



*The*  
BRITISH  
UNIVERSITY  
IN EGYPT

Faculty of Informatics and Computer Science  
Software Engineering

Automated Detection of Mental Disorders  
from Text using Natural Language Processing

**By: Mariam Yasser**

Supervised By

**Dr Amr Ghoneim**

**June 2023**

## **Abstract**

Nowadays, mental illness is highly prevalent and has a significant impact on individuals' well-being as well as society's health. It is a complex condition influenced by a range of factors including individual risk factors, economic circumstances, and clinical associations. To capture these relationships found in various forms of textual data such as social media posts, interviews, and clinical notes, natural language processing (NLP) techniques have shown promising advancements. These advancements aim to facilitate active mental healthcare and aid in early diagnosis. According to recent statistics, millions of people worldwide are affected by one or more mental disorders. Early detection of mental illness can have positive effects on disease progression and treatment outcomes. In the aftermath of the pandemic, there has been a significant increase in attention towards mental health, with people becoming more aware of its importance. Social media platforms have become outlets for individuals to express their emotions, share their life experiences, and reflect upon their struggles. Machine learning-based models traditionally relied on feature engineering and extraction, where valuable features needed to be manually identified. However, the use of machine learning frameworks has revolutionized this process by enabling models to automatically capture meaningful features without extensive feature engineering. This advancement has led to significant improvements in various fields, including computer vision, natural language processing (NLP), and signal processing. In the domain of mental illness detection from text, machine learning techniques have gained considerable attention in recent times and have demonstrated enhanced performance.

The detection of mental illness from text has become a subject of increasing research interest. Early identification of mental disorders plays a crucial role in enhancing mental health diagnosis. In this paper, we present an overview of the current research trends in this area, exploring various data sources and focusing on stress detection. Additionally, we summarize the existing machine learning techniques that have been employed for this task

# Turnitin Report

mariam192977

## ORIGINALITY REPORT

18%

SIMILARITY INDEX

11%

INTERNET SOURCES

8%

PUBLICATIONS

11%

STUDENT PAPERS

## PRIMARY SOURCES

1

Submitted to British University in Egypt

Student Paper

3%

2

[www.nature.com](http://www.nature.com)

Internet Source

1%

3

[www.coursehero.com](http://www.coursehero.com)

Internet Source

1%

4

Michael M. Tadesse, Hongfei Lin, Bo Xu, Liang Yang. "Detection of Depression-Related Posts in Reddit Social Media Forum", IEEE Access, 2019

Publication

1%

5

[www.researchgate.net](http://www.researchgate.net)

Internet Source

1%

6

Submitted to Study Group Australia

Student Paper

1%

7

Kantinee Katchapakirin, Konlakorn Wongpatikaseree, Panida Yomaboot, Yongyos Kaewpitakkun. "Facebook Social Media for Depression Detection in the Thai Community", 2018 15th International Joint Conference on

1%

## **Acknowledge**

First of all I would like to thank my friends for being always here and for helping me and encouraging me to become who I am today and achieving what I have always wished for and my family for being always here. Although, Dr. Amr for his great work and effort for supporting me from the first day and giving me all the information that I needed and more.

# Table of Contents

## Table of Contents

Abstract.....	2
Turnitin Report.....	3
Acknowledge.....	4
Table of Contents.....	5
1 Introduction .....	7
1.1 Overview .....	7
1.2 Problem Statement.....	8
1.3 Scope and Objectives.....	8
1.4 Report Organization (Structure) .....	9
1.5 Work Methodology.....	9
1.6 Work Plan (Gantt chart).....	11
2 Related Work (State-of-The-Art).....	12
2.1 Background .....	12
2.2 Literature Survey.....	13
2.3 Analysis of the Related Work.....	14
3 Proposed solution .....	18
3.1 Dataset.....	18
3.2 Pre-processing:.....	20
3.3 Feature Extraction.....	20
3.3.1 TF-IDF .....	20
3.3.2 Word2Vec .....	21
3.4 Models .....	22
3.5 Functional/ Non-functional Requirements.....	28
3.5.1 Functional Requirements.....	28
3.5.2 Non-functional Requirements.....	28
3.6 Design / Simulation set up .....	29
4 Implementation .....	32
4.1 Naïve Bayes.....	32
4.2 Random Forest.....	33
4.3 LSTM.....	33
4.4 SVM .....	34

5	Results and Discussion .....	35
5.2	Discussion.....	40
6	Testing and Evaluation .....	40
7	Conclusion and Future Works .....	41
	References .....	42

# 1 Introduction

## 1.1 Overview

Mental health refers to our emotional, psychological, and social well-being, all of which play a significant role in our daily functioning. It is as important as our physical health at every stage of life. Mental health can change over time due to various factors. When individuals face demands or challenges that exceed their coping abilities, their mental health may be negatively affected, potentially leading to mental disorders like anxiety, depression, post-traumatic stress disorder, bipolar disorder, mood disorders, psychotic disorders, eating disorders, dementia, autism, and other related conditions. Mental illness has multiple causes, rather than a single one. Mental illness has many causes, not just one. Various factors can contribute to the probability of mental illness, which can be influenced by different circumstances. These circumstances encompass early traumatic experiences during childhood, a history of exposure to violence, encounters with other serious illnesses like cancer or diabetes, biological factors or imbalances in brain chemicals, substance abuse involving drugs or alcohol, and feelings of isolation or loneliness.

Mental health has been taking a huge concern nowadays as people are more aware from social media about the causes and the symptoms of mental disorders and they can read and learn easily to understand it even if they do not have a background on the medical forms, social media shares awareness on the importance of taking care of your mental health in order not to have mental disorders and numerous research have discovered a connection between social media use and mental health problems, as people spend all their time on social media writing what they do, what they feel, and sharing what is happening in their days with the smallest details.

Chronic or excessive stress can contribute to the development or exacerbation of mental health conditions. Stress is a natural response to challenging or demanding situations, and in moderate amounts, it can even have some beneficial effects. However, when stress becomes chronic or overwhelming, it can have negative impacts on both physical and mental well-being.

Prolonged or severe stress can increase the risk of developing mental health disorders. Stress can also worsen the symptoms of existing mental health conditions and contribute to difficulties in coping with daily life.

It's important to note that while stress is not a mental illness itself, it is closely intertwined with mental health. Managing and addressing stress effectively is crucial for maintaining overall well-being and preventing the onset or progression of mental health issues. If you're experiencing persistent stress or noticing its negative impact on your mental health, it is advisable to seek support from mental health professionals who can provide appropriate guidance and assistance.

## **1.2 Problem Statement**

As Social media got mainly all the information on mental disorders and symptoms but people still do not know how to interact with the feelings that they have and the information that they got even though how helpful it may be, they may not know that they have a disorder or a problem concerning their mental health in general. So exploiting oversharing on social media will give a good impact on people's mental health to detect any mental illness by using what they are sharing on their profiles, especially on Reddit, Twitter, Tumblr and the platforms that are mainly depending on writing and sharing their daily thoughts. This is aiding doctors and psychiatrists in detecting stress which is a main symptom of mental illnesses in the early stages and this is decreasing the percentage of people having mental disorders and suicide percentage as suicide is one of the leading causes of death in young ages. Early detection will ease this and prevent turning bad feelings that they have into disorders that take more time to be healed and more chronic thoughts and behaviours that mental disorders cause and make them do without even being aware of what they are doing sometimes these people have no one to help them. Early stages of detecting mental illness will be so beneficial for decreasing the percentage of suicide, especially these days as social media is a double weapon.

## **1.3 Scope and Objectives**

This project aims to enhance the detection of mental disorders of people by detecting stress which is a main symptom across the majority of the mental illness through text using the mental health data set, different Machine learning models such as Naïve Bayes, Random Forest, LSTM, SVM which are Natural language processing (NLP) methods demonstrate promising improvements to empower proactive mental healthcare and assist early diagnosis. Although, using two feature extraction methods to transforming unstructured text into a structured format suitable for machine learning algorithm which are TF-IDF and Word2vec. Give a narrative review of the approaches, trends, obstacles, and future directions used to detect mental illness using NLP over the last ten years.



#### **1.4 Report Organization (Structure)**

- Section 1 of the report provides a general introduction to mental health and outlines the objectives, which include achieving the highest accuracy in abnormality classifications as the final step of the project. It also presents the proposed work methodology.
- Section 2 focuses on related work from other papers, summarizing their content in the background section and including any statistical or illustrative diagrams in the literature review. These papers are further analyzed in the work analysis subsection.
- In Section 3, the report discusses the proposed solution for stress, providing a detailed explanation of how the project's approach effectively addressed the problem statement.

#### **1.5 Work Methodology**

To implement an effective project you must follow a methodology that acts as a guide for efficient software development. There are many methodologies to follow but the most popular two are Agile and Water Fall. They are the most used and the most organized two that mainly everybody uses.

If you followed the agile methodology, Your project will go through several cycles over the course of its existence. the process of developing, reviewing, receiving comments, and finally approving the work item with a yes or no answer. If so, carry out and finish the task. If the answer is no, note it, make any adjustments that are required, monitor it, update the backlog or prioritisation to reflect the new information, and then proceed to the subsequent task or sprint.

The waterfall technique will be employed as it approach to progressing tasks through various stages. This methodology involves identifying requirements, designing the implementation, implementing the work item, verifying the implementation, performing quality assurance, and ultimately maintaining the feature. By adopting the waterfall technique, the project aims to simplify the task flow and ensure systematic progress.

As shown in figure1.1 the steps of the waterfall that we did here:

