indulging into social media the advancement has been immense but as a coin has two sides this also comes with major drawbacks.

The modern-day generation of about 13-66% deals with depression related issues some way or other. Depression is a constant feeling of sadness and loss of interest, which stops you doing your normal activities. Different types of depression exist, with symptoms ranging from relatively minor to severe. Abuse. Physical, sexual, or emotional abuse can make you more vulnerable to depression later in life.

Age. People who are elderly are at higher risk of depression. With the use of modern technologies like Machine Learning and NLP we can formulate the depression detection strategies. Artificial Intelligence comprises of many subfields and one of them is ML, with ML algorithm we train data inputs and then analysis of the statistical data is done to provide an output of a specific range. Another aspect of AI is Natural Language Processing, this on the other hand allows interaction between humans and system. The major areas of NLP play role in speech recognition, sentiment analysis, automatic text summarization, machine translation etc.

There's no single cause of depression but what the data from social media shows is mind numbing, "What we found overall is that if you use fewer social media, you are actually less depressed and less lonely, meaning that the decreased social media use is what causes that qualitative shift in your well-being," said Jordyn Young, a co-author of the paper and a senior at the University of Pennsylvania.

1.2.1 Tool and Technology Used

- **Tools**
    - **SOFTWARE REQUIREMENTS**
        - Google colab
        - Python 3 or Higher
        - Jupiter Notebook
    - **HARDWARE REQUIRMENT**
        - 2GB RAM (min)
        - 200GB storage
- **Technology**
    - Machine Learning
    - Artificial Intelligence
    - Data Science

CHAPTER-2

Literature Survey/Project Design

2.1 Literature Survey

To understand the correlates of depression in individuals, there have been various studies in various fields of psychology, sociolinguistics and medicine. There has been a study about personality levels, and how personality development can prevent future episodes of depression [15]. On the other hand, negative cognitive styles were identified as a paradigm to predict the onset and duration of depression [16]. There have been studies with regard to particular age groups, and how they have a variety of symptoms according to their age [17]. Unions beyond personality and negative cognitive styles have also been researched upon, it was found out that relations of time perspective affected people gravely. Future worries were the most frequent mode of distress as studied in [18].

Examination of speech could also help in the diagnosis of depression [19]. In spite of the fact that reviews to date have improved our comprehension of components that are connected to mental problems, a striking constraint of earlier examination is that it depends intensely on little, frequently homogeneous examples of people, who may not fundamentally be illustrative of the bigger populace. Further, these examinations regularly depend on studies, depending on review self-reports about disposition and perceptions about wellbeing: a strategy that limits worldly granularity. That is, such appraisals are intended to gather undeniable level outlines about encounters throughout significant stretches of time.

2.2 Project Design

1. The first stage of our project is to gather data

   How we collected the data, and what kind of data we focused on getting. The data we collected are posts from a forum of depression and assistance called Beyond Blue which spreads awareness about mental health issues. To understand what the attributes of depressed text are, we need to compare it with undepressed text, which we found on GitHub.

2. Basic feature extraction using text data

   To gather insights from the cleaned data, we had to do some manipulations on the text, e.g., word count, char count, average word length, counting the stop words, the lexical diversity of the posts, the date that the post was written, etc.

3. Basic Text Pre-processing of text data

   First, we have raw text, that is filled with tags, numerical values, punctuations, misspellings, common non-sensical txt (/n), upper letters, spam posts, links, Frequent words removal, rare words removal, and a lot of stop words that adds noise to the context, we will need to remove them.

4. Advance Text Processing

   We used several methods such as - Tokenization, stemming, lemmatization, n-grams, term frequency (TF), inverse document frequency (IDF) and TF-IDF, part of speech tagging, sentiment analysis, topic modeling, principal component analysis, and for predicting whether a text is labeled as depressed, we built a predictor using several methods e.g., support vector machine, multinomial Naïve Bayes, and K-Nearest-Neighbors. All those methods helped us build a model, gather more insights on our data, and adjusting it the best, so our model wont overfit the data.

