

**A PROJECT REPORT ON**

**ANALYZING STRESSFUL TEXT BASED ON SOCIAL  
MEDIA POSTS**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE  
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE

Bansilal Ramnath Agarwal Charitable Trust's  
VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE– 411037.

(An Autonomous Institute Affiliated to Savitribai Phule Pune  
University, Pune.)

22	Ojas Mandlecha	11810880
1	Swaraj Kothekar	11811307
5	Yohaán Kudtarkar	11810298

## **ABSTRACT**

Every person, at some point in life, experiences stress. Some can deal with it easily, for some it may take years or even their entire life to deal with stress. Especially with the fast pace of human lives these days stress has become a very common phenomenon among individuals. Feelings of inadequacy and failure in such competitive times has increased the amount of anxiety and stress. With the rise of popularity of social media platforms people tend to express their feelings and views online giving researchers access to a wide variety of content for useful analysis. In this research we make use of the lengthy and detailed posts of very famous social media site Reddit.

We are basically making use of the three major supervised learning algorithms named Naive Bayes, SVM and Decision Tree to train our model in order to classify text based posts collected dynamically using Reddit's Pushshift API as stressful or non-stressful content. Further on we analyze the severity of the stressful text by classifying them and predicting whether it is inclining towards suicidal tendencies or not. We evaluate the performance of each model through different measures like accuracy, also the precision, the recall and the F1 score to single out the ultimate model.

# TABLE OF CONTENTS

## LIST OF FIGURES

i

<b>Sr. No.</b>	<b>Title of Chapter</b>	<b>Page No.</b>
<b>1.</b>	<b>Introduction</b>	<b>8</b>
1.1.	Overview	8
1.2.	Motivation	8
1.3.	Problem Definition and Objectives	9
1.4.	Project Scope and Limitations	9
<b>2.</b>	<b>Literature Survey</b>	<b>10</b>
2.1.	Background	10
2.2.	Literature Review	11
<b>3.</b>	<b>Software Requirements Specification</b>	<b>14</b>
3.1.	Assumptions and Dependencies	14
3.2.	Features	14
3.3.	Functional Requirements	15
3.3.1.	Functional Requirements 1	15
3.3.2.	Functional Requirements 2	15
3.3.3.	Functional Requirements 3	15
3.3.4.	Functional Requirements 4	15
3.4.	System Requirements	15
3.4.1.	Software Requirements	15
3.4.2.	Hardware Requirements	15
3.5.	System Implementation Plan	16
3.6.	SDLC Model	17
<b>4.</b>	<b>System Design</b>	<b>18</b>
4.1.	System Architecture	18

<b>5.</b>	<b>Project Implementation</b>	<b>19</b>
5.1	Overview of Project Modules	19
5.2	Tools and Technologies Used	21
5.3	Algorithm Details	25
5.3.1	Algorithm 1	25
5.3.2	Algorithm 2	26
5.3.3	Algorithm 3	27
5.3.4	Algorithm 4	28
5.4	Implementation Aspects	28
<b>6.</b>	<b>Results</b>	<b>34</b>
6.1	Outcomes	34
6.2	Screenshots	34
<b>7.</b>	<b>Conclusions</b>	<b>37</b>
7.1	Conclusion	37
7.2	Future Work	37
7.3	Applications	37
	<b>Appendix A : Plagiarism Report</b>	<b>38</b>
	<b>References</b>	<b>39</b>

## **LIST OF FIGURES**

Fig. No. 1. System Implementation Plan

Fig. No. 2. System architecture

Fig. No. 3. SVM

Fig. No. 4. Decision Tree

# **1. INTRODUCTION**

## **1.1 Overview**

Our main motive through this project is to analyze the posts of our reddit dataset for the expression of stress. We have classified the posts into stressed and non stressed posts. And finally, our objective is to analyze if these posts incline towards any self harm or suicidal thoughts or not.

## **1.2 Motivation:**

Social media has become a major part of our lives nowadays. It provides a medium to not only share glimpses of people's personal lives but also to share a piece of their mind. The freedom provided by social media has allowed people to talk freely about their thoughts with anonymity. Unfortunately, talking about mental health issues is still considered a taboo and people don't feel comfortable seeking professional help.

Additionally, in recent times ever since covid-19 came into picture and people were forced into the confinements of their home for many days at a stretch, it was observed that many individuals have been struggling with stress which took a toll on their well being. In such scenarios, social media has become an easily accessible and welcoming platform to share their experiences. This motivated us to study social media posts and try to figure out certain similarities and patterns within them.

Stress is a universal experience which finds its way into every individual's life. While a certain amount of stress can have positive effects like boosting productivity, too much stress is often related to negative health effects. Since the normal social presence has shifted online, people have shifted towards social media to express their problems and get help because of its widespread accessibility and boundless nature.

Mental health illness is an important problem to address, and the rise of social media is encouraging conversations around mental health. We aim to make use of this shift in the mode of expression to collect and study people's experiences of stress to present an analysis that might be useful in formulating ways to help those experiencing stress.

### **1.3 Problem Definition & Objectives :**

Through this project, we aim to understand and analyze stressful text based on social media posts. Our main objective is to understand and analyze the dataset (In our case, it is a Reddit dataset). We aim to analyze various posts of the Reddit dataset for different expressions and signs of Stress and finally classify the posts into Stressful content or non Stressful Content.

#### **Objectives:**

1. Through this project, we aim to analyze the posts of our reddit dataset for the expressions of stress.
2. We also aim to classify the posts into stressed and non stressed posts.
3. And finally, our objective is to analyze if these posts incline towards any self harm or suicidal thoughts or not.

### **1.4 Project Scope and Limitations:**

Social media is an easily accessible platform for people to express their views and emotions. While there are a number of social media platforms available, for this project we are focusing on only one social media site - Reddit. According to Reddit themselves, “It’s a network of communities where people can dive into their interests, hobbies, passions. There's a community for whatever you’re interested in”. We restricted ourselves to this particular social media platform due to the lengthy nature of posts on Reddit as it makes it an ideal data source for studying the

nuances of phenomena like stress. Additionally, the scope of this project is limited to posts written in the English language only. We do not intend on incorporating any other language like Hindi, French, etc, i.e., Multilingual Analysis is not in our scope. The text may have special characters but inclusion of emojis or words from any other language are strictly out of scope. Based on the posts from the Reddit dataset our model will be able to predict whether a given piece of text is stressful or non-stressful. Further on, our model will also predict the severity of this stress by predicting if a given piece of stressed text is inclining towards suicidal tendencies or not. We would also like to mention that this project is only for study purposes and it only predicts certain conditions. It does not claim to diagnose anybody of having any sort of mental illness.

## **2. LITERATURE SURVEY**

### **2.1 Background:**

Machine learning comes under the umbrella of artificial intelligence that autonomously learns from data and information using the computer algorithm. Machine learning systems do not need to be explicitly programmed and can modify and improvise the algorithm on their own. Machine learning algorithms enable computers to communicate with humans on the go, drive autonomously, create and publish reports on sports games, and track suspected terrorists. This type of Research is effectively increasing and can help with diagnosing and treating mental problems. ML technology has the potential to learn human behavioral patterns, identify mental health symptoms and risks, predict disease progression, and open up new ways to customize based on personal preferences and optimize treatments.



Unique Selling Proposition:

After reading some research papers related to our topics we came to our the conclusion that our unique selling proposition will be to analyse how intense stress leads to suicidal or self harm thoughts. Also stats issued by WHO says that more than 7 lakh plus deaths happened due to suicide every year. In 2019 use recorded a death by suicide every 11 minutes. So we thought to analyse our data to find and categorise texts as suicidal or not as early detection of such intentions can be helpful in providing necessary help.

## **2.2 Literature review:**

Most studies of detecting mental illness using natural language have focused on Text samples that overlap in time with mental illness. According to the findings of these studies, everyday language contains implicit information about stressful situations. Based on external Wikipedia corpus, using the Latent Dirichlet Allocation (LDA) model, Phan et al. [1] presented a framework for expanding short texts by attaching hidden topical names.

