*A Progress Report*

*on*

**DETECTION OF DEPRESSION ON TWITTER USING  
MACHINE LEARNING TECHNIQUES**

*carried out as part of the* ***Minor Project IT3270*** *Submitted by*

***Vanshika Goyal***

***199302049***

***Riyanshi Bohra***

***199302110***

***VI-IT***

*in partial fulfilment for the award of the degree*

*of*

**Bachelor of Technology**

In

**Information Technology**



**Department of Information Technology**

**School of Computing and IT**

**Manipal University Jaipur**

***April 2022***

**Abstract:-**

The use of Social Media Sites like Twitter is expanding extremely fast. Nowadays, people tend to rely on these social media applications to share their emotions and feelings. Thus, this readily available content has become helpful for us to analyse the mental health of such users. We can apply various machine learning techniques on social media data to extract the mental health status of a user focusing on Depression.

Detecting texts that express negativity in the data is one of the best ways to detect depression. The Ensemble Learning approach for solving this problem has been enlightened. We aim to find and implement the most appropriate approach and algorithm to solve this problem.

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Title** | **Page No.** |
| **1.** | **Introduction** | 4 |
| 1.1. | Problem Statement | 4 |
| 1.2. | Objectives | 4 |
| 1.3. | Scope of Project | 5 |
| **2.** | **Analysis** | 6 |
| 2.1. | Literature Review/Related Work | 6 |
| **3.** | **System Design & Methodology** | 7 |
| 3.1. | System Architecture | 7 |
| 3.2. | Development Environment | 8 |
| 3.3. | Methodology: Algorithm/Procedures | 8 |
| **4.** | **Work Done** | 11 |
| 4.1. | Work done till date | 11 |
| 4.2. | Work left | 11 |
| 4.3. | Month wise plan of work | 11 |
| **5.** | **Conclusion and Future Plan** | 12 |
|  | **References** | 13 |

1. **Introduction:-**
2. **Problem Statement**

“To analyse and detect depression of social media data of users like Twitter feed by using machine learning techniques.”

people increasingly rely on social

media for sharing emotions, and daily life activities thus

helpful for detecting their mental health. We aim to apply

Machine Learning Techniques on Social Data of a user like

Twitter feeds for performing analysis focusing on depression

detection.

people increasingly rely on social

media for sharing emotions, and daily life activities thus

helpful for detecting their mental health. We aim to apply

Machine Learning Techniques on Social Data of a user like

Twitter feeds for performing analysis focusing on depression

detection.

people increasingly rely on social

media for sharing emotions, and daily life activities thus

helpful for detecting their mental health. We aim to apply

Machine Learning Techniques on Social Data of a user like

Twitter feeds for performing analysis focusing on depression

detection.

People increasingly rely on social media for sharing emotions, and daily life activities thus helpful for detecting their mental health. We aim to apply Machine Learning Techniques on Social Data of a user like Twitter feeds for performing analysis focusing on depression detection.

1. **Objectives**

Detecting earlier depression can be a huge step to address the mental illness and offer support to the people suffering from this terrible mental illness.

* Data analysis of tweets by Machine Learning Techniques:

Sentiment analysis is the process to identify the tone of the text is either positive, negative, or neutral. This can allow identifying the state of the user either they are in a positive mood or negative mood. The sentiment analysis technique is applied to each tweet to identify the sentiment score and labelled them as positive, negative, or neutral. Sentiment analysis in Twitter data quantifies the status of a tweet or comment by the user by calculating the sentiment scores.

The model will be written in python and it will tell whether a given tweet is depressive or not.

* Detection of depressive tweets:

After analyzing the data and finding the sentiment score on tweets via sentiment analysis technique, the machine learning algorithm enables to classify the tweets either it is depressive or not depressive based on the sentiment score labels on tweets. We implement a machine-learning algorithm to detect depression based on the tweets data which was labelled with sentiment scores.