	LINE_NUM	WORD_COUNT	CHAR_COUNT	STOPWORDS	HASHTAGS	UPPER	NUMERICS	AVG_WORDS_LEN	LEXICAL_DIVER
count	3579.000000	3579.000000	3579.000000	3579.000000	3579.000000	3579.000000	3579.000000	3579.000000	3579.000000
mean	4.377759	251.824811	1321.981442	117.716960	0.000559	14.404023	1.444258	5.262153	31.420659
std	6.853601	164.344691	882.489163	78.463127	0.023636	11.545905	1.987920	0.347010	16.741107
min	0.000000	8.000000	42.000000	2.000000	0.000000	0.000000	0.000000	2.980287	2.470588
25%	0.000000	128.000000	674.000000	59.000000	0.000000	6.000000	0.000000	5.044603	18.782517
50%	3.000000	217.000000	1144.000000	102.000000	0.000000	12.000000	1.000000	5.231293	29.200000
75%	6.000000	363.000000	1914.000000	170.000000	0.000000	21.000000	2.000000	5.430083	43.105336
max	136.000000	1920.000000	9399.000000	1033.000000	1.000000	106.000000	37.000000	9.210526	153.262295

Table 1: Features after text data mining.

5. Targets and measurements

   Our main target is to understand what makes depressed text marked as "depressed", what are the attributes that makes it unique, to see if there is a difference between depressed and undepressed, and if we can "predict" whether a text is tagged as depressed or not. We will need to extract all the measurements we can from each dataset, and examine the difference and similarity between each measurement. The measurements we look at, are

	LINE_NUM	WORD_COUNT	CHAR_COUNT	TAG	STOPWORDS	HASHTAGS	UPPER	NUMERICS	AVG_WORDS_LEN	LEXICAL_DIVER
LINE_NUM	1.000000	0.304540	0.336998	0.040934	0.293752	0.010774	0.237253	0.131274	0.259939	0.297162
WORD_COUNT	0.304540	1.000000	0.995298	0.040611	0.985550	0.005493	0.770452	0.450092	-0.055764	0.963722
CHAR_COUNT	0.336998	0.995298	1.000000	0.044229	0.976868	0.006787	0.755202	0.448212	0.018803	0.961572
TAG	0.040934	0.040611	0.044229	1.000000	0.040384	-0.001047	0.017078	0.041013	0.019794	0.037697
STOPWORDS	0.293752	0.985550	0.976868	0.040384	1.000000	-0.000065	0.721612	0.426544	-0.089640	0.962001
HASHTAGS	0.010774	0.005493	0.006787	-0.001047	-0.000065	1.000000	0.001221	0.000663	0.004185	-0.002050
UPPER	0.237253	0.770452	0.755202	0.017078	0.721612	0.001221	1.000000	0.304342	-0.148290	0.738994
NUMERICS	0.131274	0.450092	0.448212	0.041013	0.426544	0.000663	0.304342	1.000000	-0.018163	0.366542
AVG_WORDS_LEN	0.259939	-0.055764	0.018803	0.019794	-0.089640	0.004185	-0.148290	-0.018163	1.000000	-0.044370
LEXICAL_DIVER	0.297162	0.963722	0.961572	0.037697	0.962001	-0.002050	0.738994	0.366542	-0.044370	1.000000

Table 2: Description of the parameters of the text.

CHAPTER-3

Functionality/Working of Project

3.1 Basic feature extraction

As we can see, there is a clear difference between depressed and undepressed text, looking at the lexical diversity, we can see that depressed people have larger vocabulary, additionally, they use more words, hence, got more stop words.