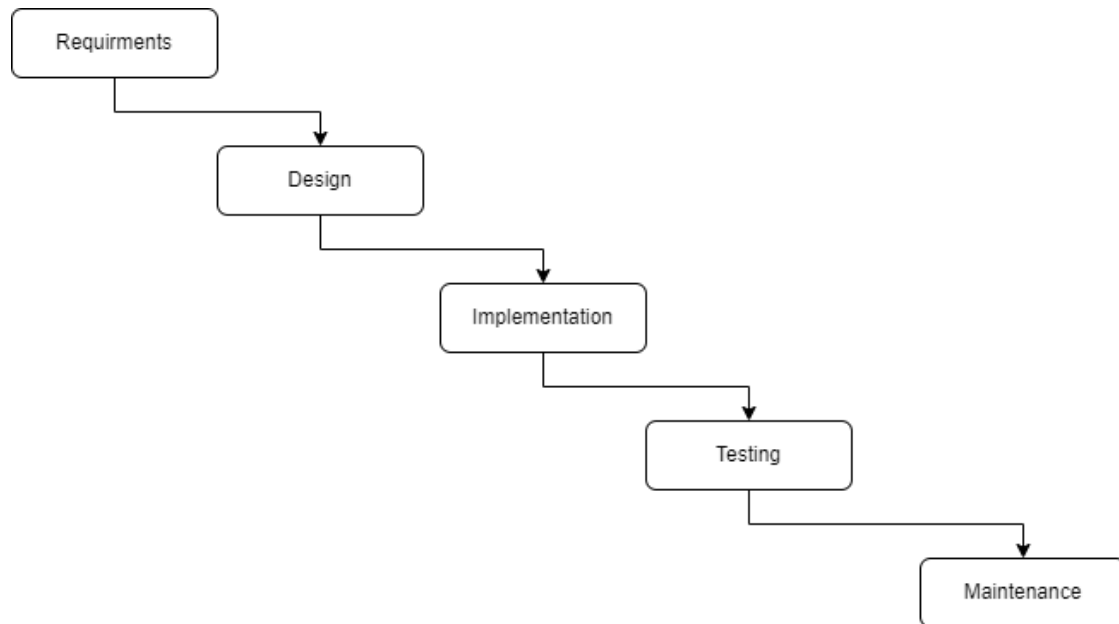


Figure 1

1.6 Work Plan (Gantt chart)

DETECTING MENTAL DISORDERS USING NLP

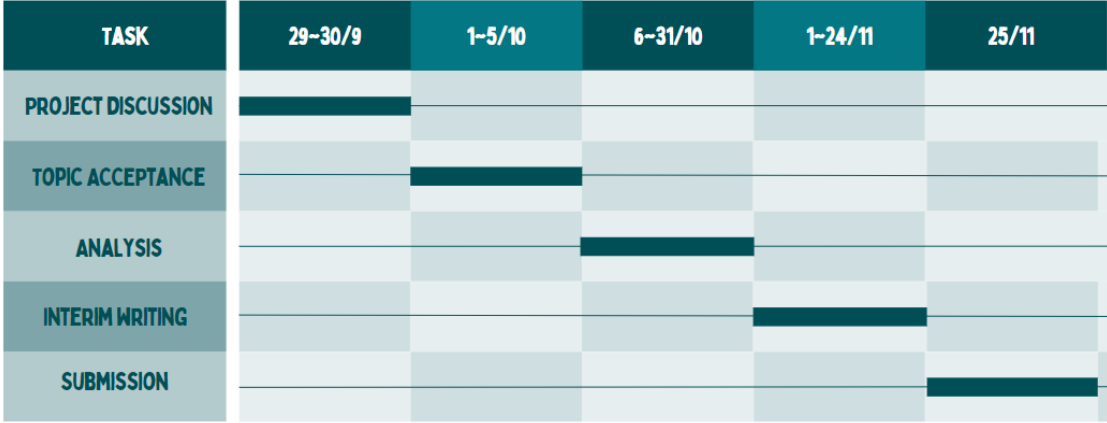


Figure2

## 2 Related Work (State-of-The-Art)

### 2.1 Background

Only a small percentage of those who need treatment for mental illness receive it. Early warning symptoms can be recognised in patients and promptly treated to prevent relapse and hospitalisation. Additionally, targeting demographically-based suffering clusters can help governments target each group with appropriate vigilance programmes, plan early-stage medical assistance for the parties in need, and allocate the resources required to lessen the burden of mental illness in their particular region. It takes initiative from those who are in need, access to medical treatments, and time commitment from qualified professionals to identify those with mental illness. It's possible that these resources won't always be accessible. It is customary to rely on clinical data, which are typically gathered after the sickness has manifested and been reported. [1][2] In addition, wanting to be isolated is a common characteristic in people with mental conditions. Due to this isolation, many with mental illnesses look for online spaces like Reddit and Twitter to publicly converse about their problems. These shared experiences turn into publicly available data that offers trustworthy perceptions of a user's character and personality through linguistic and behavioural patterns. Researcher interest has been ignited by such extensive data repositories in the fields of computational linguistics and psychology.[6] To avoid relapse and hospitalisation, patients can identify early warning signals and treat them right away. Additionally, focusing on demographically based suffering clusters can assist governments in identifying each group and implementing effective vigilance programmes, planning early-stage medical assistance for those in need, and allocating the resources necessary to lessen the burden of mental illness in their particular area. [3]Identification of persons with mental illness requires initiative from those who need it, availability of medical care, and time commitment from trained specialists. These resources might not be available at all times. Clinical data are often acquired after the illness has developed and been reported, and they are typically relied upon.[2]

## 2.2 Literature Survey

It has been concluded in this paper “Detection of Depression-Related Posts in Reddit Social Media Forum”, many researchers have introduced novel approaches to future healthcare solutions and techniques for early depression detection systems as a result of their online activity. By utilising several Natural Language Processing (NLP) methods. They attempted to use text classification methods to get a higher level of performance improvement. Multiple iterations of computational linguistics depression detection tasks are compiled in a meta-analysis by Guntuku et al. Calvo et al. have authored yet another intriguing review of mental health support and intervention in social media. whose analysis of the taxonomy of data sources, NLP methods, and computational. Through precise feature selection and their many applications, this research seeks to find a solution to a performance gain. combinations of features to characterise the content of the posts, we first select the best language features used for a depression diagnosis. Second, we examine the correlation strength, hidden topics, and word frequency that were taken out of the text. Studies on depression and other mental health illnesses have faced additional difficulties as a result of the rise of social media and the Internet. With a rich amount of text data and social metadata to capture users' behavioural inclinations, online spaces like Facebook, Twitter, and Reddit have offered a new platform for innovative study. We only take into account words for subject selection if they appear in at least 10 or more postings. Every post is to be further tokenized and stemmed as one document. This method enables us to compute the themes across the collection of documents and annotate them in accordance with the subjects that are detected. All stop words are eliminated prior to the topic modelling procedure.[12]

It has been concluded in this paper “Facebook Social Media for Depression Detection in the Thai Community” that mental illness is treatable. The duration of treatment would be shortened with early detection and intervention. Sadly, the rate of access to treatment is low. According to reports, fewer than 50% of people with this mental condition received mental health services. Lack of understanding and awareness of depression, a poor perception of mental health services, and a shortage of mental health professionals are some of the challenges. A cutting-edge technology and proactive approach must be employed to help raise the rate of accessibility to mental health services. Over the course of a year, information about Twitter users' social involvement, emotions, linguistic preferences, ego networks, and mentions of antidepressant drugs was gathered. They discovered certain noteworthy online behaviour that may be used to anticipate the development of depression. Table II displays the evaluation findings. The SVM model's accuracy was marginally higher than the majority vote which served as the benchmark for appraisal. The findings of the trial indicate that depression might be predicted using Facebook behavioural data, including messages and actions. However, because Facebook has restricted their ability to acquire personal information and the procedure for doing so has grown more challenging, the sample size of this research is quite small. As a result, not all significant aspects may be included in the study's findings. Additionally, as it was necessary to translate language-related features from Thai to English to analyse the process, some errors may have occurred throughout this procedure. It's possible that crucial sentiment polar words were dropped during translation.[10]

It has been concluded in this paper “Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing”. The numerous text posts are beneficial. determine the causes of stress and whether a user is exhibiting symptoms of mental stress. Using machine learning techniques, we conducted tests in this research to find stressful social media posts. This paper's key contribution is the significant results obtained for text classification and stress identification by combining a variety of embedding approaches with well-known machine learning models. The labelled corpus is used to train a model that can distinguish between stressful and non-stressful texts and make precise predictions. SVM to enhance social media mental stress analysis. The purpose of this study, according to the authors, is to provide the groundwork for future studies investigating neural network-based models and pre-trained language models for mental stress analysis. Additionally, the creation of standards Datasets for Reddit that are comparable to those for other social media platforms may be useful in classifying postings into stress levels and gaining more understanding of the effects of mental stress on users of social media. Overall, the suggested model has the potential to significantly reduce mental health issues among the majority of social media users by recognising the symptoms and offering help and support to deal with them.[11]

### **2.3 Analysis of the Related Work**

No.	Year	Name	Dataset	Algorithm	Objective
1	2022	Using Social Media for Mental Health Surveillance: A Review	“Reddit Self-reported Depression Diagnosis (RSDD) dataset.	1. NLP 2. Machine learning	In this article, we examine the research on using social media text for mental health monitoring. In the subsections that follow, we'll talk about the value of mental health surveillance with an emphasis on depression and suicide.
2	2020	A deep learning model for detecting mental illness from user content on social media	Posts of Mental-health-related Subreddits	1. SMOTE 2. CNN 3. NLP	In this work, a deep learning model was created to analyse posting data to determine a user's mental condition.
3	2021	Mental Health Intent Recognition for Arabic-Speaking Patients Using the Mini International Neuropsychiatric Interview (MINI) and BERT Model	Tunisian Darija	1. MINI 2. BERT models 3. ML 4. NLP	In this work, a deep learning model was created to analyse posting data to determine a user's mental condition.
4	2022	Natural language processing applied to mental illness detection: a narrative review	Reddit Self-reported Depression Diagnosis" (RSDD)	1. SVM 2. AdaBoost 3. NLP	The terms "machine learning approaches" in this research refer to both deep learning-based methods and traditional feature engineering-based methods.

5	2022	Detecting the presence of mental illness using NLP sentiment analysis	Sentiment Dataset Analysis and Visualization:	1. NLP 2. ML	It discussed how to forecast a user's mental condition while still protecting their privacy by using data mining from numerous sources, including social media and mobile devices.
6	2017	Natural language processing in mental health applications using non-clinical texts	The gold standard	1. Text classification 2. NLP	In this study, we explore the potential of NLP methods in the field of mental health.
7	2018	Facebook Social Media for Depression Detection in the Thai Community	social engagement, emotion, language styles,	1. NLP 2. SVM 3. Random Forest	We wanted to create a new psychiatric tool called a detection algorithm. This study tested whether algorithm could tell from a person's Facebook posts if they are depressed or not.
8	2023	Machine Learning-Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing	mental_health	1. ML 2. NLP 3. XGBoost 4. NB 5. SVM 6. 5. LR	In this study, we create Social Network Mental Disorder Detection, a machine learning framework for identifying SNMDs (SNMDD)
9	2016	Detection of Depression-Related Posts in Reddit Social	Reddit dataset	1. Ada 2. Boost	In this study, we looked for emotional performance



		Media Forum		3. SVM 4. NLP 5. SVM	enhancement strategies for depression identification and attempted to detect the existence of depression in Reddit social media.
10	2019	Depression Detection by Analysing Social Media Posts of Users	SNS users	1. SVM 2. NB 3. ME 4. NLP	In this study, the user's social media posts are analysed using a machine learning approach to determine the user's level of depression.