1. **Scope of Project**

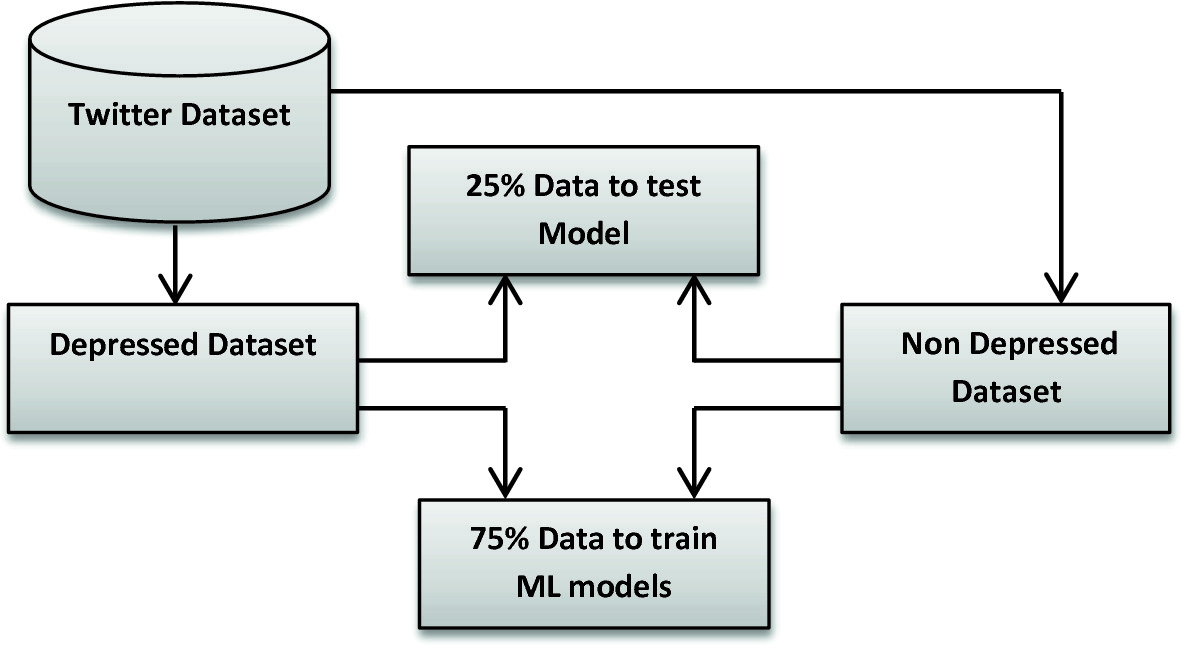
The areas of study for our project are machine learning and artificial intelligence using python. Using a machine learning approach to detect depression will surely help social media users for detecting and predicting depression risk. A machine learning approach like supervised learning can analyse and build a model on social media posts. There are many factors like users' posts, tweets, replies, post time, emotions, etc. which can contribute to detecting depression. Before making a model, we will do exploratory data analysis on our dataset to thoroughly understand it.

1. **Analysis: -**
2. **Literature Review**

In machine learning, there are supervised machine learning classification algorithms for example Support Vector Machine, K Nearest Neighbor, Naive Bayes, Decision Tree, etc. Based on the different tasks and available data we can use those algorithms. The accuracy of the model can be improved using ensemble learning methods.

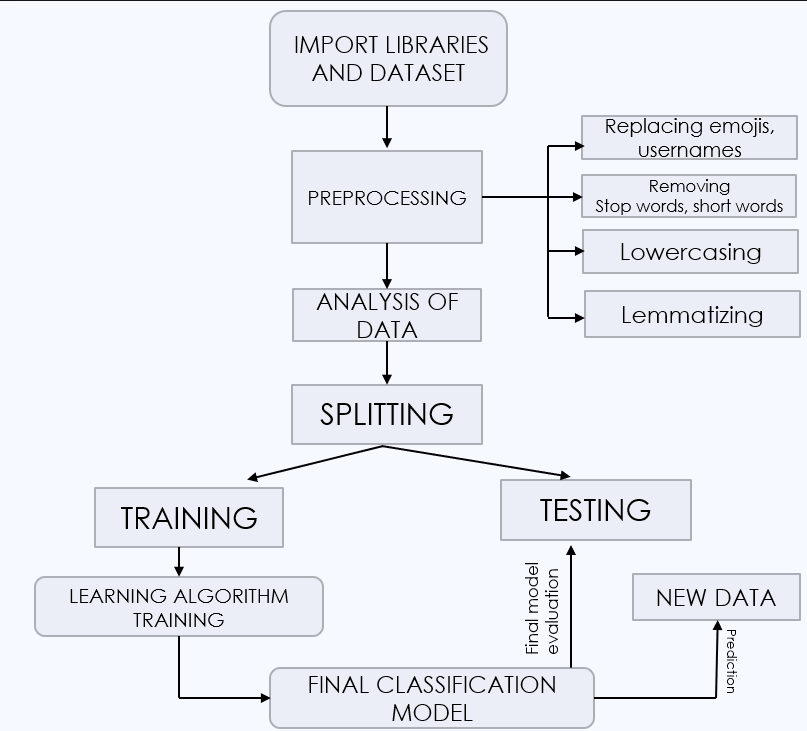
It is often difficult, to point out the real cause of depression. In some cases, there exists no such cause, however, the patient feels depressed and has a lack of interest in doing anything.