In today's world, many of us rely on social media platforms to connect with one another and discuss the stressful experiences we are or have had. De Choudhary et al. (2013) [2] used a CES-D (Center for Epidemiological Studies Depression) scale, SMDI (Social Media Depression Index), PCA (Principal Component Analysis), and SVM (Support Vector Machine) classifier to achieve their goal of measuring depression.

In addition, Saravia et al. (2016) [3] examined and detected mental illness in social media, which helped predict depression. The CES-D Scale, the TF-IDF, the Sentiment 140 API, the PLF, and the Random Forest Classifier were also employed.

Text classification has been accomplished using a variety of techniques. Rakshitha et al. (2018) [4] used machine learning approaches to determine the mental health of a person based on their intuitive well being. They selected specific keywords which they called as emotional keywords which portrayed rational stability but showed

discouragement and unhappiness. The dataset consisted of tweets with fields like user name, user description, user language, friends count, followers count, etc. They used two different approaches for classification. The first approach was by using the MonkeyLearn API for text classification by using the Multinomial Naive Bayes algorithm. The second approach they used was building their own classifiers in Python using ML algorithms like Support vector machines. Both these approaches determined whether the specific text is positive class or negative class.

Using deep learning approaches, Lee et al. (2020) [5] developed six binary classification models, each of which categorizes a user's specific post into one of the following subreddits like depression, Anxiety, bipolar, BPD, schizophrenia and autism. XGBoost and convolutional neural networks (CNN) were employed. They excluded the posts of users who have written across various subreddits.

Various studies have used various approaches and algorithms to predict mental illness from posts. Tiwari et al.(2021) [6] developed machine learning models using five algorithms namely, Naïve Bayes, Decision Tree, Support Vector Machines, k-nearest Neighbors and Random Forest for predicting the data. They observed and compared the accuracy scores of these five algorithms and found that Decision Tree gave the most accurate results.

Some studies have also attempted to determine the level of depression represented by the text. Aldarwish et al. (2017) [7] used social media posts to predict depression using the BDI-II Questionnaire. RapidMiner, Naive Bayes, and SVM classification were also used to create a depression model.

Concerns concerning suicide-related communication on social media believe that suicidal utterances on social media platforms are symptoms of true suicidal anguish in vulnerable individuals who post this material; consequently, the emotive nature of suicide talk on social media must be identified and addressed.

Burnap et al. [8] developed a number of machine classification models built with the aim of classifying text relating to communications around suicide on Twitter. The classifier distinguishes between content that is more worrying, such as suicidal ideation, and other suicide-related themes, such as reporting a suicide, memorialising,

campaigning, and providing assistance. It also looks for sarcastic references to suicide. In response to the loud nature of social media, where short, informal spelling and grammar are regularly used, they devised a set of regular expression (RegEx) and pattern matching rules.

The dataset used for research also plays an important role. The dataset used in [9] has been created by collecting Reddit posts belonging to different categories of subreddits like interpersonal conflict, mental illness and financial need wherein people are likely to share their experiences with stress. The primary goal of this work being prediction of stress, posts were divided into smaller segments for easy handling. A subset of these segments were annotated as stressful and non-stressful by human annotators from Amazon Mechanical Turk. This data was analyzed for word categories and lexical diversity. Various lexical, syntactic, social media features were added to the dataset. Various models were trained using this data - logistic regression, Naïve Bayes, Support Vector Machines (SVMs), decision trees, Perceptron, and Convolutional Neural Network (CNN). The parameters for these models were tuned using 10-fold cross-validation, and different combinations of input and features were used to obtain results.

### **3. SOFTWARE REQUIREMENTS SPECIFICATION**

Before starting to build any project, it is necessary to delineate a proper layout as to what exactly the team wishes to build and the different tools the team plans on using. Similarly, for this project too, we defined our objectives, the different requirements, the scope and the points we decided were not in scope for us.

#### **3.1 Assumptions and Dependencies :**

We have chosen Reddit as our social media platform and use its posts for analysis. Hence, we assume that whatever people have written in their posts and the issues they have claimed in their posts are true. We are also assuming and taking into consideration posts written in the English language only.

#### **3.2 Features**

A few features of our model include :

1. Identifying stress markers: Based on the traditional definition of stress the model will be able to identify stress markers that may include but are not limited to words like stress, worry, anxious, tense.
2. Classifying text: Following identification of stress markers the model will classify a given piece of text as stressful or non-stressful.
3. Analyzing severity: Further on, our model will assess the intensity of stress by analysing a given piece of stressed text for signs of inclination towards suicidal tendencies.

### **3.3 Functional Requirements**

Functional Requirement 1: Read the dataset and retrieve each post successfully.

Functional Requirement 2: Preprocess each and every text in the dataset using text preprocessing.

Functional Requirement 3: Classify the preprocessed text into stressful content or non-stressful content.

Functional Requirement 4: Analyze severity of text by identifying if a given piece of text is inclining towards suicidal tendencies or not.

### **3.4 System Requirements**

#### **3.4.1 Software Requirements**

Software requirements include:

1. Operating System: Windows
2. Platform: Jupyter Notebooks

#### **3.4.2 Hardware Requirements**

Hardware requirements include:

1. Intel core i5 processor or above
2. Minimum 8 gb RAM
3. 1 gb storage or above

### 3.5 System Implementation Plan

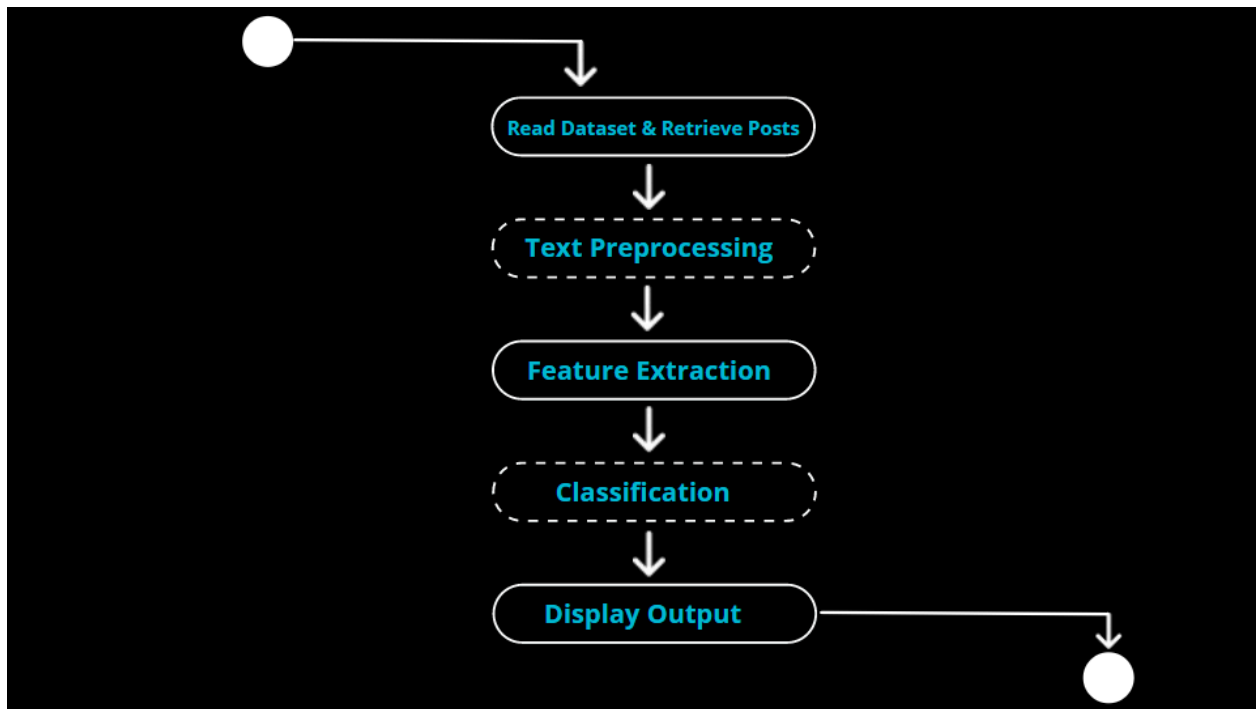


Fig no. 1. System Implementation Plan

The first step of the system implementation plan will be to read the dataset properly and retrieve each post successfully. Following that each post must go under text preprocessing which include cleaning the data, casing, stopword removal and word lemmatization. After text preprocessing the next step will be feature extraction. Machines can not understand text but only numbers and a machine learning model can not work on raw text directly hence we transform words into feature sets that improves accuracy of the learning algorithm and also shortens the time. Following this we classify Reddit text based posts into stressed or non stressed and finally display the output in form on CSV or Excel file.