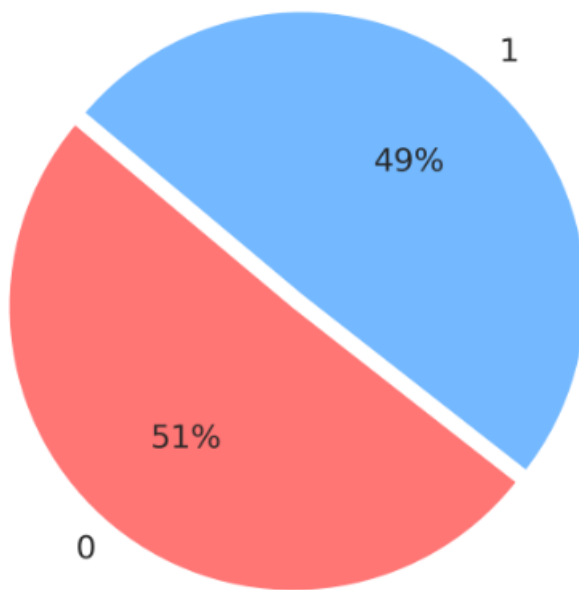
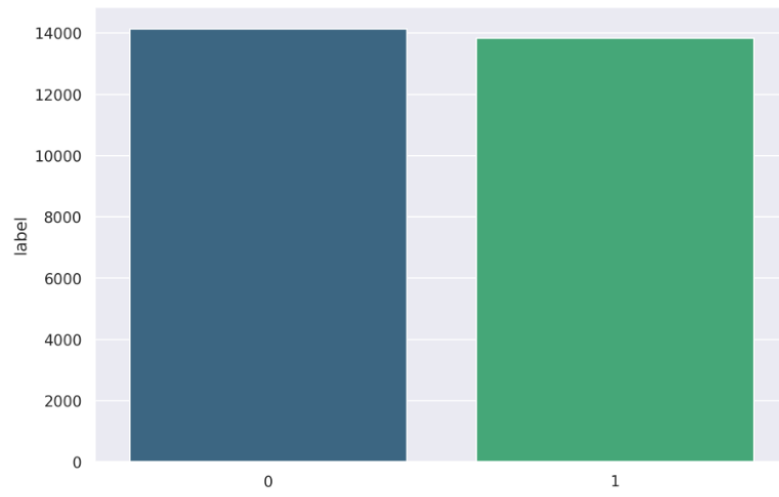
### 3 Proposed solution

#### 3.1 Dataset

In this report "Mental Health" dataset will be used that is available on Kaggle, created by Reihan Enamdar, and is designed to facilitate research and analysis in the field of mental health. This dataset contains text data related to people with anxiety, depression, and other mental health issues. The corpus collected from Reddit posts consists of two columns: one containing the comments, and the other containing labels indicating whether the comments are considered diagnosed or not. The corpus can be used for a variety of purposes, such as sentiment analysis, toxic language detection, and mental health language analysis. The data in the corpus may be useful for researchers, mental health professionals, and others interested in understanding the language and sentiment surrounding mental health issues. The dataset contains 27,978 instances.

1	text	label
2	dear american teens question dutch person heard guys get way easier things learn age us sooooo thth graders like right guys learn math	0
3	nothing look forward lifei dont many reasons keep going feel like nothing keeps going next day makes want hang myself	1
4	music recommendations im looking expand playlist usual genres alt pop minnesota hip hop steampunk various indie genres artists people like cavetown ali	0
5	im done trying feel betterthe reason im still alive know mum devastated ever killed myself ever passes im still state im going hesitate ending life shortly aft	1
6	worried year old girl subject domestic physicalmental housewithout going lot know girl know girl etc let give brief background known girl years lives uk liv	1
7	hey rredflag sure right place post this goes im currently student intern sandia national labs working survey help improve marketing outreach efforts many	1
8	feel like someone needs hear tonight feeling right think cant anything people keep puting listen this its your life everyone else living it someone tells unable	0
9	deserve liveif died right noone would carei real friendsi always start conversations get dry responses i feel comfortable around females emotional abuse n	1
10	feels good ive set dateim killing friday nice finally know im gonna it bye	1
11	live guiltok made stupid random choice its getting me basically molested relative super erratic thing haunting right now random walk home randomly assa	1
12	exercise motivated ngl cant wait get shape know gonna overnight im happy right now	0
13	know youd rather laid big booty body hella positive cuz got big booty	0
14	even time fuck supposed mean	0
15	usual hollywood stereotyped everyone movie but one classic uptight white collar banker russian woman well done even facial expressions great language	0
16	think it nearly unbelievable film could made death penalty one worlds controversial topics offends neither against testament tim robbins extraordinary int	0
17	trying rd time k krma special	0
18	guy coming sure wear f hey guy friend coming tomorrow im excited im sure wear ive known since middle school weve talking couple months and honest k	0
19	one best episodes entire xfiles series creepy beyond words tension suspense episode well executed entire minutes managed almost scary entire movie ep	0
20	good byehey you know sure hell know me goodbye probably mean anything you plus bother read rules sub may may taken down hard getting harder weak	1
21	tried put sugar coffee back spoon happy monday everyonestay safe sunflowers one days	1
22	sure leave noteso im struggling place leave note obvious family die find read it especially includes instructions funeral i want wasting money it shit like that	1

Datasets were split into 20% training and 80% testing, the testing contains 5596k instances and the training contains 22382k from Reddit posts. The label column is the classification target class where 0 indicates that there is no stress, and 1 indicates that there is stress. Total of [0,14139] [1, 13838].



### 3.2 Pre-processing:

The pre-processing of the data is one of the most significant processes in machine learning. In this step, we clean the text to remove any extraneous noise and maintain only the relevant information. Previous research has shown that pre-processing can improve classification outcomes.

The first thing that needs to be done is to Convert all text to lowercase helps in normalizing the text by reducing the variation in character cases. It ensures that the model treats the same word or phrase in different cases as identical because even though "Paper" and "paper" have the same meaning, the vector model will treat them as two distinct words, improving the accuracy. Then eliminating any extraneous information from the text, such as URLs, and punctuation. These are essentially noise to the model and don't provide any useful information. After that eliminating stop words, which are often used in papers but indicate little significance. Lemmatization is used to reduce words to their base form and reduces the overall vocabulary size by grouping different inflected forms of a word.

Using stemming to change each word into its original form, so words like "learning" will become "learn" and "caring" will become "care." Tokenizing tweets is the process of dividing sentences into words, which is necessary in natural language processing (NLP) because it makes the model to comprehend the context of the words or interpret the meaning through an analysis of the provided word order. Keras Library will be used.

### 3.3 Feature Extraction

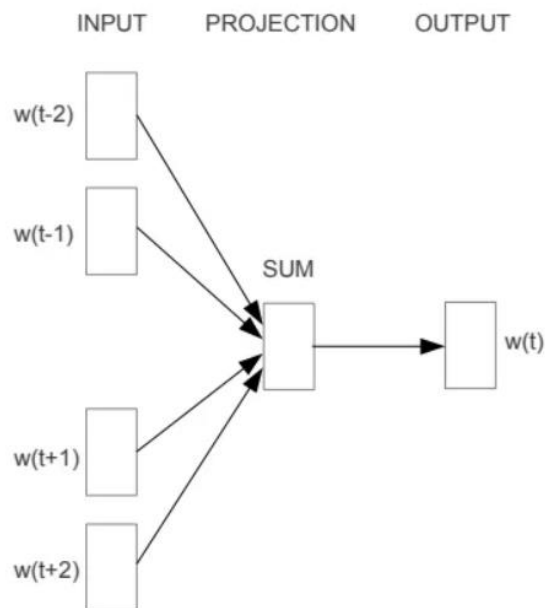
#### 3.3.1 TF-IDF

TF-IDF calculates a numerical value for each term in a document, indicating how relevant it is to the document and the overall corpus. The calculation involves two main components:

- The Term Frequency (TF) component counts the number of times a term appears in a document. It is computed by dividing the total number of terms in a document by the number of times a given term appears in that text. TF assumes that phrases that appear more frequently in a document are more important and gives them heavier weights.
- Inverse Document Frequency(IDF) component assesses a term's rarity across the whole corpus of documents. It is determined by dividing the total number of documents by the proportion of documents that contain the phrase. IDF gives less frequent phrases greater weights because it believes they contain more discriminative information.[15]

The TF-IDF score for a term in a document is obtained by multiplying its TF by its IDF. The final score reflects how significant the term is in the text and corpus. Terms with higher TF-IDF scores are thought to be more relevant to the content and are frequently employed for information retrieval, keyword extraction, and document categorization tasks.

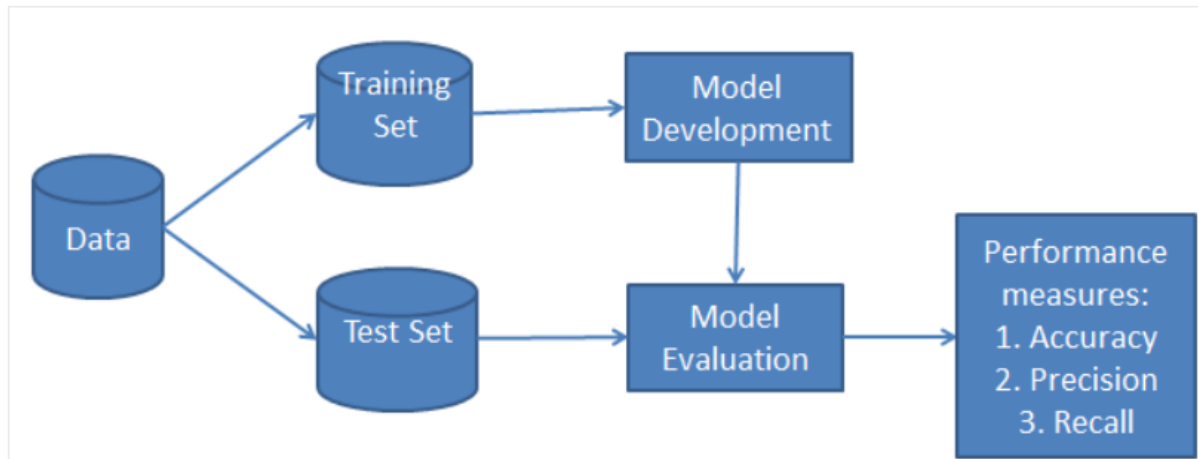
### 3.3.2 Word2Vec



It is a neural network-based model that learns word embeddings, which are dense vector representations of words. Word2Vec's goal is to represent words so that comparable words are clustered together more closely in the vector space. This is accomplished by using a huge corpus of text data to train the model. In order to forecast the likelihood of neighbouring words given a target word (continuous bag of words, CBOW), Word2Vec takes into account the context in which words appear. Word2Vec modifies the word embeddings during training. As a result, words that frequently occur in comparable circumstances have vector representations that are similar. The ability of Word2Vec to identify semantic connections and comparisons between words is one of its main benefits. This is accomplished by performing vector arithmetic operations on the word embeddings.[14]

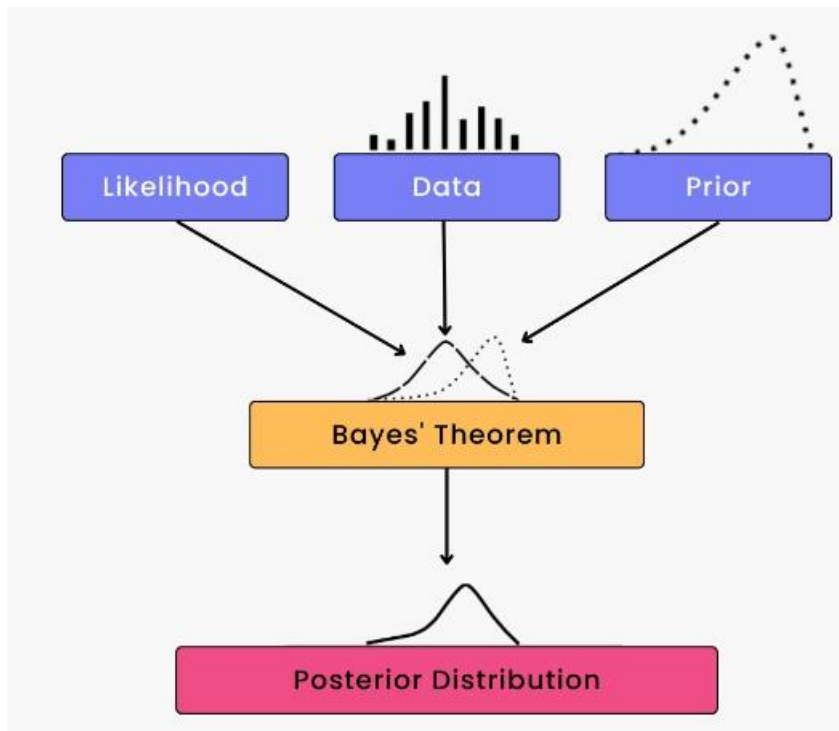
### 3.4 Models

For machine learning models and algorithms to accomplish the required classification task, they need to be provided with numerical inputs to grasp the provided classification rules. In light of this, feature engineering becomes crucial for text classification in NLP. Depending on the technique employed to modify the raw text and apply the embeddings, this can result in several benefits, including enhanced efficiency and making it simpler for the algorithm to find patterns. Models used in this paper are Naïve Bayes, Random Forest, LSTM, and SVM.