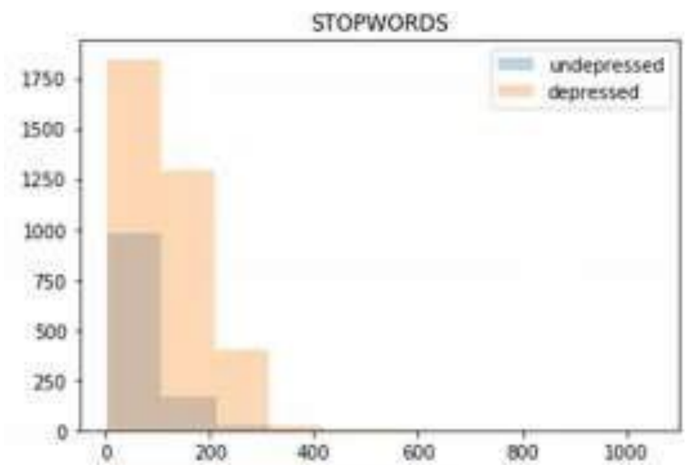

Figure Count of depressed and undepressed stop words.

.

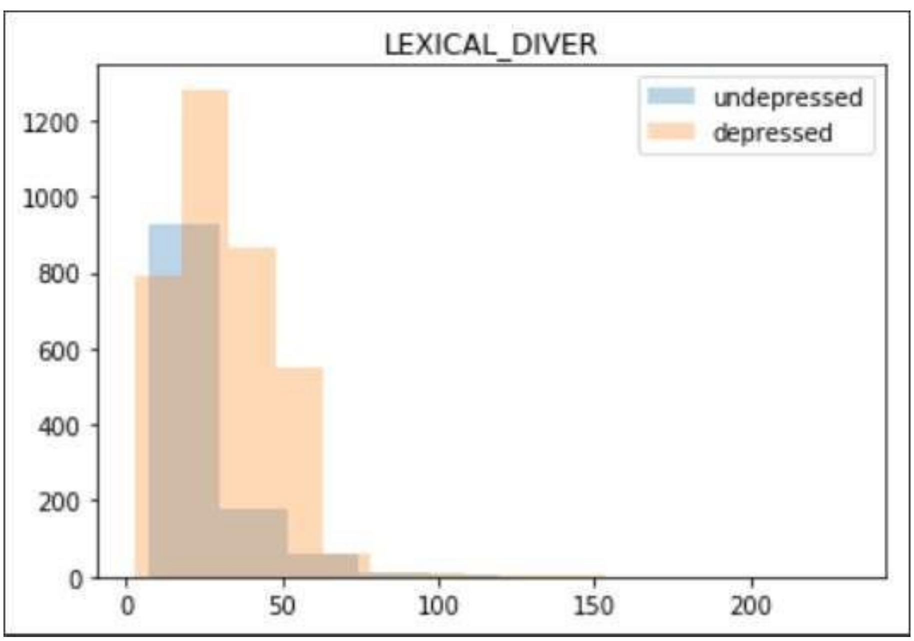

Figure 2: Lexical diversity of depressed and undepressed text.