### 3.6 Software Development Life Cycle (SDLC)

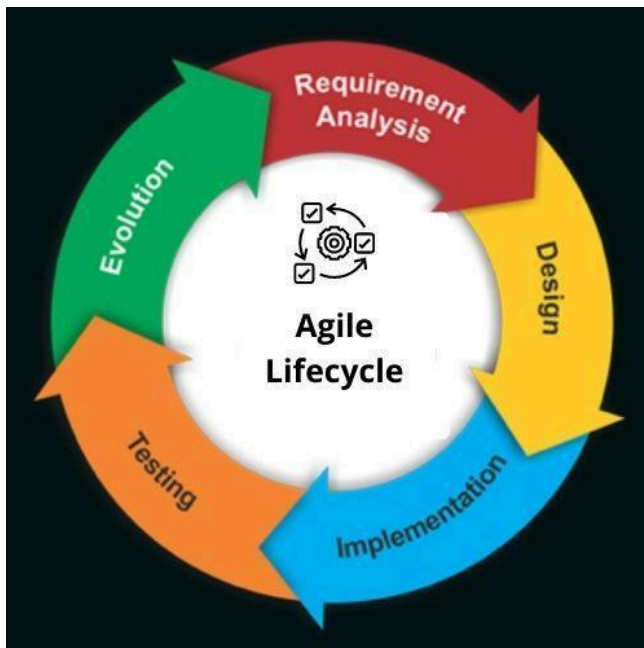


Fig no. 1. Software D

Software Development life cycle (SDLC) is a development model used in project management. It gives us a flow of stages that includes an information system describing the process followed for development of the project which ranges from an initial feasibility study to the maintenance of completed projects.

In this project we have used the Agile methodology of SDLC. The first phase of the methodology begins with Requirement Analysis where we make note of and gather all necessary requirements that we'll need throughout the project. Next phase is Design where we design what our model would look like and also decide its various attributes and behaviors. Next comes Implementation where we write code for our model and implement the already decided attributes and behaviors. After that comes Testing where we test our model for various test cases and check how accurately the model works and make necessary changes. Last is the Evolution phase where we update our models according to the various inputs we got from previous phases which helps our model evolve.