#### 3.4.1.1 Naïve Bayes:

Naive Bayes is a straightforward and easy-to-understand algorithm. The parameters needed for classification can be estimated with only a small amount of training data. It scales effectively with the number of training cases and features. Naive Bayes frequently works well in a variety of real-world applications, particularly in text categorization and spam filtering tasks, despite its simplicity. Naive Bayes can handle irrelevant features well since it makes the feature independence assumption. Calculating the naïve Bayes requires first determining the prior probabilities of each class based on the frequency of occurrence in the training data when given a training dataset with labelled examples. Naive Bayes determines the probability of each feature in the dataset occurring given each class. This is accomplished by determining the conditional probabilities of each feature in the training data for each class.[21]Next, Naive Bayes applies Bayes' theorem to combine the class probabilities and feature probabilities to determine the posterior probability of each class given the observed features. The classification step is the final step. After calculating the posterior probabilities, Naive Bayes chooses the class with the highest probability as the projected class for a particular instance.



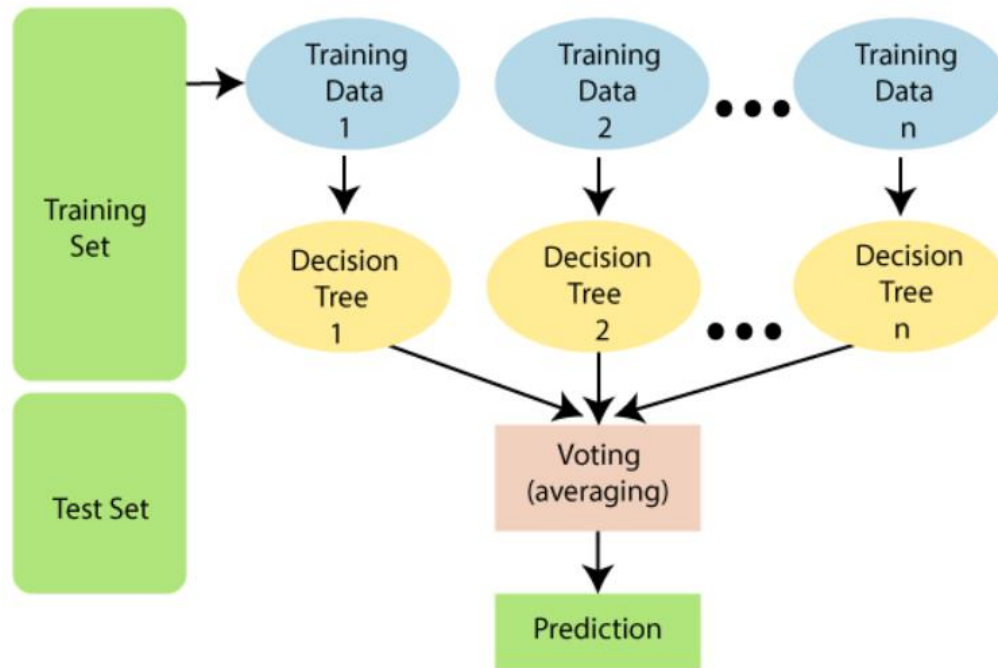
How it is calculated:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In conclusion, the ability of Naive Bayes to effectively handle datasets with text features makes it essential for stress detection. It is useful for real-time applications because of its efficiency and simplicity. Accurate classification is facilitated by the algorithm's capacity to represent the probabilistic links between words and stress labels. Naive Bayes may efficiently find signs of stress in textual data, offering insightful information. Because of its interpretability, users can better grasp the causes of stress, which facilitates intervention and support. Naive Bayes performs well in text classification tasks despite assuming feature independence, making it a useful tool for stress detection applications.

### 3.4.1.2 Random forest:

To produce more precise and reliable forecasts, it combines the predictions of various decision trees. A random subset of the training data is used to construct each decision tree in the random forest, and the forecasts of all the trees are combined to provide the final prediction. Continuous variables in dataset can be handled well with random forest.[20]



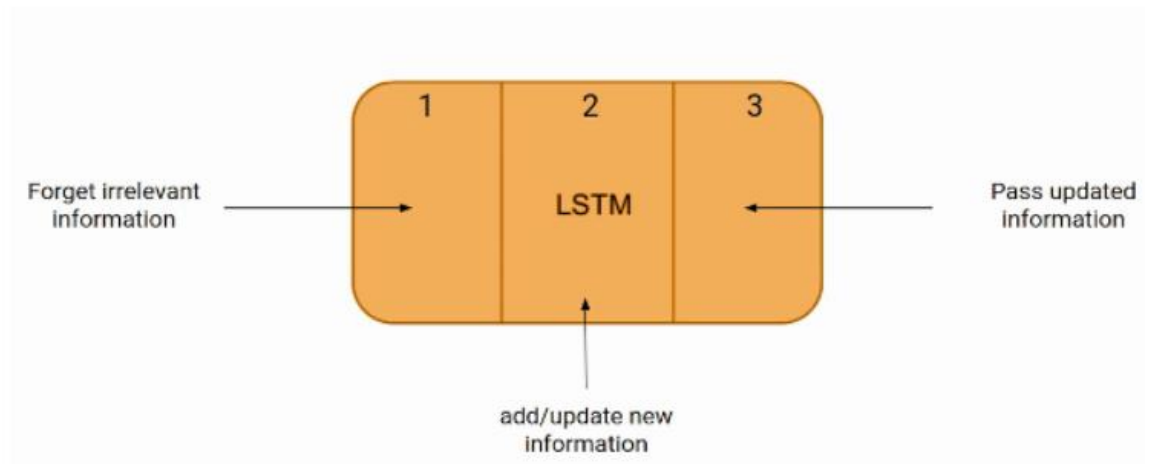
The first step is Data Preparation is a training set and a testing/validation set was created from the dataset. A set of features and an associated target variable make up each data point, Random Subset Selection is where A random subset of the training data is chosen for each tree in the random forest, often using a technique known as bootstrapping (sampling with replacement), Then the construction of the Tree by using the chosen subset of the training data, a decision tree is formed. A feature is selected at each node of the tree to divide the data according to some criterion (for example, Gini impurity or information gain). Up until the tree is entirely formed or a halting requirement is satisfied, this procedure is repeated recursively. The final step is Ensemble Prediction where the Predictions are made by each tree individually using the testing/validation set or out-of-bag samples (data not used during training), following the construction of all the trees. By combining the predictions from various trees, either through majority voting (for classification) or averages (for regression), the final prediction is established.

Overall, Random Forest is a valuable tool for identifying project stress due to its capacity for managing complex interactions, robustness against overfitting, feature importance analysis, handling noisy data, and provision of interpretability. It enables project managers to gather knowledge, make wise choices, and take preventative action to properly control stress levels.



### 3.4.1.3 LSTM:

Long Short-Term Memory (LSTM), is a recurrent neural network (RNN) architecture that is popularly used for time series analysis and sequence modelling tasks. Traditional RNNs frequently suffer from the vanishing gradient problem, which is addressed especially by LSTM networks. The basic goal of an LSTM network is to add memory cells that enable the network to retain data for extended amounts of time. Input, forget, and output gates, as well as other gates that regulate information flow, are present in these memory cells. LSTMs are capable of efficiently capturing and retaining pertinent patterns in sequential data by selectively updating and forgetting information.



It contains:

**Memory Cells:** The memory cell is an essential part of LSTM since it aids in the network's ability to store knowledge over time. The memory cell has a self-loop that enables it to maintain information while only updating or forgetting specifically relevant information.

**Gates:**

LSTM uses three different kinds of gates to regulate the information flow inside memory cells:

**Forget Gate:** Chooses which data to remove from the memory cell. It receives as inputs the previous hidden state and the current input and produces a forget vector that is element-wise multiplied by the prior cell state.

**The input gate** selects the fresh data to be stored in the memory cell. It is made up of two components: a tanh layer that generates a vector of potential new values and an input gate that selects which values should be updated. An update vector is produced by combining these two parts with a sigmoid layer.

**Output Gate:** Selects the data that should be output from the memory cell. To create an output vector, it combines the updated memory cell state, the prior hidden state, and the current input.

The following computing stages are performed by the LSTM for each input in the sequence:

- a. Input processing: The input gate and forget gate both take as inputs the current input and the prior hidden state. These gates determine which data should be kept or deleted from the memory cell.
- b. Memory Cell Update: The new memory cell state is decided by the input gate and the forget gate. The new memory cell state and the current input are combined to update the cell content.
- c. Output Generation: The memory cell's output gate chooses which data to output. The output is computed using the new memory cell state and the current input. The output can also be used for forecasts or additional processing before being passed on to the following time step as the concealed state.

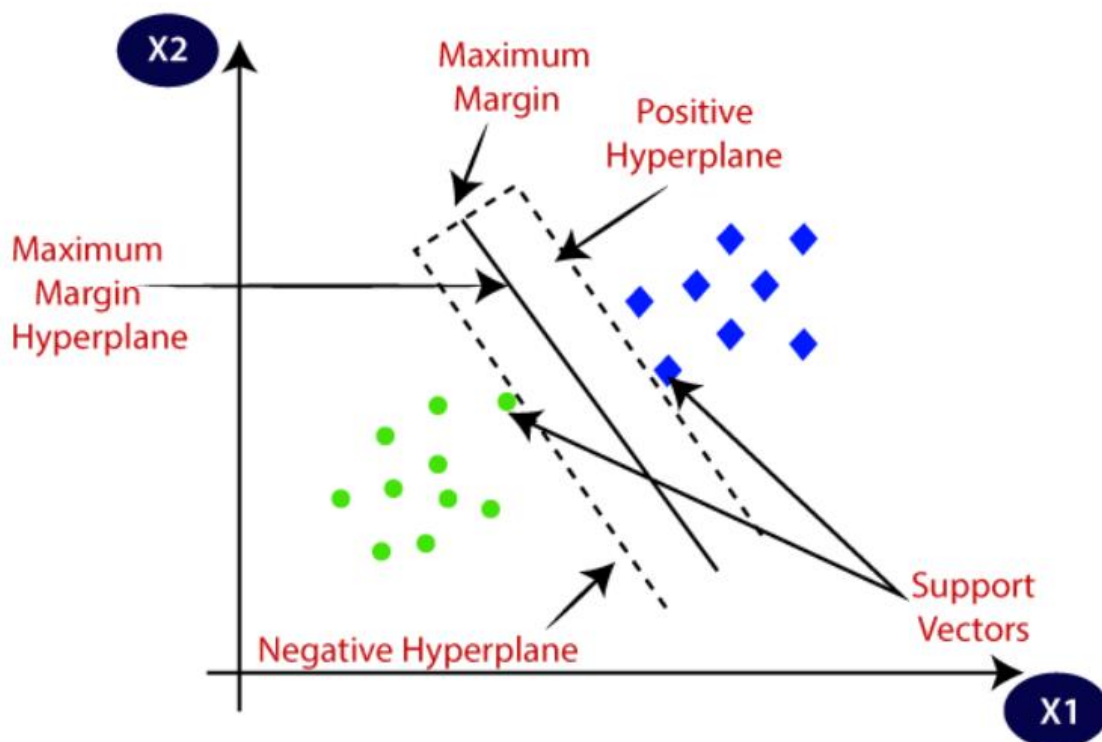
Backpropagation Through Time (BPTT): The backpropagation through time technique is used to train an LSTM network. In BPTT, the network's parameters are updated using an optimization approach like gradient descent after computing the network's gradients with respect to the loss at each time step.