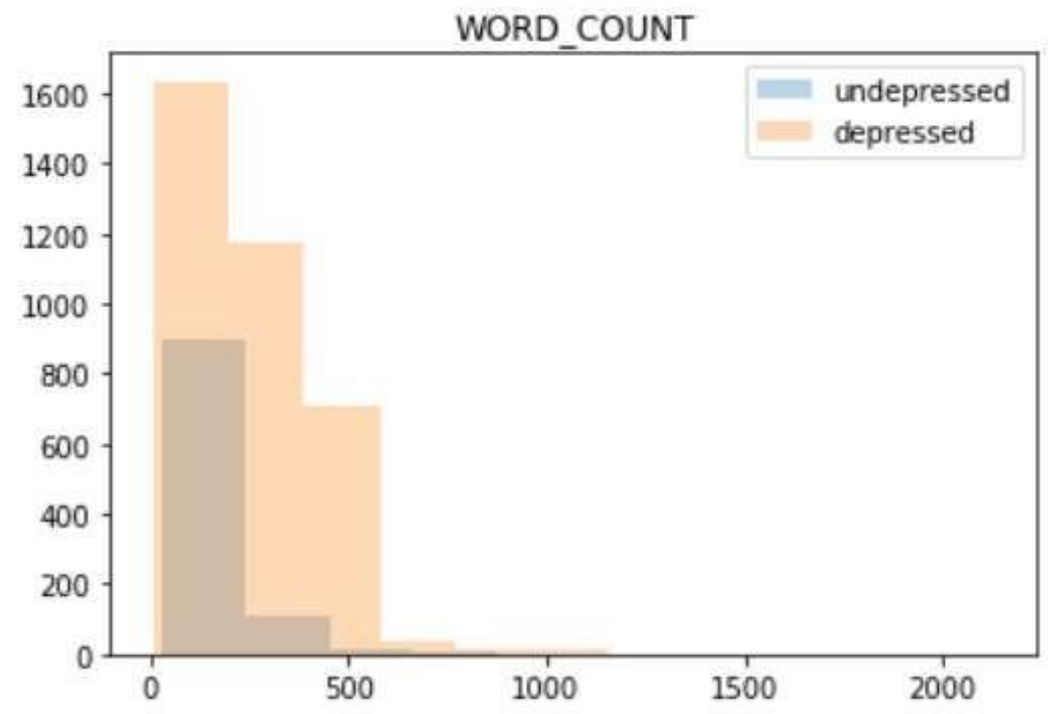

Figure 3: Word count of depressed and undepressed text.

3.2 Analyzing the date

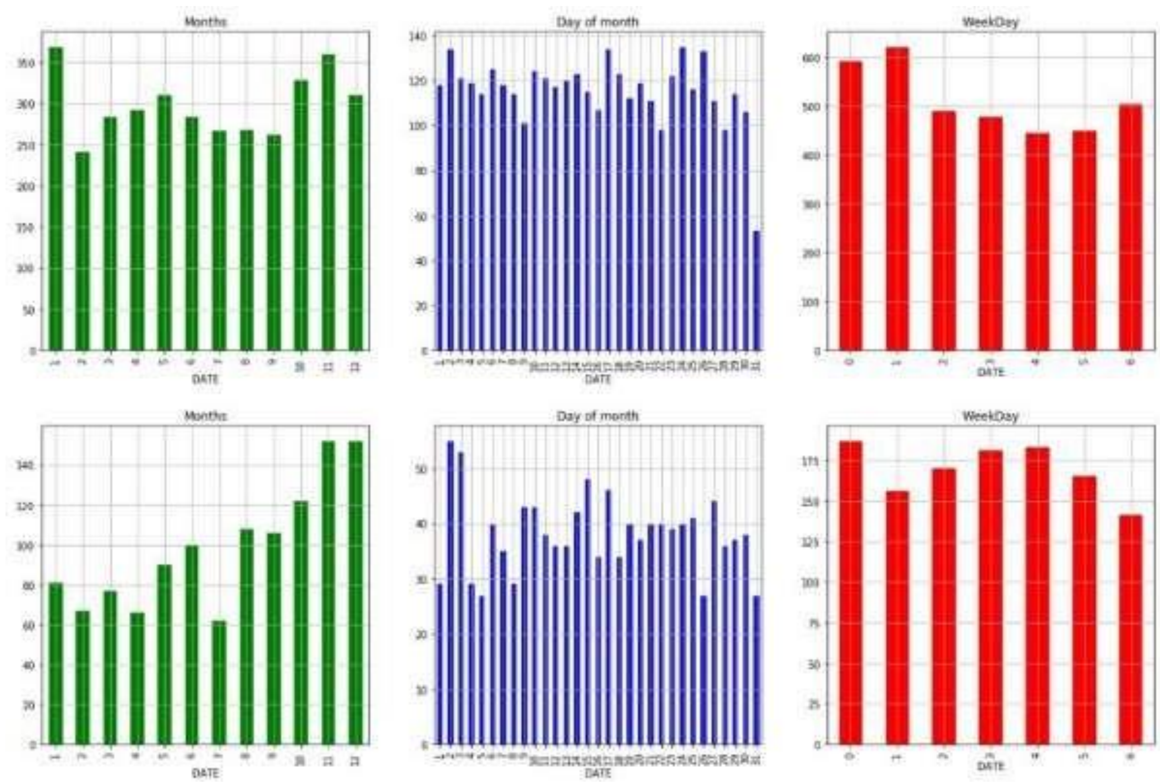

Figure 4: Graphs of analysis of posts according to their dates.