## 4. SYSTEM DESIGN

### 4.1 System Architecture

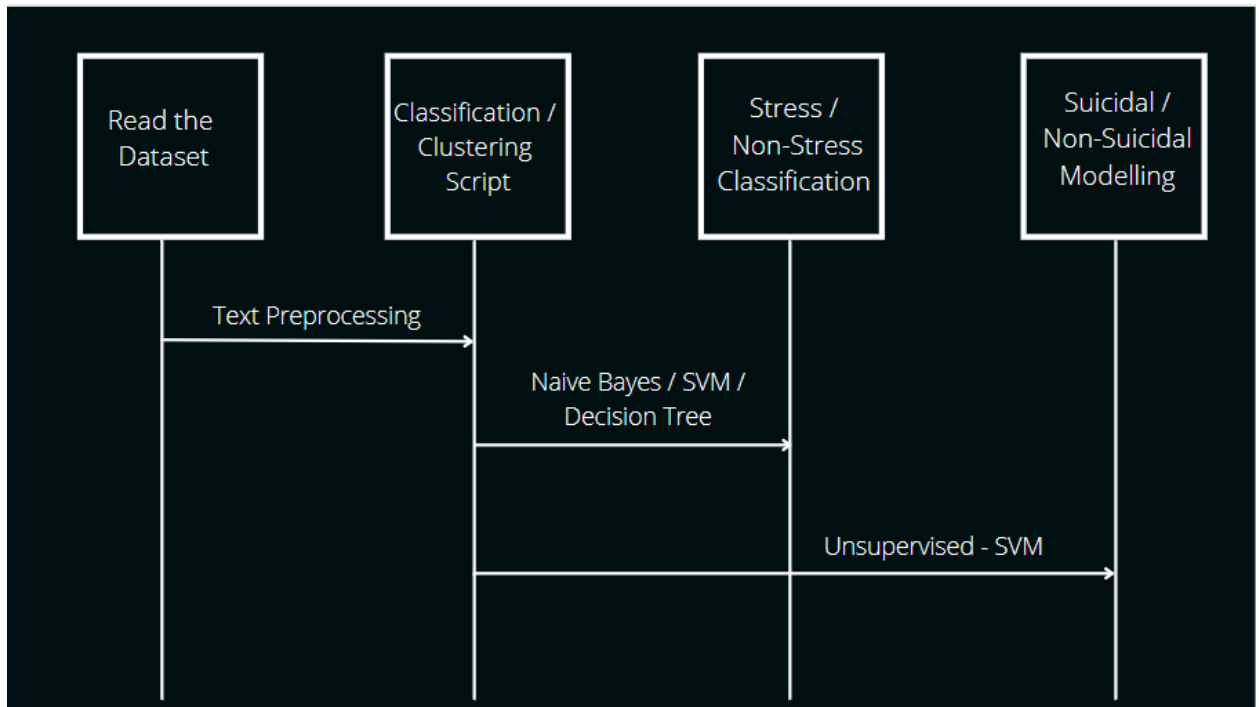


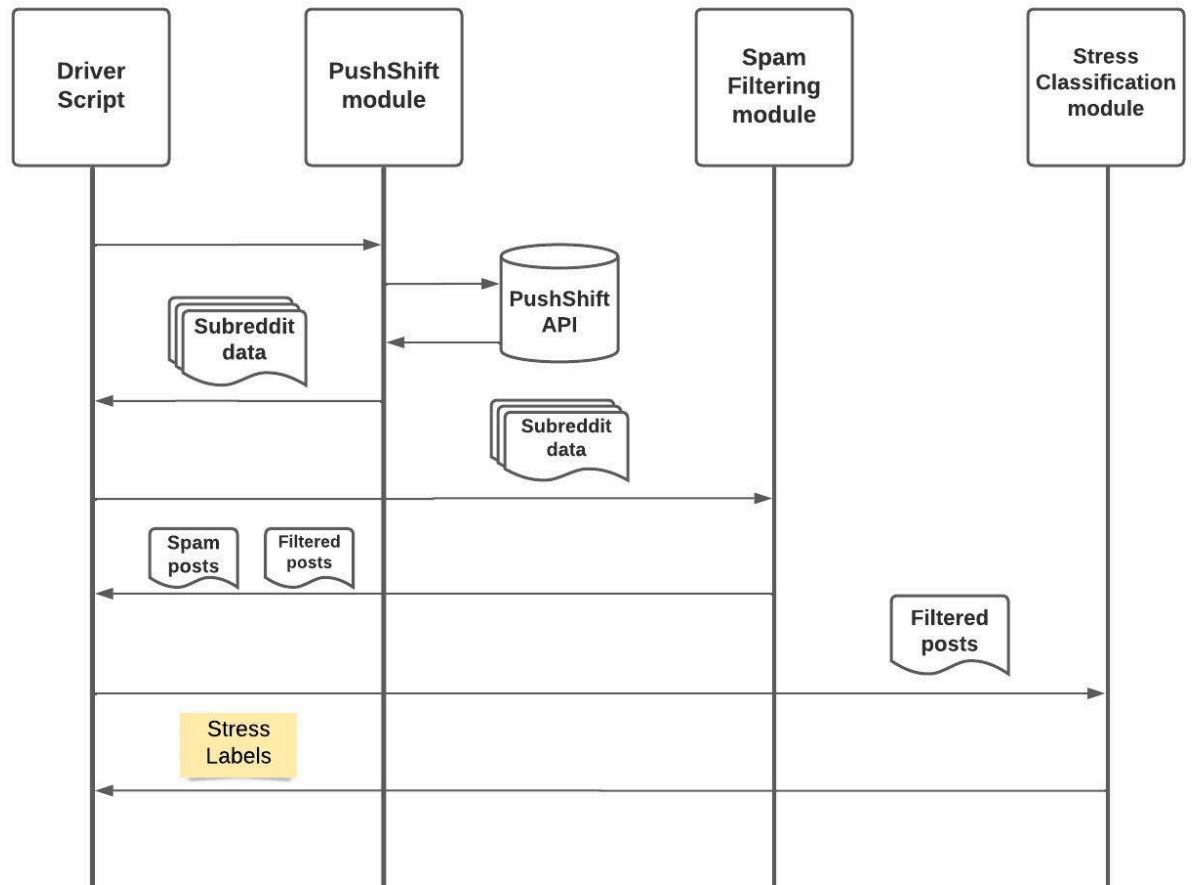
Fig No. 2. System Architecture

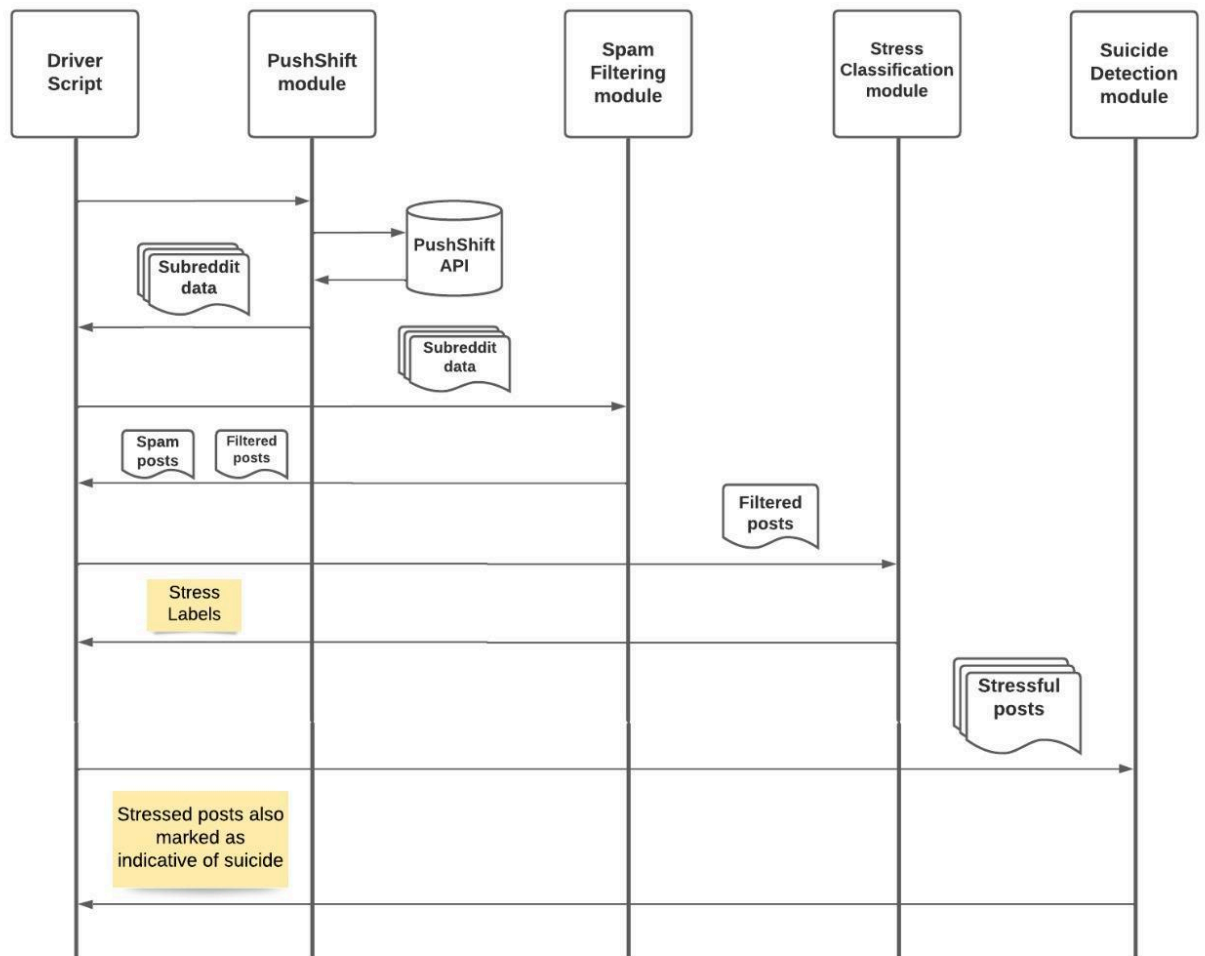
System architecture of our project involves selecting a dataset that provides a large number of posts, after such a dataset is loaded and read properly, first and foremost we must clean the data using text preprocessing techniques and then use appropriate classifier algorithm to classify the posts as stressed or non stressed. Finally based on the severity of stressed text the model must identify if a stressed post is inclining towards suicidal tendencies or not using unsupervised learning approach. We evaluate the accuracy of each algorithm at each step.



## 5. PROJECT IMPLEMENTATION

### 5.1 Project modules overview





## **5.2 Tools and Technologies Used**

Maintaining a project is a very difficult task in itself. Good for us, there are various tools that help us get things done and accomplish tasks. There are some tools that require a computer with the supporting software, while others can be used in a manual manner. A system that applies technology by taking any type of input and modifying it in accordance with the application of the system, and manufacturing results is known as a technology system. In this particular project, we have used many different kinds of technologies and tools for the completion of this project. Following is the list of Technologies that we have used in our project:

### **OPERATING SYSTEM : WINDOWS**

Windows is a functional graphics program which was made by Microsoft. It allows its users to view and save files, run various softwares, play different games, watch interesting videos, and provides a way to keep users connected to the Internet. It was made to be used by both home computer users and professional workers. It includes many features that help users. Windows offers a Control Panel feature that integrates many of the tools to configure and manage resources on its computer. Windows is easy to use as it provides a simple interface. Microsoft has improved security features on Windows in recent years.

### **LANGUAGE: PYTHON 3**

Python's a very advanced language of programming. Python is a high-quality, translated, interactive and focused language. It is built to be read easily and efficiently. It uses words of english more regularly as well as uses punctuations & has a lower grammatical structure than other languages. Supports efficient and systematic planning methods with OOPs. Can be used as a writing language or as a means of

integrating into a byte-code to create great applications. It provides superior quality data types & also provides dynamic type testing too.

It has automatic garbage collection. Python is a language which can be used with other languages such as C++, C, CORBA, and Java.

## NLTK (NATURAL LANGUAGE TOOLKIT)

NLTK is a medium for structuring various python applications to operate with human language. Provides easy-to-use sites for more than 50 companies and dictionary tools such as WordNet, as well as a list of text-based libraries, token, stamping, marking, marking, segmentation, and semantic thinking, NLP dynamic industrial libraries, and forum active chat. NLTK has been described as "an amazing tool for teaching, and working on, computer languages using Python," and "an amazing library to play in the native language." Python Indigenous Language Processing provides an effective introduction to the language processing system. Written by the different NLTK creators, it instructs the reader on the basics of writing Python programs, working with companies, classifying text, analyzing language structures, and more. The online version of this particular book has also been updated to Python 3 and NLTK 3.

## IDE: JUPYTER NOTEBOOK

A server client application called Jupyter Notebook App allows you to edit and use notebook documents with a web browser. The Jupyter Notebook Application can basically be installed on a desktop that does not require net (internet) access (as it is already described in this document) or it can also be installed on a different remote server and accessed through the Internet. Apart from editing or displaying or using the manual documents, the Jupyter Notebook Application has a "Dashboard", a "control panel" that displays local files and allows them to open brochure documents or close their characters. Jupyter Notebooks is an emerging project from the IPython project. The name, Jupyter, comes from a number of supported programming languages: Julia

Ships, Python, and R. Jupyter has the Python kernel, which basically allows the programmer to write your programs with Python, but currently there are over 100 characters you have.