In this project, we analyze the dataset by first preprocessing the data, and then creating Word clouds for Positive and Negative tweets and see which words occur the most. Then we split the data into Training and Testing set. We train the data in different classifiers and check which classifier has the most accuracy.



1. **System Design & Methodology: -**
   1. **System Architecture**

Flowchart Diagram:

****

* 1. **Development Environment**

We will achieve our above-mentioned objectives by using a programming language called Python and by using Machine Learning techniques & algorithms. Following are the detailed software requirements for the development of our project:

* Kaggle(for the dataset)
* Jupyter Notebook which can be opened through Anaconda prompt.
* Wordcloud, a visualization tool.

Minimum System Requirements: -

* Processors: Intel Atom® processor or Intel® Core™ i3 processor
* Disk space: 1 GB
* Operating systems: Windows\* 7 or later, macOS, and Linux
* Python\* versions: 2.7.X, 3.6.X
  1. **Methodology: Algorithm/Procedures**

The framework begins with data collection by using Twitter scrapper tools and is stored in a. csv file. The raw data will be cleaned and we start to do data pre-processing. The data will be normalized. Next, the data will be analyzed by using sentiment analysis to obtain a score of words. The data are fed into different classifiers. The data is then split into the train and test set. The training data is used for model development to make sure the classifier learns. The test data will feed into the model once the model learned about the data for evaluation. The accuracy results of the classifiers will be compared to determine which algorithm outperforms well.

The methodology of the project can be broadly classified into the following steps-

1. Data extraction and Data collection

The dataset used for this project contains 1,600,000 tweets extracted using the **Twitter API**. The tweets have been annotated **(0 = Negative, 4 = Positive)** and they can be used to detect sentiment.

The Dataset contains the following fields:

1. sentiment: the classification of the tweets *(0 = negative, 4 = positive)*
2. ids: tweet id
3. date: tweet date
4. flag: The query.
5. user: username of tweeps
6. text: the text of the tweet

For this project only the sentiment and text fields are required, so we drop the rest.

Furthermore, the sentiment field will be changed to reflect new sentiment format as follows (0 = Negative, 1 = Positive)

1. Data Preprocessing

**Text Preprocessing** at this stage, the text is preprocessed and prepared for analysis and model development.

The Preprocessing steps taken are:

1. Lower Casing**:** Each text is converted to lowercase.
2. Replacing URLs: "http", "https" & "www" are replaced with "URL".
3. Replacing Emojis: Emojis are replaced with a pre-defined dictionary alternative
4. Replacing Usernames: Usernames are standardized.
5. Removing Non-Alphabets: Only retaining Digits, Alphabets and space.
6. Removing Consecutive letters: 3 or more consecutive letters are shortened
7. Removing Short Words: Words with length less than 2 are removed.
8. Removing Stopwords: Stopwords are the English words which does not add much meaning to a sentence.
9. Lemmatizing: Lemmatization is the process of converting a word to its base form.
10. Data Visualization

Here, we're going to use visualization to quickly explore the preprocessed data in a bid to uncover any implicit patterns, spot anomalies, and check assumptions. We'll plot **Word Clouds**for**Positive and Negative** tweets from our dataset and see which words occur the most.

1. Data Analysis

In this project, we analyze the dataset by first preprocessing the data, and then creating Word clouds for Positive and Negative tweets and see which words occur the most. Then we split the data into Training and Testing set. We train the data in different classifiers and check which classifier has the most accuracy.

1. Classification using Machine Learning Algorithms

In this project, we used 3 machine learning models:

* Bernoulli Naive Bayes (BernoulliNB)
* Linear Support Vector Classification (LinearSVC)
* Logistic Regression (LR)

The evaluation metric used in this project is the Precision-Recall Metrics. We also plotted the Confusion Matrix to get an understanding of how our model is performing on both classification types.

1. Training and Testing of Data

The Preprocessed Data is split into;

* **Training set:** Containing 90% of the data.
* **Test set:** Containing 10% of the data