Top row- depressed posts, bottom row- neutral posts. In the graphs above, we noticed some interesting info about the time the posts were written, looking at the month, we noticed that depressed people write more during the end of the year and the first month, and during the year have "ups and downs", while undepressed writer's activity is weak during the entire year, and gets stronger in the end of the year. By looking at the weekday, where Sunday is 6, depressed people are more active in the start of the week (Sunday, Monday, Tuesday), while in undepressed posts, we see that the activity is high at the end of the week, and Sunday.

1. **Bi- grams**

   By looking at Bi-grams, we could "sense" the text and get the main idea, we had to look manually for bi-grams that contribute our understanding, and the strong bi-grams were not very helpful. Adding bi-grams to our 3d projection, and topic modeling, we noticed that it groups the posts (after dimensional reduction), and made the clustering worse, we decided not to use this feature LDA and PCA.

   Statistical language models, in its essence, are the type of models that assign probabilities to the sequences of words. You can think of an N-gram as the sequence of N words, by that notion, a 2- gram (or bigram) is a two-word sequence of words like "please turn", "turn your", or" your homework", and a 3-gram (or trigram) is a three-word sequence of words like "please turn your", or "turn your homework". As the name suggests, the bigram model approximates the probability of a word given all the previous words by using only the conditional probability of one preceding word. In other words, you approximate it with the probability: P (the | that). Now, that we understand the underlying base for N-gram models, you'd think, how can we estimate the probability function. One of the most straightforward and intuitive ways to do so is Maximum Likelihood Estimation (MLE). The N-gram model, like many statistical models, is significantly dependent on the training corpus. As a result, the probabilities often encode particular facts about a given training corpus. Besides, the performance of the N-gram model varies with the change in the value of N.

403	suicide,or
460	hard,to
379	to,feel
327	my,depression
409	the,past

5514	I,have
4374	I,am
3101	and,I
2923	I,feel
2606	I,was
2439	I,don't
2115	to,be

Table 3: Bi-grams

2. Part of speech tagging

POS tagging is a measurement that helped us understand the morphologic and syntactic properties. In the graph above, we see the use of part of speech within the depressed and undepressed posts, writers use a lot of NN (nouns) and JJ (adjectives), as well as VB (verbs), and we couldn't find a distinct difference between the texts. Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context. As earlier mentioned, the process of assigning a specific tag to a word in our corpus is referred to as part-of-speech tagging (POS tagging for short) since the POS tags are used to describe the lexical terms that we have within our text. Part-of-speech tags describe the characteristic structure of lexical terms within a sentence or text; therefore, we can use them for making assumptions about semantics.

Other applications of POS tagging include:

- Named Entity Recognition
- Co-reference Resolution
- Speech Recognition

When we perform POS tagging, it's often the case that our tagger will encounter words that were not within the vocabulary that was used. Consequently, augmenting your dataset to include unknown word tokens will aid the tagger in selecting appropriate tags for those words.