## PUSHSHIFT API

The Pushshift.io Reddit API was designed and built by the / r / datasets mod team to improve functionality and allow you to search for Reddit ideas and posts. This RESTful API provides the full functionality of Reddit data retrieval, and includes the ability to create powerful data integrations. With this API, you can quickly find your favorite dates and find interesting links. Reddit There are two ways to access the comment and outbound database. One is to use the API directly via <https://api.pushshift.io/> and the other is to access the Elasticsearch search engine via <https://elastic.pushshift.io/>. This document describes both methods and provides examples of how to use the API effectively. This document also describes the use of API parameters for targeted searches.

## LIBRARIES

Pandas:

Pandas is basically a Python library for the analysis. Founded by Wes McKinney in 2008 due to the need for very powerful and volatile value analysis tools, Panda has become one of Python's most popular libraries. There is a very active donor community. Pandas is based on two major Python libraries. Matplotlib for displaying data and NumPy for math operations. Pandas acts as a threshold for these libraries, giving you access to many of the Matplotlib and NumPy methods with a little code. For example, the panda's `.plot ()` combines multiple matplotlib modes into one, allowing you to edit the chart in just a few lines.

Numpy:

It also has functions for working on line algebra, fourier transform, and matrices. NumPy was founded in 2005 by Travis Oliphant. Numpy is completely open source and can be used for free. NumPy stands for Numerical Python. In Python we have lists that serve the purpose of the array, but are slow to process. NumPy aims to provide something up to 50x faster than a standard Python list. NumPy's list item is called ndarray, it provides many support functions that make working with ndarray much easier. Arrays are widely used in data science, where speed and resources are very important.

Sklearn:

Scikit-learn (Sklearn) is a very useful and powerful mechanical library in Python. Provides a selection of effective machine learning tools and mathematical modeling that includes division, retranslation, merging and reduction in size with a virtual interface in Python. Scikit-learn is probably the most useful typewriter library in Python. The sklearn library contains many efficient machine learning tools and mathematical modeling that include division, regression, aggregation and size reduction.

## 5.3 Algorithm Details

### 5.3.1 Algorithm 1 -

This is the part where Reddit's Pushshift API is used to incorporate dynamic, real time data. Reddit is divided into communities called *subReddits*. Each subReddit has submissions that are posted by users and can be commented by other users and can be upvoted or downvoted.

In order to have in depth analysis of Reddit, we need to access all of its submissions, comments and users' information. To do this, we use an API called "PushShift".

Using PushShift's API we can either

"search\_comments": Which will return a list of comments associated with a search term.

"search\_submissions": Which will return a list of posts from the selected subReddit.

First of all we import the relevant modules like pandas, requests, json, csv, time and date time. Then, we can access the PushShift API through building an URL with the relevant parameters without even needing Reddit credentials. Now with parameters, we will access the depression subReddit, between 2 dates written in unix timestamps and search for all concerned submissions

The URL we created returns a JSON page of our results. In Python, JSON objects can translate to dictionary types and in this case would be held under the dictionary key "data", followed by a list of nested dictionaries

For our particular case, we use these parameters :

- size — increase limit of returned entries to a particular number
- after — where to start the search
- before — where to end the search
- title — to search only within the submission's title
- Body - to give the text of a particular submissions

Dept. of Mechanical Engineering Vishwakarma Institute of Technology, Pune

- subReddit — to narrow it down to a particular subReddit

Once we get our search results, we want key data for further analysis including: Submission Title, URL, Body, Author, Submission post ID etc. You can search for all submissions between any two time frames. There are other parameters such as subCount which tracks the number of total submissions we collect and Sub-stats which is the dictionary where we store our data.

Then we run a loop that gets all the data until the function returns 0 results. Once we have our data we can upload that into a CSV file for further analysis.

### 5.3.2 Algorithm 2 -

After thoroughly reading and going through a lot of posts from the dreaddit dataset, we observed that only those posts who had the usage of personal pronouns like I, he, myself, her, etc could be categorized as stressed or non stressed because only those posts were talking about a personal experience about individuals facing some actual issue. Rest of the posts who didn't contain any personal pronouns were either making a generic statement about some stressful issue, or about invitations for events organized to address such issues or statements that made absolutely no sense and contained unnecessary language. So, we referred to the former set of posts as non spam posts and the latter set as spam posts which brings us to the next part of our implementation - Spam Filtering.

After we receive posts from Reddit, we run the code to filter these posts into spam or not spam by generating two different datasets for the same. In order to accomplish this we tokenized all the posts and further used the method of POS tagging to tag each word of the posts in different parts of speech like pronouns, verbs, adjectives, nouns etc. Next, after tagging we differentiated the posts containing words tagged as proper nouns from posts containing no words tagged as proper nouns. We called the former set of posts as non spam dataset and the latter set of posts as spam dataset and saved these two datasets in CSV format.

Dept. of Mechanical Engineering Vishwakarma Institute of Technology, Pune



### **5.3.3 Algorithm 3 -**

After getting the filtered posts containing non spam texts, we next run our model on this dataset to classify posts into stress and non stress. But before that we incorporate text preprocessing.

In order to do this, we first cleaned all posts where regular expressions have been used to remove the punctuation marks, other unnecessary symbols, currency signs, website links etc. Following this, we did casing wherein we converted all characters in either of the cases - Lowercase or Uppercase so that all words are treated equally. After that was the stopwords removal process. Stopwords are english words which do not add much meaning to a sentence and can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc. So, such words were removed for our convenience. Next we did lemmatization. A lemma is the canonical form or dictionary form of a set of words. For example, runs, running, ran are all forms of the words run, therefore run is the lemma of all these words. So, all the words were converted into their root form. Next, we implemented the TF-IDF vectorizer to transform text into machine understandable representation of numbers used to fit ML Algorithms for prediction. It basically counts different words' number of occurrences in different statements. TF-IDF considered the overall document weightage of the word and hence it helped us to deal with most frequent words like anxiety, ptsd, stress, etc in our case.

After all of this, we incorporated three different classifiers - Naïve Bayes, SVM and DT and adjusted their parameters so as to get the desired result of classification of posts into stressed and non-stressed. The performance metrics were decided and finally the precision, recall, fi score and support values for all stressed and non-stressed posts, using the three classifiers were displayed.

### **5.3.4 Algorithm 4 -**

Next part of our implementation is analysing the severity of stress to check whether it is inclining towards suicidal tendencies. For this we took the help of two labelled datasets from the subreddits - casual conversations and suicidal watch and using these posts, we trained our model in order to predict whether a stressed post is inclining towards suicidal tendencies or not. We used the SVM model for its prediction and lastly, stored the results obtained in a CSV / Excel file. For its implementation as well, all the steps like cleaning, casing, stopwords removal, lemmatization, vectorization were done and finally the desired output was produced.

## 5.4 IMPLEMENTATION ASPECTS

Following are the various aspects of our implementation:

1. Text Preprocessing:

Steps included in text preprocessing after loading the dataset are:

- a. Cleaning: All the punctuations that are present as part of the text are deleted in this step. Python's string library contains some predefined collection of punctuations such as `'!"#$%&'()*+,-./:;<?@[\\]^_`{|}~'`
- b. Casing: This is one of the most common preprocessing stages, in which text is transformed to the same case, ideally lower case. However, you do not need to perform this step every time you work on an NLP problem because lower casing can result in information loss in some cases.
- c. Stopword removal: Stopwords are common words that are omitted from the text because they bring no value to the analysis. These terms seem to have little meaning. The NLTK library provided a set of terms that are considered stopwords in English. I me, my, myself, we, our, you, you're, you've, you'll, you'd, your, he, most, other, some, such, no, nor, not, only, own, so, then, too, very, can, will, just, don't, should, now, d, ll, m, o, re, ve, y, ain't, could, didn't] However, using the provided list as a stop word is not compulsory because they should be properly selected based on the project.

- d. Lemmatization: The stemming or diminishing of words to their root/base form is also known as the text standardization phase. Words like 'programmer,' 'programming,' and 'programme,' for example, will be stemmed to 'programme.' But then again, stemming has the disadvantage of compromising the meaning of the root form or not reducing the term to a suitable English word.

## 2. Feature Extraction:

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced sets of features should then be able to summarize most of the information contained in the original set of features. In this way, a summarized version of the original features can be created from a combination of the original set.