Long-term dependencies in sequential data can be efficiently captured by LSTM by utilizing memory cells and gates. The gates regulate the information flow, enabling the network to update, delete, and output pertinent information only when necessary. Because comprehending context and long-term interdependence is essential for tasks like language modelling, machine translation, speech recognition, and sentiment analysis, LSTM is a good fit for these kinds of applications.[17]

Due to its recurrent nature, LSTM may take into account the full input, collecting relationships between words or events that could point to patterns associated with stress. The network can store information for extended periods of time because of the memory cells in LSTM. This is especially helpful in the identification of stress because stress-related indications may appear at various points in a sequence or may accumulate over time. These long-term dependencies can be captured by LSTM, and forecasts can take historical data into account.

#### 3.4.1.4 SVM:

SVM is utilized for classification and regression problems. It can be expanded to handle multi-class classification, its main use is the solution of binary classification issues. Finding the optimum hyperplane that separates the data points of distinct classes with the greatest margin is the basic goal of SVM. The hyperplane functions as a decision boundary for classification in the setting of fresh, unobserved data points.[18]The selection of the ideal hyperplane is the most crucial step in SVM. The hyperplane that maximizes the margin between the two classes, i.e., has the greatest separation between the nearest data points of each class, is the ideal hyperplane. Support vectors are the locations that are closest to the hyperplane, hence the term "Support Vector Machine." SVM employs a method known as the "kernel trick" to transform the initial input data into a higher-dimensional feature space, where it is simpler to discover a linear separation, in order to locate the best hyperplane. The kernel method implicitly maps non-linearly separable data to a higher-dimensional space, which enables SVM to handle it well.[19]



Data pre-processing: SVM needs training data that has been labelled, with each data point having a class label associated with it. Feature Selection: Identify the relevant features that are most informative for the classification task. The construction of the hyperplane that maximizes the margin between the classes is how the SVM learns the parameters during the training phase. The support vectors and hyperplane are located by the solution of an optimization problem. Testing Phase: After the SVM has been trained, it may be put to use to categorize fresh, unknown data points by determining which side of the hyperplane they fall on.[19]

Some key advantages of SVM include its ability to handle high-dimensional data, resistance to overfitting (if properly regularized), and effectiveness even with a small amount of training data. Understanding the context and patterns in a series of events or textual material is frequently necessary for stress detection.

### **3.5 Functional/ Non-functional Requirements**

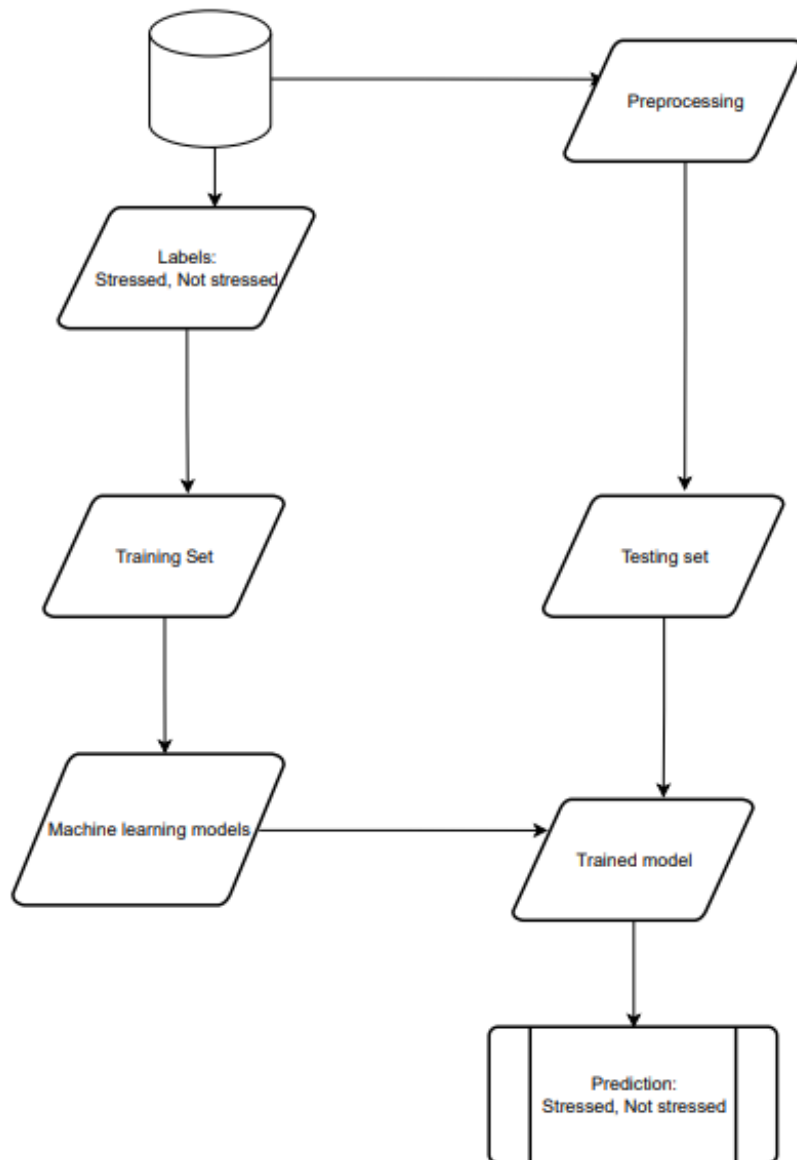
#### **3.5.1 Functional Requirements**

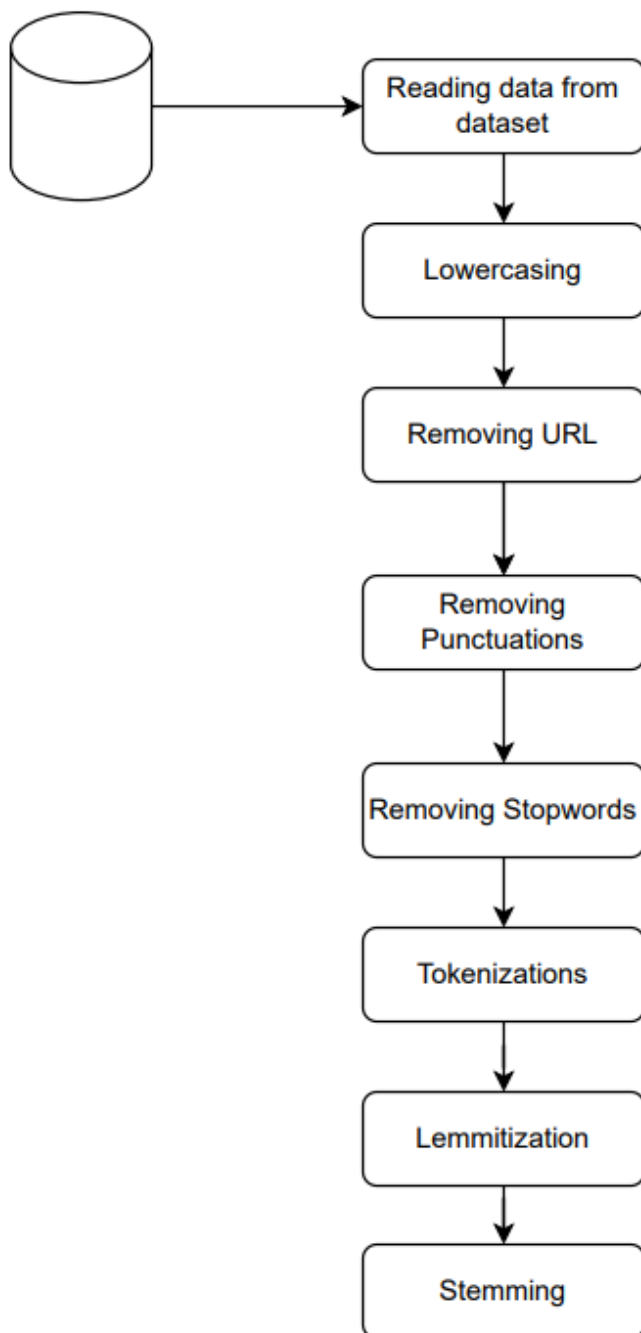
1. The user will find everything on the home page, it won't take much time to find the functionality that the user wants to find.
2. System will be fast, will take seconds to run the function.
3. The user will get a pop-up window after finishing each step to make sure that it is done.

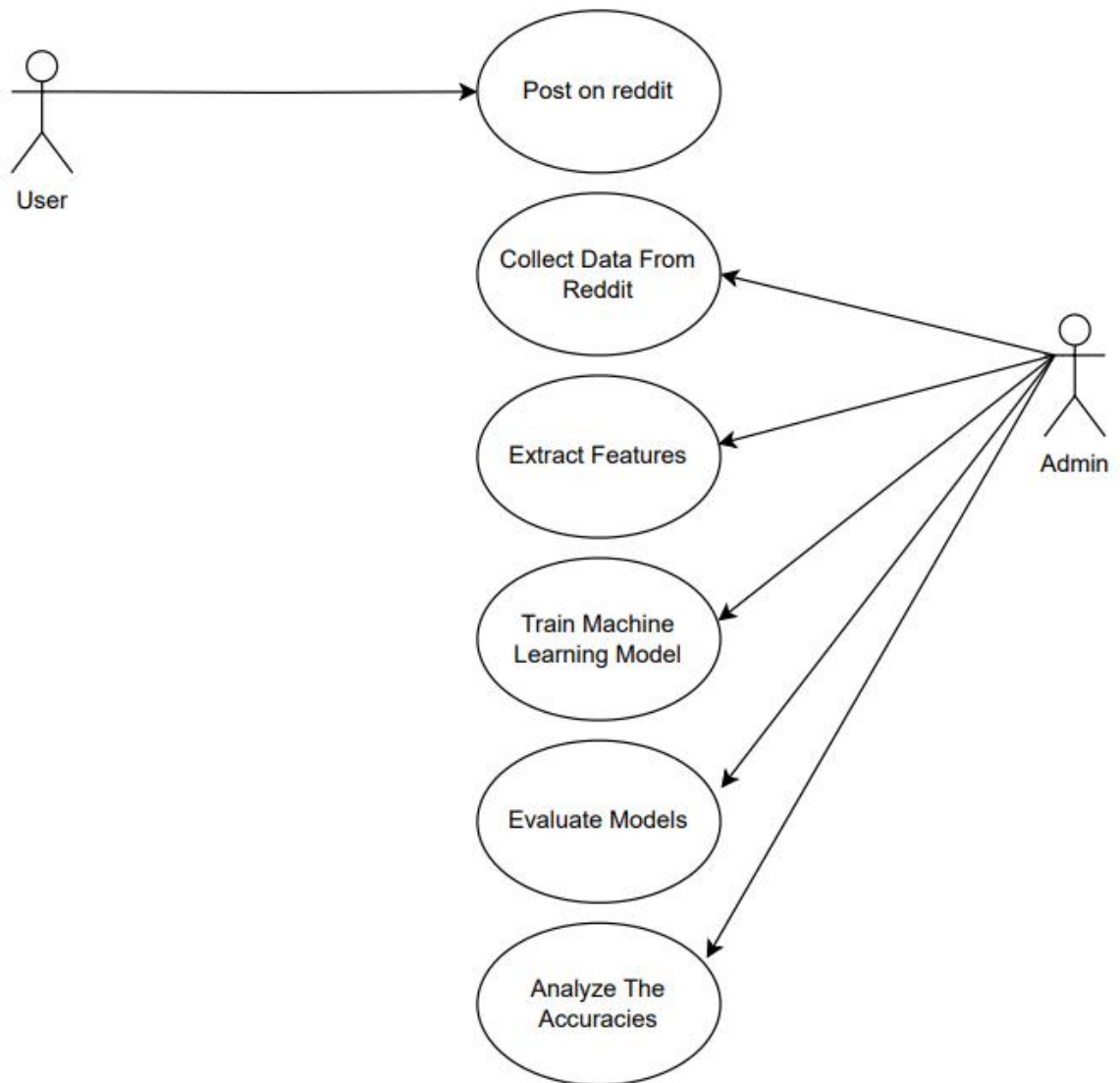
#### **3.5.2 Non-functional Requirements**