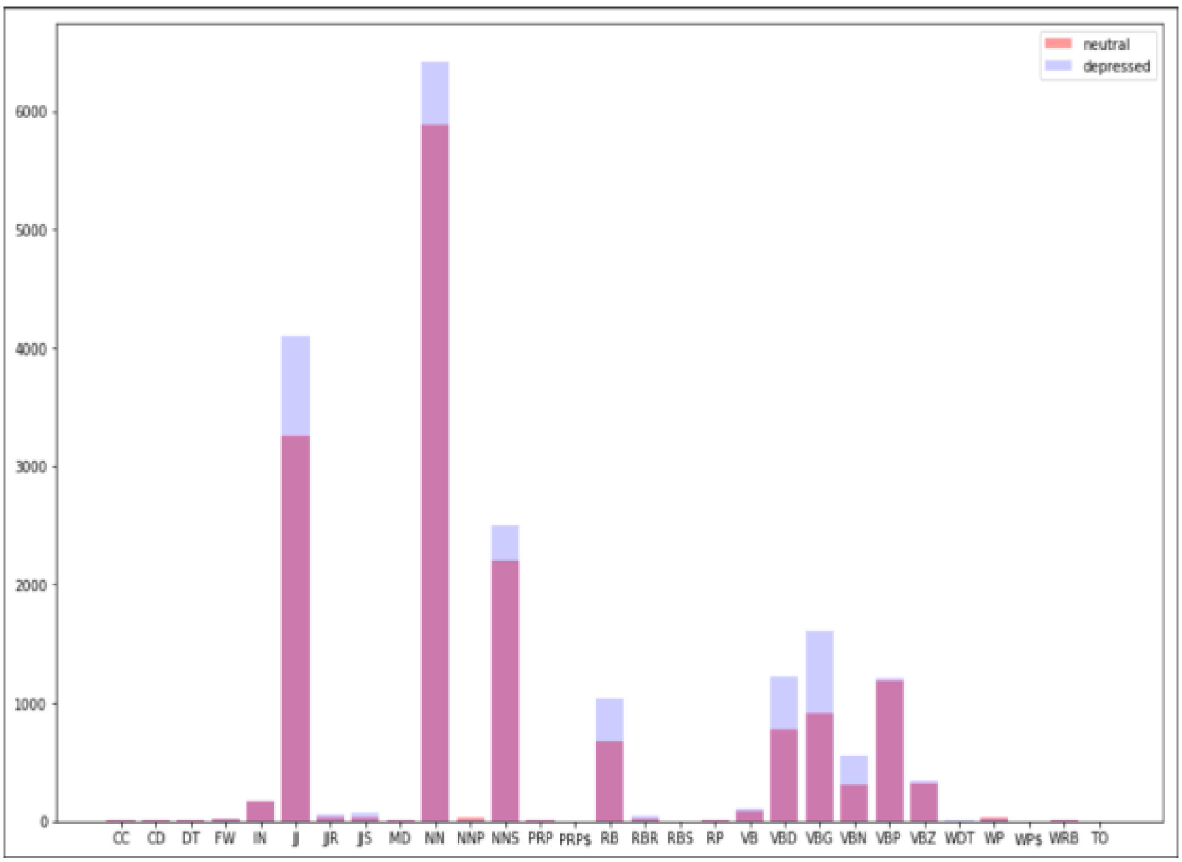

Figure 5: Neutral and depressed POS tagging measurement.

1. **Word cloud**

In this section we took the most common words and we tried to understand the general subject. We can see that there are different subjects, but the general topic is the struggle dealing with depression, while the undepressed cloud (the black one) doesn't have a specific subject.

Figure 6: Two words clouds, (top) word cloud is the cluster of data by topics. (down) word cloud shows positive topics.