- 3. Support Vector Machine: SVM stands for Support Vector Machine and is one of the most widely used Supervised learning algorithms for Classification and Regression problems. However, it is mostly utilized in Machine Learning for Classification problems. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that new data points can be readily placed in the correct category in the future. A hyperplane is the name for the optimal choice boundary. The extreme points/vectors that help create the hyperplane are chosen via SVM. Support vectors represent the extreme examples, which is why the technique is called Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

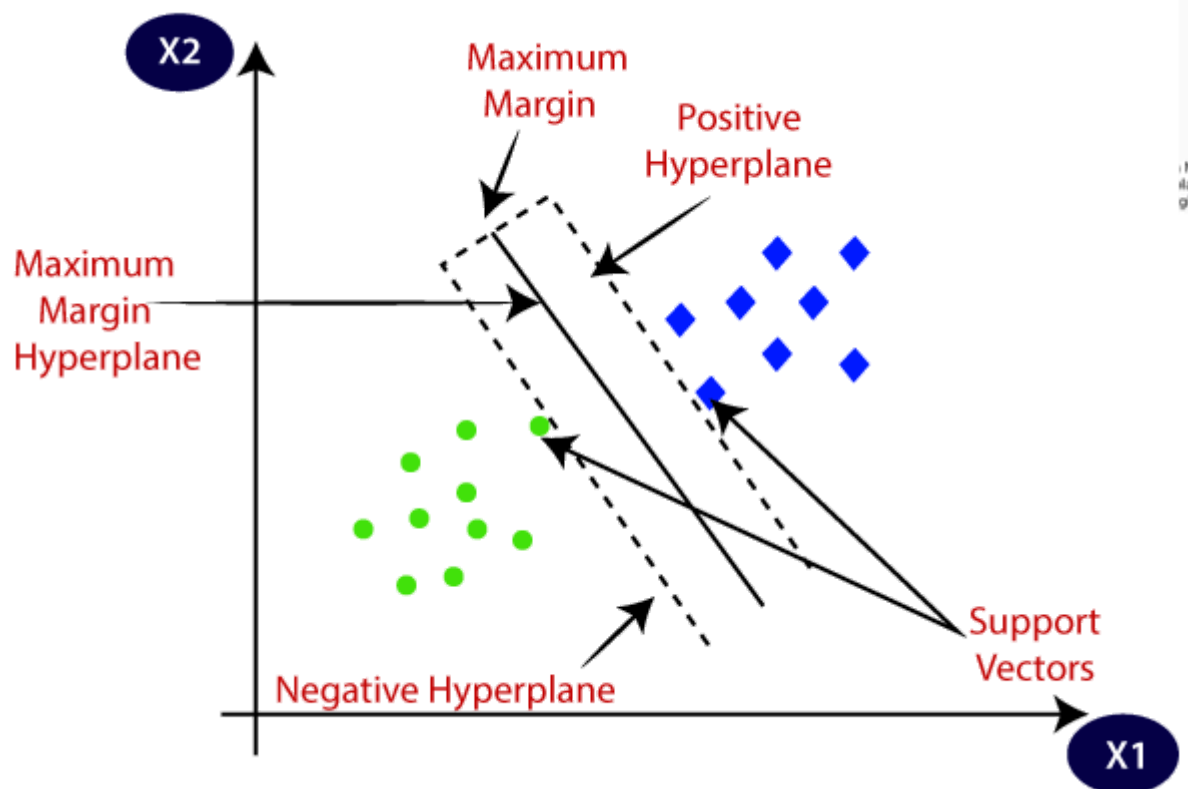


Fig no. 3. SVM

4. **Naive Bayes Classifier:** Naive Bayes Algorithm is a supervised learning algorithm which is based on Bayes Theorem and is used in classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other. Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

$P(A|B)$  is Posterior probability: Probability of hypothesis A on the observed event B.  $P(B|A)$  is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.  $P(A)$  is Prior Probability: Probability of hypothesis before observing the evidence.  $P(B)$  is Marginal Probability: Probability of Evidence.

5. Decision Tree: Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

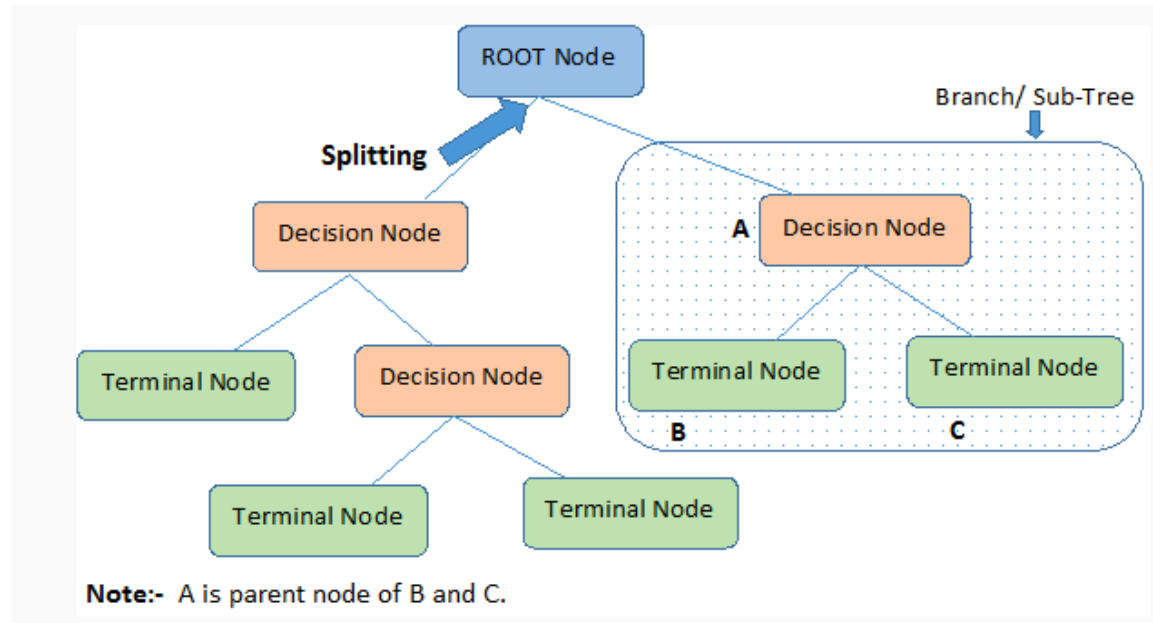


Fig no. 4. Decision Tree

Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example.

Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node.

6. For implementing the suicide classification model, a dataset containing clinically verified suicidal posts and clinically healthy posts from reddit has been used. This dataset is then trained using the SVM algorithm. The model will then classify any given text as suicidal or non suicidal.

**DATASET:**

Dreaddit is a dataset of extensive social media posts divided into five categories, each of which includes stressful as well as non-stressful texts and various ways of expressing stress, with a subset of the data elucidated by human annotators. The dataset used in this study was compiled by collecting Reddit posts from various subreddits such as interpersonal conflict, mental illness, and financial need, where people are likely to share their experiences with stress. Our dataset is made up of 3.5K posts from five different Reddit communities, which were also labeled by Amazon Mechanical Turk. Reddit is a social media website where users post in different subreddits, or topic-specific communities.

Because of the lengthy nature of reddit posts, it is considered as an excellent source of data for studying the nuances of phenomena like stress. To collect stress expressions, we chose subreddits where redditors are most likely to discuss stressful topics:

- Interpersonal conflict: (adversity and social domains) People who posted in the abuse subreddit are primarily considered as survivors of an abusive relationship or situations where they share their stories and offer support, whereas people who posted in the social subreddit discuss about difficulties in a relationship (often but not always romantic) and seek advice on how to handle the situation.
- Mental illness: (Post-Traumatic Stress Disorder (PTSD) and Anxiety domains) People who posted in these subreddits seek advice on how to cope with mental illnesses and its symptoms, different ways to handle such situations, seek diagnoses and extend support to such people among other things.
- Financial need: (Financial domain) Posters in financial subreddits typically seek financial or material assistance from other posters. We include ten subreddits from the five domains of PTSD, social, anxiety, abuse and financial that were scraped using the PRAW API between January 1, 2017 and November 19, 2018; 3.5k posts in total. In our dataset, the average length of a post is 420 tokens, which is much longer and extensive than most microblog data such as Twitter data.