1. Usability: Making sure that the system is easy to use.
2. Testability: Testing the system after implementing new features to make sure that it is functioning well.
3. Completeness: Making sure that all the data inputs are completed.
4. Accessibility: Users can use the services in the system easily. Users can easily know what this system does without taking much time when reading the name of the service.

### 3.6 Design / Simulation set up



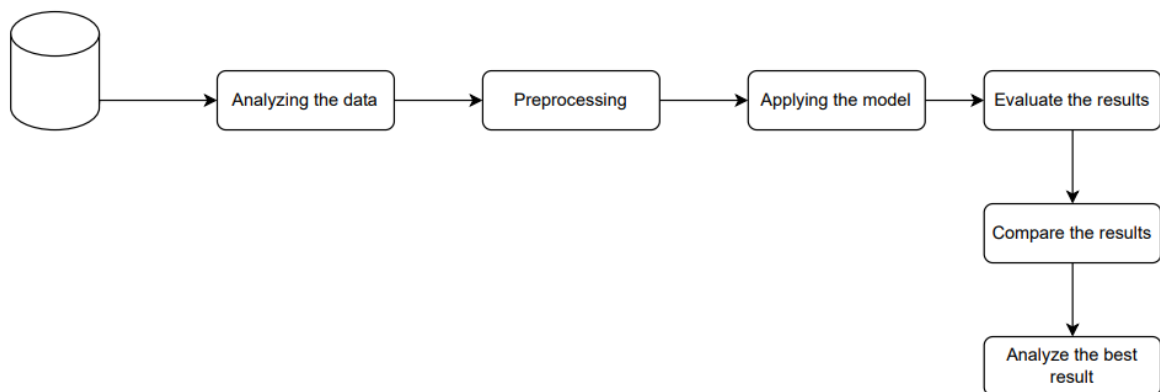




## 4 Implementation

This section will discuss the flow of the system of each situation and explain the algorithm of how the system works and get its results. The implementation of the system was done using Colaboratory using Python programming language.

First of all the dataset has been read, explored, and analysed by counting the instances and differentiating and counting the different labelled and mentioning what they mean and how to work which each individual. After that the pre-processing part is done, converting all letters to lowercase, removing URLs and punctuation, applying stemming, lemmatization, and stop words.



```
x = data["text"]
y = data["label"].values

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size= 0.2, random_state= 42, stratify = y)
```

This code prepares the data by separating the input features (text) and the target variable (labels) and then splits the data into training and testing sets, ensuring that the class distribution is maintained in both sets using the `train_test_split()` function from `sci-kit-learn`. It sets the random seed for reproducibility to 42, ensuring that the same split will be generated if the code is run multiple times.

**Applying the models on the pre-processed data after:**

### 4.1 Naïve Bayes

```
lr = MultinomialNB()
metrics_lr = train_model(lr, vector_size=vector_size)
print("")
plot_metrics(metrics_lr)
```

- first Creating a new Scikit-Learn `MultinomialNB` class object.
- The `MultinomialNB` classifier (nb) is trained using the training data by the `train_model` function.
- The result of this function is stored in the variable `metrics_lr` to be printed by `plot_metrics` function.



## 4.2 Random Forest

```
rf = RandomForestClassifier(n_estimators= 300)
metrics_rf = train_model(rf)
```

- n\_estimators setting it to 300, meaning that the random forest will consist of 300 decision trees.
- Train\_model function is responsible for training the RandomForestClassifier (rf) on the training data.

## 4.3 LSTM

```
vocab_size = 5000
embedding_size = 32
epochs=50
```

- vocab\_size represents the size of the vocabulary, which is the number of unique words in the text data, setting it to 5000.
- Embedding size represents the dimensionality of the word embeddings, setting it to 32.
- Epochs represent the number of training epochs, setting it to 50.

```
model.compile(loss='binary_crossentropy', optimizer='SGD', metrics=[tf.keras.metrics.Recall(), 'accuracy'])
print(model.summary())
```

- The loss function is set to binary\_crossentropy.
- The optimizer is set to 'SGD' (Stochastic Gradient Descent).
- The metrics used for evaluation are recall and accuracy.

```
es = EarlyStopping(monitor = 'val_loss', patience=5)
batch_size = 64

history = model.fit(X_trn, y_trn,
                    validation_data=(X_vld, y_vld),
                    batch_size=batch_size, epochs=epochs, verbose=1,
                    callbacks = [es])
```

- The EarlyStopping is used to monitor the validation loss and stop the training if the loss does not improve after a certain number of epochs
- The batch\_size specifies the number of samples per gradient update.
- Model.fit is called to train the model using the training data and validate it using the validation data.
- The training is performed for the specified number of epochs. The verbose parameter is set to 1 to display the progress during training.
- The history object captures the training history, including the loss and metrics values at each epoch.

## 4.4 SVM

```
for c in [0.01, 0.1, 1, 10, 100]:  
    svm = SVC(C=c, probability=True)  
    svm.fit(X_train[:1000], y_train[:1000])  
    y_pred = svm.predict(X_test[:1000])
```

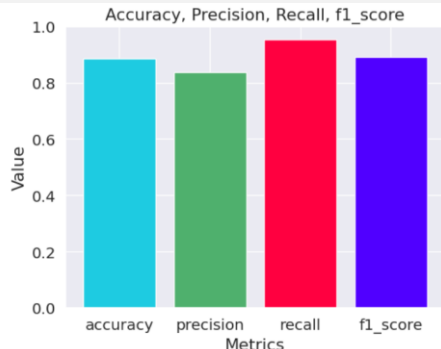
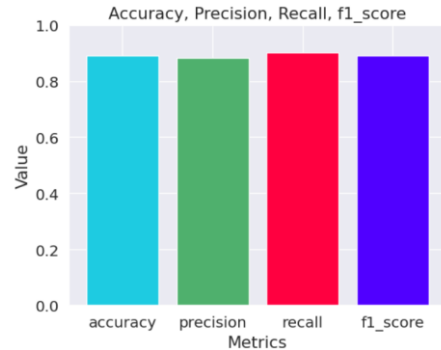
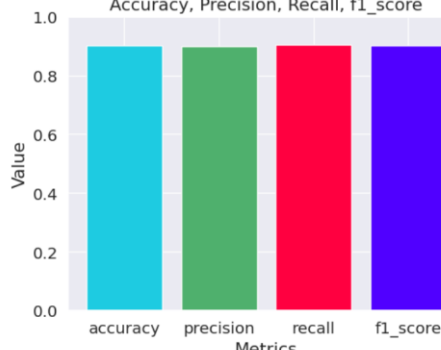
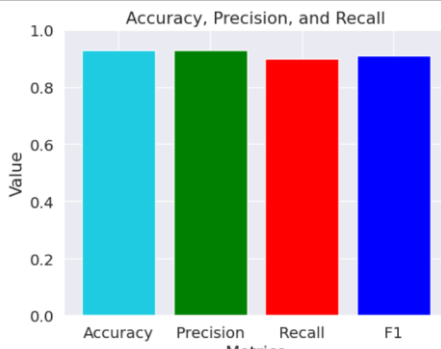
- First line initiates a loop that iterates over the list of C values ,allowing training the SVM classifier multiple times with different C values.
- For each iteration a new svm object is created.
- Fitting the training data into the classifier.
- Predecting the labels for the test set and storing it into y\_pred

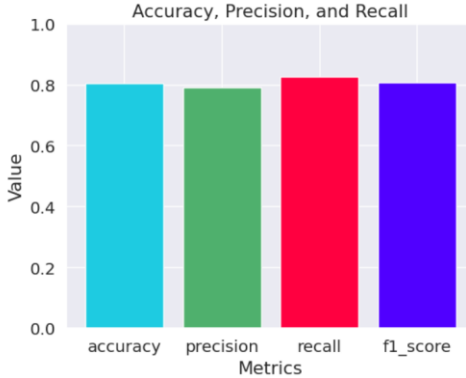
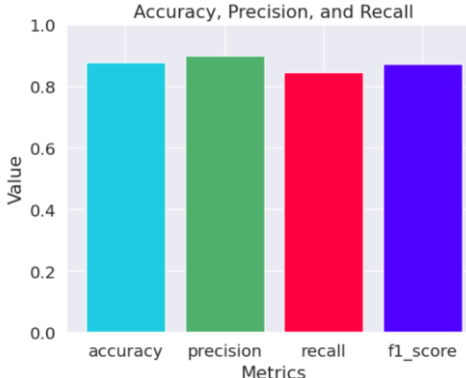
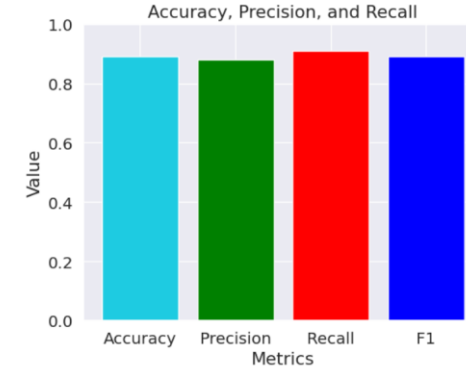
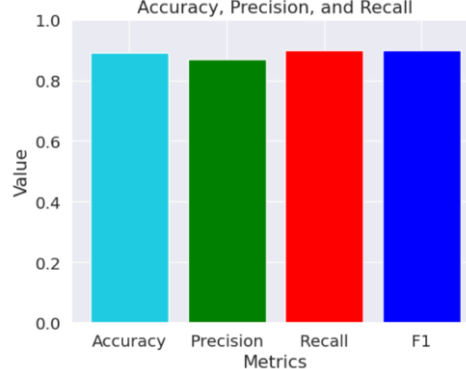
### Word2Vec

```
vector_size = 300
```

With Word2Vec we initialize the vector\_size and add it to the models.