2. Sentiment Analysis

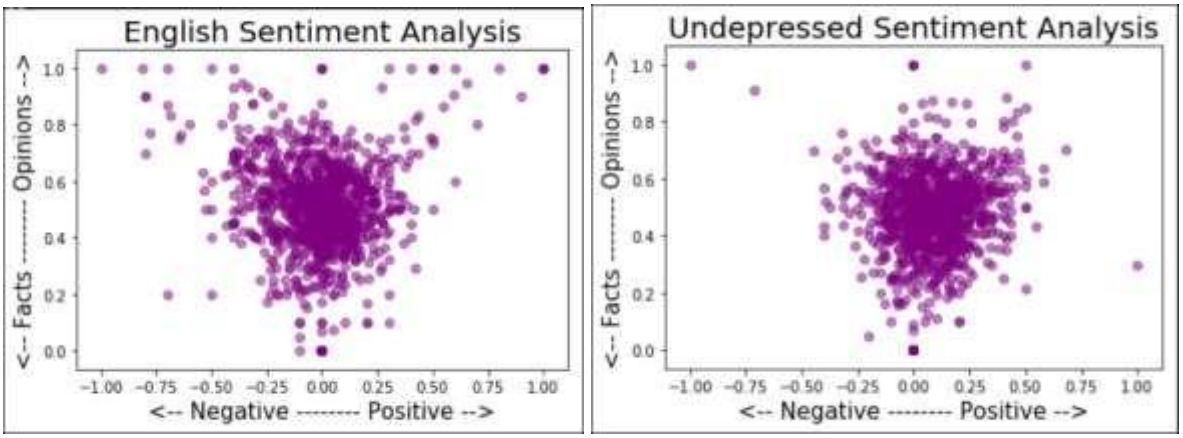

Figure 7: Classification of posts based on their polarity and with respect to the subjectivity of facts and opinions.

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment. Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs.

The scatter plots above, shows us the polarity and subjectivity of the depressed and undepressed posts, we see that the English (depressed) posts are tend to be more negative then undepressed, and less positive. here are the differences –

- Depressed: -
  - Polarity (over 0.25) = 71 (positive)
  - Polarity (under -0.25) = 115 (negative)
  - Subjectivity (over 0.6) = 279 (opinion)
  - Subjectivity (under 0.4) = 255 (fact)
- Undepressed –
  - Polarity (over 0.25) = 110 (positive)
  - Polarity (under -0.25) = 56 (negative)
  - Subjectivity (over 0.6) = 274 (opinion)
  - Subjectivity (under 0.4) = 256 (fact)

Chapter 4

Results and Discussion

4.1 Result

we show that our proposed method can significantly improve the accuracy rate, we got the highest score with a 99.69% success rate, from this paper

4.2 Discussion

- Early identification, intercession, and proper treatment can advance abatement, forestall backslide, and decrease the enthusiastic and monetary weight of the illness.
- Several previous works have highlighted the importance of early detection in improving outcomes related to major depressive disorder.

- A portion of the advantages of early finding and intercession incorporate the decrease of intermittent scenes and backslides, expansion in friendly capacity and usefulness, diminished truancy and better possibility of abatement

- These days, in the time of computerized environments, wherein we appreciate large number of flourishing and dynamic better approaches for imparting over the Internet, everybody is progressively going to stages like web-based media in look for psychological wellness advice.

- Mood tracking using mobile devices is an active area of research, leveraging emerging mobile sensing technologies. Previous work shows that mood influences our behavior and plays a significant role in our daily lives.

- Mental wellbeing plays a profound role in people's health and their quality of life. Mood is considered to be a compartment of cognitive aspect of human nature and can therefore be identified as a situational impairment which potentially influences cognitive aspects of interaction with the mobile device.

- Early finding and treatment of PPD can significantly improve the result, forestall backslide, and limit the related passionate and monetary weight.

- Depression is a crucial issue that influences an enormous level of the populace all throughout the planet. It not just influences the prosperity and usefulness of people yet in addition purposes weighty financial weight on the general public.

- Depression is curable disease. An early detection and intervention would shorten the treatment course.

- Early detection of depression can help mitigate these threats, but most studies on early detection rely on diagnoses from patients' self-reported surveys and experiences, which are extremely costly in terms of both time and money, and a major portion of countries with primary health care service do not have the support for these diagnoses. Fortunately, social media may provide us with a solution to this problem, as many studies have successfully leveraged the contents of social media to analyze and predict users' mental well-being.