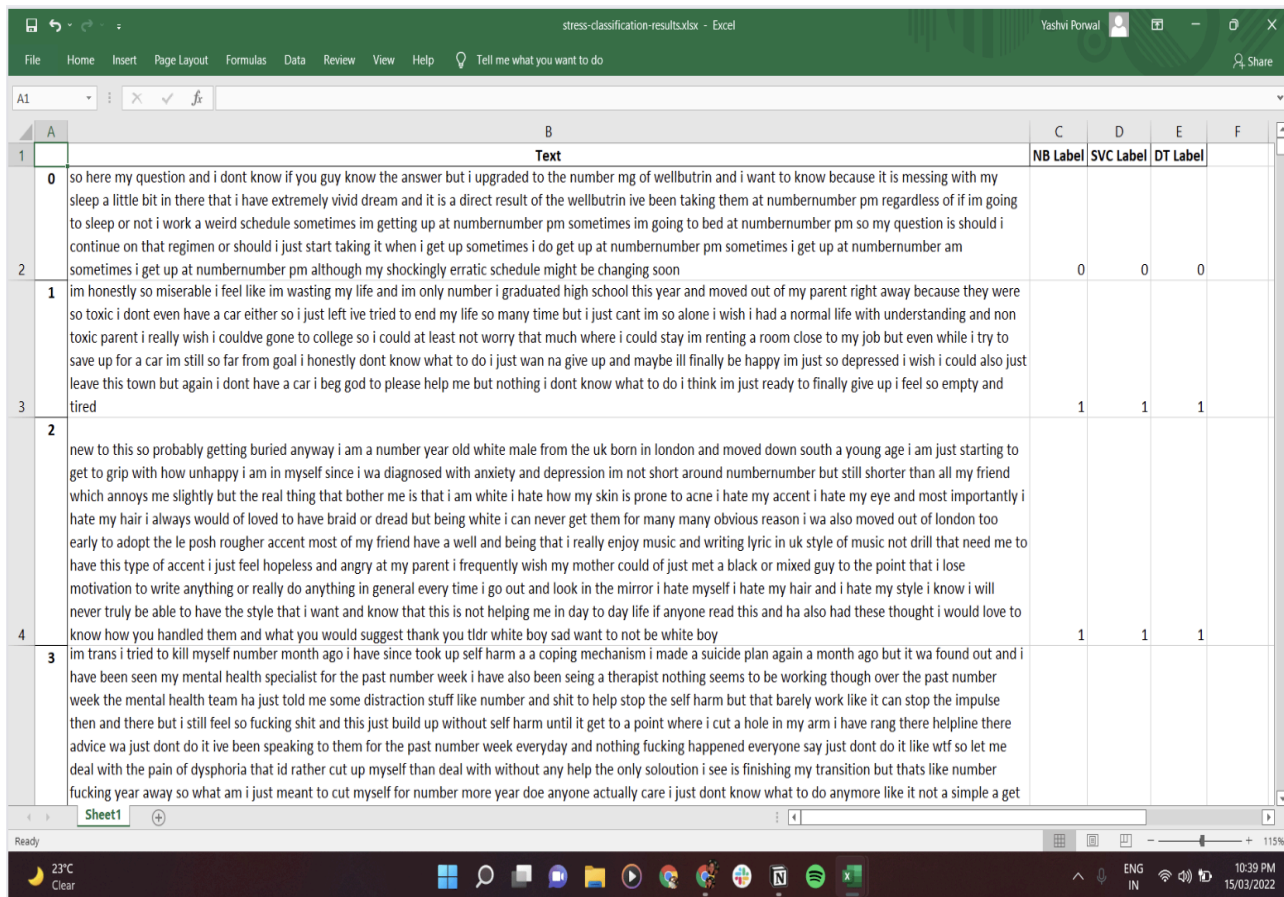
## 6. RESULTS

## 6.1 Outcomes

1. Analyzing Dreddit dataset
2. Incorporating dynamic and real time data
3. Classifying text into stress or non stress
4. Analyzing severity of stress and classifying it as suicidal or non suicidal.

## 6.2 Screenshots

- Stress classification results by three models



	A	B	C	D	E	F
		Text	NB Label	SVC Label	DT Label	
1	0	so here my question and i dont know if you guy know the answer but i upgraded to the number mg of wellbutrin and i want to know because it is messing with my sleep a little bit in there that i have extremely vivid dream and it is a direct result of the wellbutrin ive been taking them at numbernumber pm regardless of if im going to sleep or not i work a weird schedule sometimes im getting up at numbernumber pm sometimes im going to bed at numbernumber pm so my question is should i continue on that regimen or should i just start taking it when i get up sometimes i do get up at numbernumber pm sometimes i get up at numbernumber am				
2	1	im honestly so miserable i feel like im wasting my life and im only number i graduated high school this year and moved out of my parent right away because they were so toxic i dont even have a car either so i just left ive tried to end my life so many time but i just cant im so alone i wish i had a normal life with understanding and non toxic parent i really wish i couldve gone to college so i could at least not worry that much where i could stay im renting a room close to my job but even while i try to save up for a car im still so far from goal i honestly dont know what to do i just wan na give up and maybe ill finally be happy im just so depressed i wish i could also just leave this town but again i dont have a car i beg god to please help me but nothing i dont know what to do i think im just ready to finally give up i feel so empty and tired	0	0	0	
3	2	new to this so probably getting buried anyway i am a number year old white male from the uk born in london and moved down south a young age i am just starting to get to grip with how unhappy i am in myself since i wa diagnosed with anxiety and depression im not short around numbernumber but still shorter than all my friend which annoys me slightly but the real thing that bother me is that i am white i hate how my skin is prone to acne i hate my accent i hate my eye and most importantly i hate my hair i always would of loved to have braid or dread but being white i can never get them for many many obvious reason i wa also moved out of london too early to adopt the le posh rougher accent most of my friend have a well and being that i really enjoy music and writing lyric in uk style of music not drill that need me to have this type of accent i just feel hopeless and angry at my parent i frequently wish my mother could of just met a black or mixed guy to the point that i lose motivation to write anything or really do anything in general every time i go out and look in the mirror i hate myself i hate my hair and i hate my style i know i will never truly be able to have the style that i want and know that this is not helping me in day to day life if anyone read this and ha also had these thought i would love to know how you handled them and what you would suggest thank you tldr white boy sad want to not be white boy	1	1	1	
4	3	im trans i tried to kill myself number month ago i have since took up self harm a a coping mechanism i made a suicide plan again a month ago but it wa found out and i have been seen my mental health specialist for the past number week i have also been seing a therapist nothing seems to be working though over the past number week the mental health team ha just told me some distraction stuff like number and shit to help stop the self harm but that barely work like it can stop the impulse then and there but i still feel so fucking shit and this just build up without self harm until it get to a point where i cut a hole in my arm i have rang there helpline there advice wa just dont do it ive been speaking to them for the past number week everyday and nothing fucking happened everyone say just dont do it like wtf so let me deal with the pain of dysphoria that id rather cut up myself than deal with without any help the only soloution i see is finishing my transition but thats like number fucking year away so what am i just meant to cut myself for number more year doe anyone actually care i just dont know what to do anymore like it not a simple a get	1	1	1	

- Spam Filtering





WPS Office File Edit View Insert Format Tools Data Chart Workspace Window Help 90% Fri 13 May 00:14

Home suicide-classification-results Guest

Menu Home Insert Page Layout Formulas Data Review View Tools

Calibri 11 A<sup>+</sup> B I U Merge & Center Wrap Text General

Paste Copy Format Conditional Formatting Format as Table Symbol AutoSum Filter Sort Format Fill

B4 fx ive lost a lot of motivation over the last few month i really just need a break if i could get over the idea of upcoming bill and put my head down im sure i could get somewhere sadly it never a simple a that thats what im holding onto im not sure i could handle anything happening to them