## 5 Results and Discussion

Model Name	Results for TF-IDF	Graph
Naïve Bayes	Accuracy:0.88% Precision:0.84% Recall:0.95% F1:0.89%	
Random Forest	Accuracy:0.89% Precision:0.88% Recall:0.90% F1:0.89%	
LSTM	Accuracy:0.90% Precision:0.90% Recall:0.89% F1:0.90%	
SVM	Accuracy:0.93% Precision:0.90% Recall:86% F1:88%	
Models	Accuracy WITH Word2V	Graph

Naïve Bayes	Accuracy:0.81% Precision:0.79% Recall:0.83% F1:0.81%	 <table><thead><tr><th>Metric</th><th>Value</th></tr></thead><tbody><tr><td>accuracy</td><td>0.81%</td></tr><tr><td>precision</td><td>0.79%</td></tr><tr><td>recall</td><td>0.83%</td></tr><tr><td>f1_score</td><td>0.81%</td></tr></tbody></table>	Metric	Value	accuracy	0.81%	precision	0.79%	recall	0.83%	f1_score	0.81%
Metric	Value											
accuracy	0.81%											
precision	0.79%											
recall	0.83%											
f1_score	0.81%											
Random Forest	Accuracy:0.88% Precision:0.89% Recall:0.84% F1:0.87%	 <table><thead><tr><th>Metric</th><th>Value</th></tr></thead><tbody><tr><td>accuracy</td><td>0.88%</td></tr><tr><td>precision</td><td>0.89%</td></tr><tr><td>recall</td><td>0.84%</td></tr><tr><td>f1_score</td><td>0.87%</td></tr></tbody></table>	Metric	Value	accuracy	0.88%	precision	0.89%	recall	0.84%	f1_score	0.87%
Metric	Value											
accuracy	0.88%											
precision	0.89%											
recall	0.84%											
f1_score	0.87%											
SVM	Accuracy:0.89% Precision: 0.88% Recall:0.91% F1:0.89%	 <table><thead><tr><th>Metric</th><th>Value</th></tr></thead><tbody><tr><td>Accuracy</td><td>0.89%</td></tr><tr><td>Precision</td><td>0.88%</td></tr><tr><td>Recall</td><td>0.91%</td></tr><tr><td>F1</td><td>0.89%</td></tr></tbody></table>	Metric	Value	Accuracy	0.89%	Precision	0.88%	Recall	0.91%	F1	0.89%
Metric	Value											
Accuracy	0.89%											
Precision	0.88%											
Recall	0.91%											
F1	0.89%											
LSTM	Accuracy:0.89% Precision:87% Recall:0.90% F1:90%	 <table><thead><tr><th>Metric</th><th>Value</th></tr></thead><tbody><tr><td>Accuracy</td><td>0.89%</td></tr><tr><td>Precision</td><td>87%</td></tr><tr><td>Recall</td><td>0.90%</td></tr><tr><td>F1</td><td>90%</td></tr></tbody></table>	Metric	Value	Accuracy	0.89%	Precision	87%	Recall	0.90%	F1	90%
Metric	Value											
Accuracy	0.89%											
Precision	87%											
Recall	0.90%											
F1	90%											

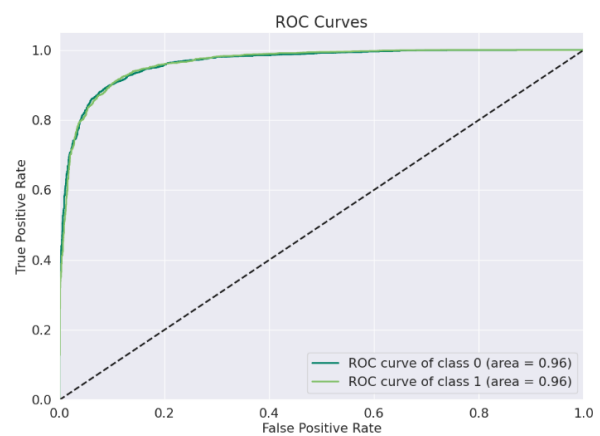
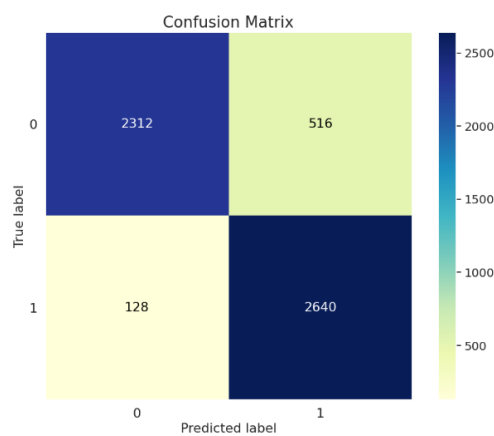
## Comparison:

### Confusion matrix indicator:

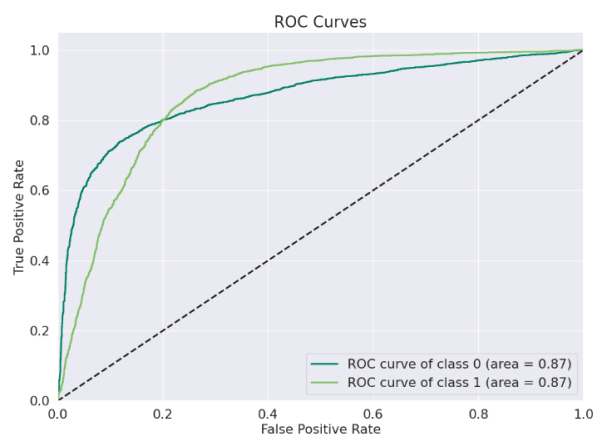
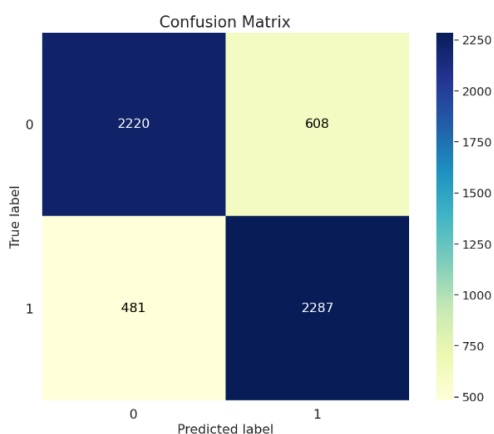
As both true positive and true negative are high and false positive and false negative this shows that the data collected contains positive and negative instances are being classified correctly by the model and that the model successfully captures the patterns and characteristics linked to each class. Which is shown in all the models below.

Naïve Bayes:

TF-IDF

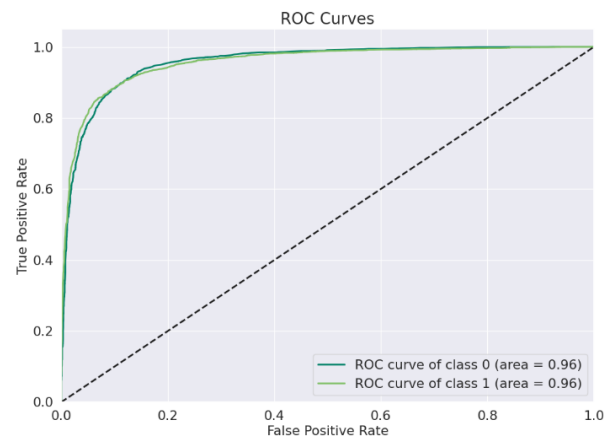
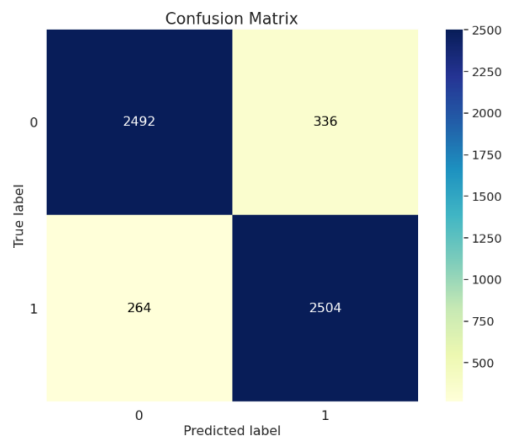


Word2Vec

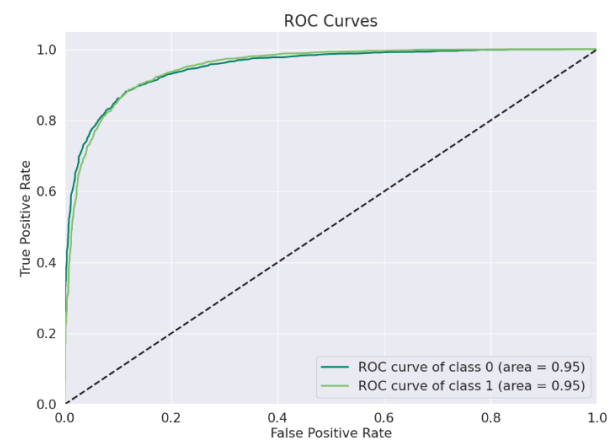
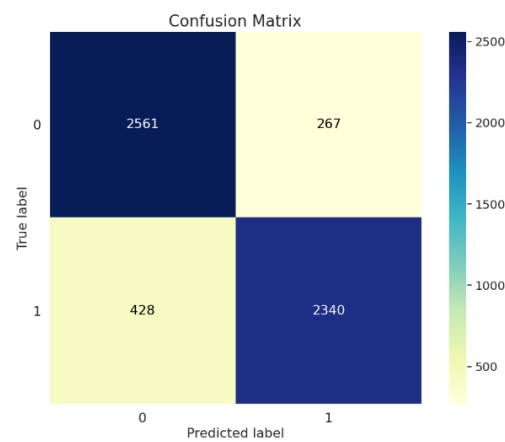


Random Forest:

TF-IDF

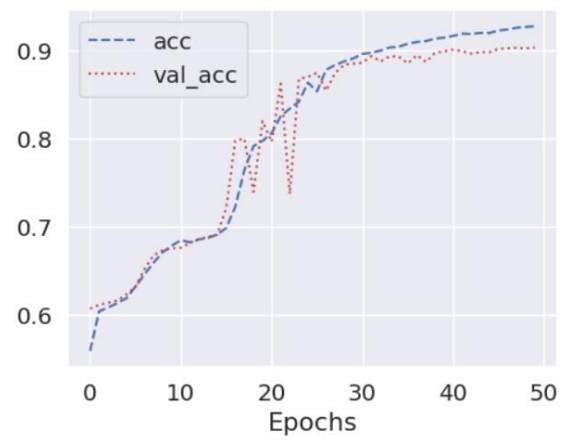
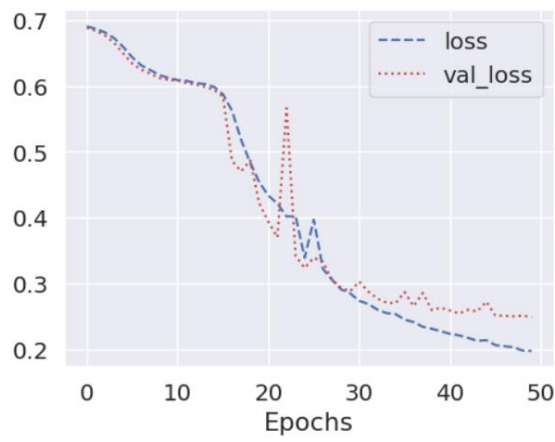