Chapter 5

5.1 Conclusion

After examining all these features, we wanted to see if we can predict whether a post should be marked as depressed or not, we built a predictor using cross validation with several models (SVC, Multinomial Naïve Bayes, K-Nearest-Neighbors), we show that our proposed method can significantly improve the accuracy rate, we got the highest score with a 99.69% success rate, from this paper we see, that diagnosing depression can be much easier for authorities if they will work with data mining tools.

5.2 Future Scope

Since this is prediction model on the text analysis in future this model can be enhanced to examine real life habits, different languages, and over voice analysis. This project has a lot of scope in the field of psychology, where it makes it easier to detect depression and at a really early stage. This study can be extended to detecting depression from other sources of texts also, like people's chats, or people's confessions. It can also be extended to detect depression from speech. There is already a lot of study in the field of speech recognition, so we can also implement depression recognition from speech.

Speech recognition is one such technology that is empowered by AI to add convenience to its users. In speech recognition, the computer takes input in the form of sound vibrations. This is done by making use of an analog to digital converter that converts the sound waves into a digital format that the computer can understand. Detecting depression from speech would be a really huge step in the field of machine learning and psychology.

The mass adoption of artificial intelligence in users' everyday lives is also fueling the shift towards voice applications. The number of IoT devices such as smart thermostats, appliances, and speakers are giving voice assistants more utility in a connected user's life. Smart speakers are the number one way we are seeing voice being used; however, it only starts there. Many industry experts even predict that nearly every application will integrate voice technology in some way in the next 5 years.

REFERENCES

[1] G. 2. D. a. I. I. a. P. Collaborators., "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories.," The Lancet, 2018.

[2] W. H. Organization, "Depression," 2020.

[3] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang and A. Ulhaq, "Depression detection from social network data using machine learning techniques," Health Information Science and Systems, 2018.

[4] M. D. Choudhury, M. Gamon, S. Counts and E. Horvitz, "Predicting Depression via Social Media," AAAI, p. 10, 2013.

[5] S. Seraj, K. Blackburn and J. W. Pennebaker, "Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup," Proceedings of the National Academy of Sciences, 2021.

[6] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu and Z. Bao, "A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network," in Springer, Berlin, Heidelberg, 2013.

[7] B. Hosseinifard, M. H. Moradi and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal,," Computer Methods and Programs in Biomedicine,, vol. 109, no. 3, pp. 339-345, 2013.

[8] M. J. Patel, A. Khalaf and H. J. Aizenstein, "Studying depression using imaging and machine learning methods,," NeuroImage: Clinical,, 016.

[9] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in Association for Computational Linguistics, 2002.

[10] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture, 2003.

[11] D. Malvern, B. Richards, N. Chipere and P. Durán, Lexical Diversity and Language Development, 2004.

[12] S. Feldman, "NLP Meets the Jabberwocky: Natural Language Processing in

Information Retrieval," Information Today, Inc., 1999.

[13] H. M. Wallach, "Topic modeling: beyond bag-of-words," in ICML '06: Proceedings of the 23rd international conference on Machine learning, 2006.

[14] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, 2018.

[15] R. Cloninger, D. Svrakic and T. Przybeck, "Can personality assessment predict future depression? A twelve-month follow-up of 631 subjects," Journal of affective disorders, 2006.

[16] L. B. Alloy and M. S. Robinson, "Negative Cognitive Styles and Stress-Reactive Rumination Interact to Predict Depression: A Prospective Study," Cognitive Therapy and Research, pp. 275-291, 2003.

[17] G. Livingston, A. Mann and B. Blizard, "Does sleep disturbance predict depression in elderly people? A study in inner London.," British Journal of General Practice, 1993.

[18] E. Åström, R. Adolfsson , M. Carelli and M. Rönnlund, "Depressive symptoms and time perspective in older adults: Associations beyond personality and negative life events," Aging and Mental Health, 2018.

[19] O. TE, T. GJ. and . R. SD, "The language of paranoia," The American Journal of Psychiatry, 1982.