	A	B	C	D	E	F	G	H	I
		Text	Label						
1									
2	0	so let me sparkle some context here in argentina it is tradition to travel after we finish secondary school our equivalent to the high school here it	0						
3	1	removed	0						
4	2	ive lost a lot of motivation over the last few month i really just need a break if i could get over the idea of upcoming bill and put my head down im	1						
5	3	no my family want me gone my ex husband is a douchenozzle and would use this against me if he found out i have no friend the only people wh	1						
6	4	i dont think so because your sacrifice wouldnt right wrong i understand your desire for secrecy certain event have left me with selective mutism	1						
7	5	i can try to fool myself it will work yes actually i feel that way most of the time self trust a self esteem never thought of it that way maybe it that i t	1						
8	6	to answer this properly id first have to define better well go with the standard tof a more excellent or effective type or quality with that being said	1						
9	7	it not the end it just feel that way or at least it doesnt have to be you have an entire lifetime to fix these thing and while i personally have never bc	1						
10	8	youre not being a bitch and you dont have to charm anyone youre depression and you may just want someone to listen to you there nothing you	1						
11	9	i had similar thought and when i wa young and i actually did some bad stuff to some people the thing that got me trough wa my endless imagina	1						
12	10	yes that is true but our parent are the only people who will honestly be there for u and love u no matter what maybe tell them youve been struggl	1						
13	11	to everyone who responded to this post thank you my name zack it wa very nice of you to try n stop me but it too late if this doesnt work a rope	1						
14	12	our situation sound very similar you are right we have to keep hope it just difficult when it seems so small i try my best to be his friend and be the	1						
15	13	gt no one give a single shitthat is factually incorrect we are a bunch of internet stranger who have nothing to gain by giving a shit about you yet v	1						
16	14	dont know there a dumb a it sound i feel hyperactive behavior i dont even deserve help this is the third time in a row that any relationship i have	1						
17	15	no gt and theyre causing you to break into tear this got me i neither viewed it a an inducable variablefact for my reason to create an image of my	1						
18	16	perhaps the way you are learning is the problem i know when i first started learning to program what helped me wa picking a project i wanted to	1						
19	17	i dont know what youre going through but i live my life number hour at a time some day those number hour feel like number youre going to pain	1						
20	18	trying to talk to my friend he drunk at another friend and ha no transportation what frustrates me the most is he said to call i needed him and whe	1						
21	19	then let explore that because selfhate isnt a simple thing ive been there i really have and it awful but you need to look at why you hate yourself s	1						
22	20	ive been thinking about the reddit community and how lovely you guy are in my opinion ha the least toxic community of any social medium platf	0						
23	21	there are so many people in the world youre not alonehttpwww.viruscomixcommonstrepenciesjpg and you have so much those people waiting t	1						
24	22	you need a sponsor like recovering drug addict but someone who is also struggling with mental illness being alone is the difference between life	1						
25	23	yeah it help a lot thanks where do you live im studying in an asian medical school though im just a beginner if possible i can help you out by intro	1						
26	24	she ha parkinson im so sorry to hear that about your mom i can not imagine we always think we have endless time and we never do i need to ap	1						
27	25	all i had wa her cellphone number so when i called them all they were able to do wa call her she didnt answer and got anger at me thats all there	1						
28	26	hey im free for a whilewhats the newsdo you want to talk about anything specific or do you wan na just chat about thing that are interesting or ur	1						
29	27	let me tell you something about life it not an easy road and it full och shit and fuckup but there more to it dont think about the past it will only dra	1						
30	28	i know there probably were and will be a million post like these but i just want people to know that theyre not alone people are not alone everyon	1						
31	29	i lost my job at a science museum when covid hit i wa crushed but my manager put me in contact with a parent looking for a tutor a the pandem	0						
32	30	ive tried to do work and have sat down and looked at it for hour but i just can not find a way to focus and get it done i want to believe thing will g	1						
33	31	rate your day out of number my day wa a numberrwork wa good load to do and the workhours flew by afterwards rushed home for to m	0						
34	32	just wondering i wa thinking of starting out and wa wondering what is the best possible site to get some exposure i know i wont have number or	0						
35	33	maybe those thing became le interesting maybe you should make plan that are very specific instead of seeing the world or traveling to africa ch	1						
36	34	ive had mental depression and severe anxiety mental depression disorder for seven year but ive managed by with antidepressant and therapy tl	1						
37	35	i have one question for you whatre your interest specific please you are not in prison my god you are not in a federal maximumsecurity prison cc	1						
38	36	my brother committed suicide three year ago his numberth birthday would have been tomorrow look someone who is serious about committing	1						
39	37	i get what you mean and no i dont have a single person ive been to the psych ward number time and have seen many professional putting on a	1						

Sheet1

Local backup on

## 7. CONCLUSIONS

### Conclusion

We presented a machine learning model that reads and classifies text based posts from

Dept. of Mechanical Engineering Vishwakarma Institute of Technology, Pune

social media platform - Reddit into stressful or non stressful content. We incorporated various unsupervised learning algorithms and compared their accuracy scores to outline the best classifier. We also made use of Reddit's real time data to make our research more dynamic. Further on we also studied intriguing relations between stress and suicide by analyzing the severity of stressful text and checking if they are inclining towards suicidal tendencies.

### **Future Work**

Analysis of our data and models shows that stress detection is a highly lexical problem that benefits from domain knowledge. We believe there is a scope of betterment and improvement. This dataset can be used in the near future to contextualize stress, and offer simplification using the content features of the text available. We could also analyze posts to understand the cause of stress and possible coping mechanisms adopted by people. In future, we could also try to extend this analysis to various social media platforms and make it centric towards an individual's social media activity. We intend to present our findings throughout the project by incorporating various data visualization techniques which will help to determine the interactive factors that influence mental health across people's lifespans.

### **Applications**

The work done here can have different kinds of applications in various areas which may include: diagnosing physical illness and diagnosing mental illness, determining public mood and worries in the area of economics and politics, and keeping a track of the effect of the disasters. It can also be used to detect early signs of depression in social media users and help them by providing useful support links. Use of such technology to monitor people's mental health online can help with a number of issues including the rise in teenage suicides.

### **Appendix A: Plagiarism Report**

## PAPER NAME

Project Report\_G11\_PR.pdf

## WORD COUNT

6652 Words

## CHARACTER COUNT

35117 Characters

## PAGE COUNT

36 Pages

## FILE SIZE

2.4MB

## SUBMISSION DATE

Jun 1, 2022 2:22 PM GMT+5:30

## REPORT DATE

Jun 1, 2022 2:23 PM GMT+5:30

**14% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 5% Submitted Works database

**Excluded from Similarity Report**

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less than 10 words)

Summary

**References:**

- [1] Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.
- [2] De Choudhury, M., Counts, S., & Horvitz, E. (2013, May). Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47-56). ACM .
- [3] Saravia, E., Chang, C. H., De Lorenzo, R. J., & Chen, Y. S. (2016, August). MIDAS: Mental illness detection and analysis via social media. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1418-1421). IEEE .
- [4] Rakshitha C L , Gowrishankar S. (2018).Machine Learning based Analysis of Twitter Data to Determine a Person's Mental Health Intuitive Wellbeing. In International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018)
- [5] Lee, Jieon & Park, Eunil & Han, Jinyoung. (2020). A deep learning model for detecting mental illness from user content on social media. Scientific Reports. 10. 10.1038/s41598-020-68764-y.
- [6] Tiwari, Pradeep & Sharma, Muskan & Garg, Payal & Jain, Tarun & Verma, Vivek & Hussain, Afzal. (2021). A Study on Sentiment Analysis of Mental Illness Using Machine Learning Techniques. IOP Conference Series: Materials Science and Engineering. 1099. 012043. 10.1088/1757-899X/1099/1/012043.
- [7] Aldarwish, M. M., & Ahmad, H. F. (2017, March). Predicting depression levels using social media posts. In 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS) (pp. 277-280). IEEE .
- [8] Burnap, Pete & Colombo, Gualtiero & Amery, Rosie & Hodorog, Andrei & Scourfield, Jonathan. (2017). Multi-class machine classification of suicide-related communication on Twitter. Online Social Networks and Media. 2. 32-44. 10.1016/j.osnem.2017.08.001.
- [9] Turcan, Elsbeth & McKeown, Kathy. (2019). Dreddit: A Reddit Dataset for Stress Analysis in Social Media. 97-107. 10.18653/v1/D19-6213.