Word2Vec

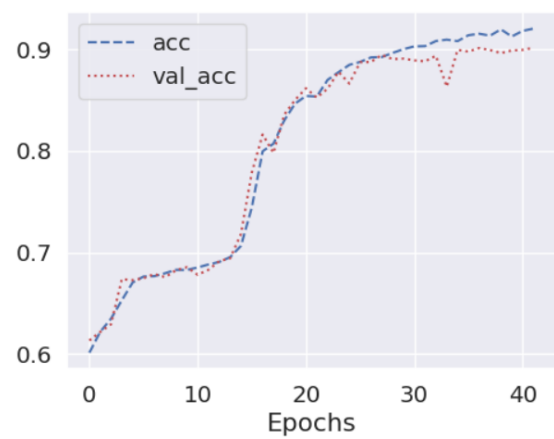
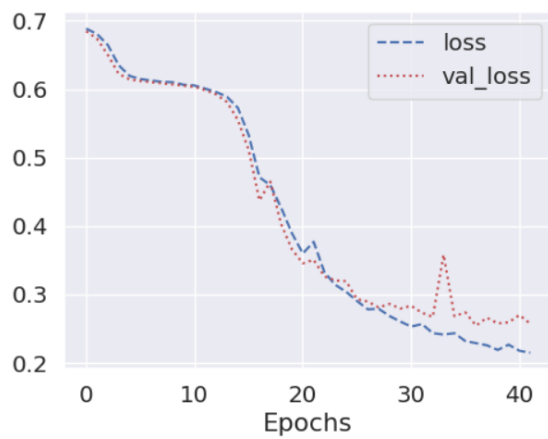


## LSTM Epochs:

TF-IDF



Word2Vec



## 5.2 Discussion

According to this study TF-IDF and Word2vec has been applied on the four models Random Forest, Naïve Bayes, LSTM, SVM. With TF-IDF the accuracies were (Naïve bayes 0.88%, Random Forest 0.89%, LSTM 0.90%, SVM 93%). Showing that SVM got good result which is 93% the highest one comparing to the other 3 models. With applying word2vec the accuracies was (Naïve bayes as it got 0.81%, Random Forest 0.88%, LSTM 0.89%, SVM 89%). The highest accuracy was with SVM as well and LSTM got same accuracy 0.89%, and Naïve Bayes accuracy decreased from 0.89% with TF-IDF to 0.81% with Word2Vec. In this study TF-IDF shows a promising results comparing to word2vec and SVM as well was giving the best accuracy upon all the models with TF-IDF and Word2Vec. This accuracies comparing to the ones in paper “Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing” , using the same dataset mental\_health, their highest accuracies was with SVM model and TF-IDF feature extraction. SVM got accuracy 0.94% with TF-IDF proving that it is the best model and with word2vec it got 91%. Naïve bayes gives accuracy 0.89% with TF-IDF and with word2vec gives 83%. The other two models was LR and XGBoost which shows low accuracies, so trying two other models was a good decision as RF and NB shows a better accuracies than the other two models in the paper.

## 6 Testing and Evaluation

### Ensemble:

The test was making ensemble to see the results from combining the four models together and test if it will increase or decrease the accuracy of the results. It gets accuracy with TF-IDF is 65% and 90% with Word2Vec.



## 7 Conclusion and Future Works

### 7.1 Conclusion

Four well-known machine learning models—Naive Bayes, Random Forest, LSTM, and SVM—were used to analyse social media posts and classify them according to signs of stress. In order to prepare the textual data for machine learning algorithms, NLP techniques were used to convert the unstructured text into structured form including tokenization, stemming, and removing stop words, to transform the unstructured text into a structured format suitable for machine learning algorithms using TF-IDF and Word2Vec as well for feature extraction. Using a deep learning model as well. Each model was trained and tested using a labelled dataset made up of reddit platform postings from people posts showing if this person is “stressed” or “Not stressed”. Results showed that SVM got the highest accuracy than other models, showing its effectiveness in identifying stress in social media posts. This implies that SVM can properly classify data using linguistic features that are derived from the text as well as find patterns. SVM thus becomes a potential model for identifying stress in posts on social media. The findings have consequences for those working in the field of mental health as well as researchers because automated stress and mental illness identification can offer important information about population-level mental health trends, early intervention, and focused mental health interventions.

It's crucial to remember that this study has some limitations. Furthermore, because the study's main focus was on stress detection, its findings might not apply to other mental health issues. Multi-class classification can be tried instead of only having two classes for the classification being either "stressed" or "Not-stressed" in order to try to further define and detect different types of mental illness. Future studies should analyse the generalizability of the results across various groups and examine the applicability of these models to more types of mental health disorders. The models' accuracy and resilience in identifying mental illness from social media posts may be improved with further development and modification, as well as the addition of new data sources and attributes. Moreover, using more deep learning models. Overall, the study shows how the analysis of social media data can be used to enhance mental health treatment through the use of NLP and machine learning.

## References

Use the *Reference* paragraph style to enter and cross-reference document references. Books **Error! Reference source not found.**, standards **Error! Reference source not found.**, reports **Error! Reference source not found.**, journal articles [1], conference papers **Error! Reference source not found.**, and web pages [2] are conventionally presented in slightly different ways.

- [1] Ottawa, R.S.U.of et al. (2021) Using social media for Mental Health Surveillance: A Review: ACM computing surveys: Vol 53, no 6, ACM Computing Surveys. Available at: <https://dl.acm.org/doi/abs/10.1145/3422824> (Accessed: November 25, 2022).
- [2] Kim, J., Lee, J., Park, E., & Han, J. (2020, July 16). *A deep learning model for detecting mental illness from user content on social media*. Nature News. Retrieved November 25, 2022, from <https://www.nature.com/articles/s41598-020-68764-y>
- [3] Mezzi, R., Yahyaoui, A., Krir, M. W., Boulila, W., & Koubaa, A. (2022, January 23). Mental health intent recognition for Arabic-speaking patients using the Mini International Neuropsychiatric Interview (Mini) and Bert Model. MDPI. Retrieved November 25, 2022, from <https://www.mdpi.com/1424-8220/22/3/846>
- [4] *Open access original research can natural ... - homepage / BMJ open*. (n.d.). Retrieved November 25, 2022, from <https://bmjopen.bmj.com/content/bmjopen/12/2/e052911.full.pdf>
- [5] Hassan, A., Ali, M. D. I., Ahammed, R., Bourouis, S., & Khan, M. M. (2021, December 13). *Development of NLP-integrated intelligent web system for E-Mental Health*. Computational and Mathematical Methods in Medicine. Retrieved November 25, 2022, from <https://www.hindawi.com/journals/cmmm/2021/1546343/>
- [6] *Midas: Mental illness detection and analysis via social media / IEEE ...* (n.d.). Retrieved November 25, 2022, from <https://ieeexplore.ieee.org/abstract/document/7752434>
- [7] Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022, April 8). *Natural language processing applied to Mental Illness Detection: A Narrative Review*. Nature News. Retrieved November 25, 2022, from <https://www.nature.com/articles/s41746-022-00589-7>
- [8] IRJMETS - International Research Journal of Modernization in Engineering Technology and Science. (n.d.). *Irjmets*. IRJMETS International Research Journal of Modernization in Engineering Technology and Science. Retrieved November 25, 2022, from

[https://www.irjmets.com/uploadedfiles/paper//issue\\_6\\_june\\_2022/27031/final/fin\\_irjmets1656403378.pdf](https://www.irjmets.com/uploadedfiles/paper//issue_6_june_2022/27031/final/fin_irjmets1656403378.pdf)

- [9] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017, January 30). *Natural language processing in mental health applications using non-clinical texts†: Natural language engineering*. Cambridge Core. Retrieved November 25, 2022, from <https://www.cambridge.org/core/journals/natural-language-engineering/article/natural-language-processing-in-mental-health-applications-using-nonclinical-texts/32645FFCFD37C67DA62CA06DB66EB2F4>
- [10] *IEEE Xplore*. (n.d.). Retrieved November 25, 2022, from: <https://ieeexplore.ieee.org/document/8457362>
- [11] *IEEE Xplore* (no date). Available at: <https://ieeexplore.ieee.org/ielam/69/8371206/8239661-aam.pdf> (Accessed: November 25, 2022).
- [12] *IEEE Xplore Full-text PDF*: (no date). Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8681445> (Accessed: November 25, 2022).
- [13] *Depression detection by analyzing social media posts of user | IEEE ...* (no date). Available at: <https://ieeexplore.ieee.org/abstract/document/9065101/> (Accessed: November 25, 2022).
- [14] <https://www.semanticscholar.org/paper/Feature-Extraction-Methods-for-Depression-Detection-Fang-Dianatobing/4ca18e765aae8f3839d4b25a02355fc9e9f82ec8>
- [15]
- [16] Vatsal. (2023, January 29). *Word2Vec explained*. Medium. <https://towardsdatascience.com/word2vec-explained-49c52b4ccb71>
- [17] Dutta, M. (2021, July 14). *Bag-of-words vs TFIDF vectorization –a hands-on tutorial*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/07/bag-of-words-vs-tfidf-vectorization-a-hands-on-tutorial/>
- [18] Ramadhan, L. (2021, February 4). *TF-IDF simplified*. Medium. <https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530>
- [19] *Time Series - LSTM model*. Online Courses and eBooks Library. (n.d.). [https://www.tutorialspoint.com/time\\_series/time\\_series\\_lstm\\_model.htm#:~:text=An%20LSTM%20module%20has%20a,only%20a%20few%20linear%20interactions.](https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm#:~:text=An%20LSTM%20module%20has%20a,only%20a%20few%20linear%20interactions.)
- [20] *Support Vector Machine (SVM) algorithm - javatpoint*. [www.javatpoint.com](http://www.javatpoint.com). (n.d.). <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

- [21] *Support Vector Machines (SVM) algorithm explained*. MonkeyLearn Blog. (2017, June 22).  
<https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/>
- [22] R, S. E. (2023, April 26). *Understand random forest algorithms with examples (updated 2023)*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
- [23] *Naive Bayes classifier in machine learning - javatpoint*. www.javatpoint.com. (n.d.-a).  
<https://www.javatpoint.com/machine-learning-naive-bayes-classifier#:~:text=Na%C3%AFve%20Bayes%20is%20one%20of,choice%20for%20text%20classification%20problems.>
- [24] Turing. (2022, March 11). *Naive Bayes algorithm in ML: Simplifying classification problems*. Naive Bayes Algorithm in ML: Simplifying Classification Problems.  
<https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners>
- [25] *Confusion matrix in machine learning - javatpoint*. www.javatpoint.com. (n.d.-a).  
<https://www.javatpoint.com/confusion-matrix-in-machine-learning>
- [26] *Machine learning (ML) for Natural Language Processing (NLP)*. Lexalytics. (2022, July 11).  
<https://www.lexalytics.com/blog/machine-learning-natural-language-processing/>
- [27] *Natural language processing (NLP) - A complete guide*. (NLP) [A Complete Guide]. (n.d.).  
<https://www.deeplearning.ai/resources/natural-language-processing/>