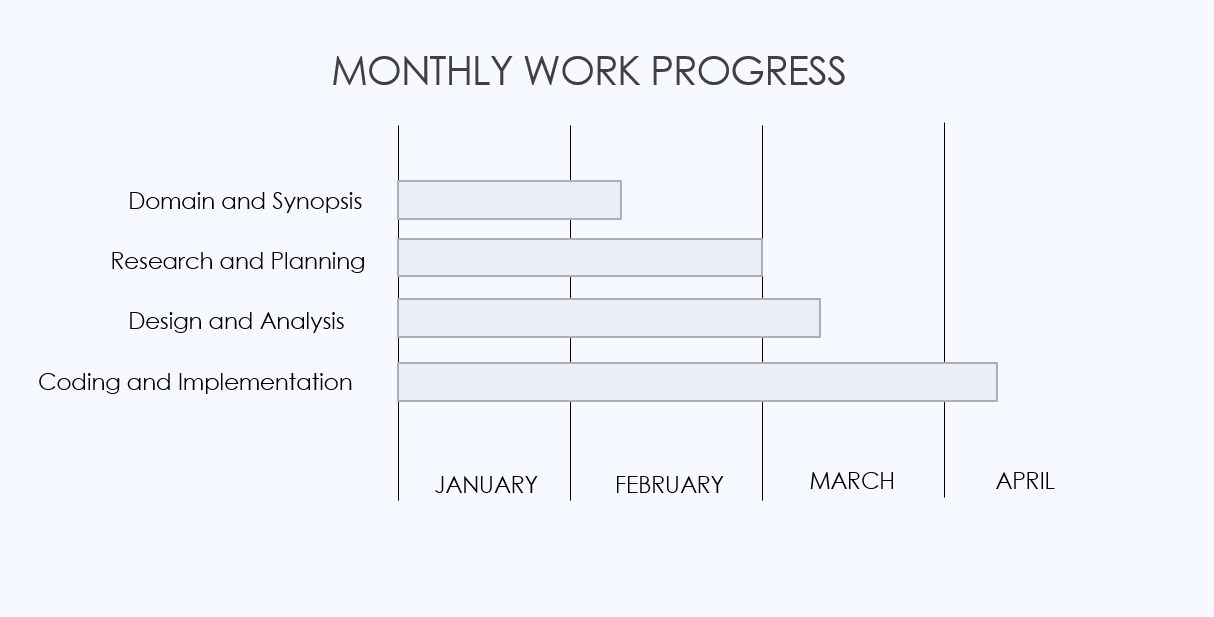
1. **Work Done: -** 
   1. **Work done till date**

We have successfully completed the Planning and Research part of our project. We found a relevant dataset from Kaggle and imported the necessary libraries in our code. We started working on the implementation of the code by doing Data preprocessing on the dataset and Data Analysis using a data visualization tool known as Wordcloud.

* 1. **Work left**

We are currently working on Model development and Evaluation. The work left includes predicting the outcome of the three classifiers and plotting the Confusion Matrix to get an understanding of how our model is performing on all classification types. The evaluation metric that we will be using in this project is the Precision-Recall Metrics. We will be comparing the performances of the different classifiers and decide on which provides us with the best accuracy.

* 1. **Month wise plan of work**



1. **Conclusion and Future Plan: -**

This project defines a classification problem as identifying whether a person is depressed or not, based on their tweets.

This study used a dataset taken from twitter to explore various methods of early detection of depression based on machine learning. We performed a thorough analysis of the dataset to characterize the subjects’ behavior based on their tweets.

Different machine learning algorithms are used for the detection of depression. Many preprocessing steps are performed, including removal of stopwords and short words and lemmatizing. This project can be considered as a step toward building a complete social media-based platform for analyzing and predicting mental and psychological issues and recommending solutions for these users.

This study can be extended in the future by considering more ML models that might have much better accuracy for the classification problem and to find a more dependable way to measure the features’ impact.

**References**

1. 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", Heal. Inf. Sci. Syst., vol. 6, no. 1, pp. 1-12, 2018.
2. Holleran SE. The early detection of depression from social networking sites. Tucson: The University of Arizona; 2010.
3. "How to Get Use & Benefit From Twitter's API", Fontanella Clint, Available: https://blog.hubspot.com/website/how-to-use-twitter-api.
4. Ang, L. (2011) Community relationship management and social media. Journal of Database Marketing and Customer Strategy Management, 18 (1), 31–38, doi:10.1057/dbm.2011.3.
5. <https://www.javatpoint.com/machine-learning>
6. A. Sistilli, Twitter Data Mining: Analyzing Big Data Using Python | Toptal, [online] Available: <https://www.toptalcom/python/twitter-data-mining-using-python>.
7. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. Predicting depression via social media. ICWSM. 13, 1-10 (2013).
8. Reece, A.G., et al. Forecasting the onset and course of mental illness with Twitter data. Sci. Rep. 7(1),13006 (2017).
9. Nadeem, M. Identifying depression on Twitter. Preprint at arXiv:1607.07384 (2016).
10. Harman, G., & Dredze, M. H. Measuring post traumatic stress disorder in Twitter. Proceedings of ICWSM, (2014).
11. Benton, A., Mitchell, M., & Hovy, D. Multi-task learning for mental health using social media text. Preprint at arXiv:1712.03538 (2017).
12. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing depression from twitter activity in Recognizing depression from twitter activity 3187-3196 (ACM, 2015).
13. Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. Monitoring tweets for depression to detect at-risk users. CLPsych, 32, (2017).