****

**University of Hertfordshire**

***School of Computer Science***

***MSc Advanced Computer Science with Sandwich***

***Placement***

**7COM1039- Advanced Computer Science Masters Project**

**Final Project Report**

**Title: Social Media Network Sentimental Analysis using Machine Learning**

**[Your name]**

**[Your email address]**

**[Your student ID]**

**Project Supervisor: [XXXX]**

# Abstract

There is various type of the social media platform taken a special place in the human nature and their behavior, and all this just gets possible because of the internet that brings peoples closer throughout the world. Different people around the world share their personal opinion on different topics, events, political agendas etc. The sentimental analysis is the trend of the many research and social media play an important role to collect the information through the help of it by using the text data mining and the information retriever. By using social media platforms, many users express their personal views and express their thoughts using Facebook, Twitter, etc. still there is various type of the problem available in the performing the sentimental analysis. Already there are various types of research proposed by using the different types of approaches but still lack a lot due to lack of appropriate methods, approaches, insufficient data and many other reasons. So to overcome all these approaches, a comparative sentimental analysis is performed to carry out the more accurate sentimental analysis outcome.

A framework is proposed to perform the sentimental analysis on the social media platform through using the different types of ML algorithms, and classifiers (**Logistic regression, long short-term memory, and random forest**) are used to simulate the results. The approach is combined with the **TF-IDF “term frequency-inverse document frequency”** proposed is developed by focusing on the pre-processing phase as well as on the feature extraction phase so that the better outcome can be calculated. The tweet data sets are taken to simulate the framework in which 1048571 numbers of Twitter are stored in the row, and six columns are available in the data sets. The simulation of the proposed model results is analyzed on the basis of the F1 score, in which the results are in the form of positive and negative aspects. After simulating the proposed approach by using the different ML algorithms, it calculated the long short-term memory algorithm provides the best results among all other algorithms.

# Acknowledgment

I would first like to thank my supervisor \_\_\_\_\_\_\_\_\_\_ for the continuous support, guidance and for valuable feedback throughout the project journey. Finally, I must express my very profound gratitude to my parents and to my spouse for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

# Declaration

This report is submitted in partial fulfillment of the requirement for the degree of Master of Science in Advanced Computer Science with Sandwich Placement at the University of Hertfordshire (UH).

It is my own work except where indicated in the report.

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website, provided the source is acknowledged.

**Table of Contents**

[Abstract i](#_Toc81055534)

[Acknowledgment ii](#_Toc81055535)

[Declaration iii](#_Toc81055536)

[List of Figures vi](#_Toc81055537)

[List of Tables x](#_Toc81055538)

[Chapter 1 Introduction 1](#_Toc81055539)

[Background information 1](#_Toc81055540)

[Problem statement 2](#_Toc81055541)

[Research questions 3](#_Toc81055542)

[Research aim 3](#_Toc81055543)

[Research objectives 4](#_Toc81055544)

[Project plan 4](#_Toc81055545)

[Overview 5](#_Toc81055546)

[Chapter 2 Literature review 6](#_Toc81055547)

[SVM based model 6](#_Toc81055548)

[Naive Bayes 8](#_Toc81055549)

[Deep learning 10](#_Toc81055550)

[Logistic regression 12](#_Toc81055551)

[Random forest classifier 14](#_Toc81055552)

[Long short term memory 16](#_Toc81055553)

[Benefits of these techniques 16](#_Toc81055554)

[Comparison table 19](#_Toc81055555)

[Chapter 3 Methodology 23](#_Toc81055556)

[Data collection 25](#_Toc81055557)

[Pre-processing 25](#_Toc81055558)

[Application of machine learning models 25](#_Toc81055559)

[Performance computation 26](#_Toc81055560)

[Chapter 4 Implementation 27](#_Toc81055561)

[Data pre-processing 27](#_Toc81055562)

[Data visualization 30](#_Toc81055563)

[Splitting of data 32](#_Toc81055564)

[Application of machine learning models 32](#_Toc81055565)

[Logistic regression model 33](#_Toc81055566)

[Random forest classifier 34](#_Toc81055567)

[Long short term memory 35](#_Toc81055568)

[Chapter 4 Results 38](#_Toc81055569)

[Data pre-processing 38](#_Toc81055570)

[Data visualization 39](#_Toc81055571)

[Splitting of data 41](#_Toc81055572)

[Application of machine learning models 41](#_Toc81055573)

[Logistic regression model 41](#_Toc81055574)

[Random forest classifier 42](#_Toc81055575)

[Long short term memory 42](#_Toc81055576)

[Chapter 5 Discussion 43](#_Toc81055577)

[Chapter 6 Conclusion and future work 45](#_Toc81055578)

[Conclusion 45](#_Toc81055579)

[Future work 46](#_Toc81055580)

[References 47](#_Toc81055581)

[Appendix 51](#_Toc81055582)

# List of Figures

[Figure 1 Work breakdown structure of the project 4](#_Toc81055583)

[Figure 2 Gantt chart 5](#_Toc81055584)

[Figure 3 Depression measurement is analyzed through sentimental analysis 6](#_Toc81055585)

[Figure 4 Multilingual textual data framework 7](#_Toc81055586)

[Figure 5 Architecture of the system 9](#_Toc81055587)

[Figure 6 Results of the proposed approach 10](#_Toc81055588)

[Figure 7 Architecture of end to end opinion retrieval system 11](#_Toc81055589)

[Figure 8 Proposed framework model 12](#_Toc81055590)

[Figure 9 Flow of the proposed system 14](#_Toc81055591)

[Figure 10 Pre-processing and feature extraction framework 15](#_Toc81055592)

[Figure 11 Final architecture design 26](#_Toc81055593)

[Figure 12 Import of the Pandas library 27](#_Toc81055594)

[Figure 13 Information about the size of the dataset 27](#_Toc81055595)

[Figure 14 Collection of feature data 28](#_Toc81055596)

[Figure 15 The development of the logic that this dataset 28](#_Toc81055597)

[Figure 16 Selection of features 29](#_Toc81055598)

[Figure 17 Data cleaning 29](#_Toc81055599)

[Figure 18 Removal of all kinds of URLs 29](#_Toc81055600)

[Figure 19 Removal of all the numeric data 29](#_Toc81055601)

[Figure 20 The construction of the frame for the data 30](#_Toc81055602)

[Figure 21 The calculation of all the positive and negative words 30](#_Toc81055603)

[Figure 22 The development of the word cloud using the data 31](#_Toc81055604)

[Figure 23 The seaborn library for the presentation of data collected from the dataset 31](#_Toc81055605)

[Figure 24 The library of wordcloud which provides the designing of the image 31](#_Toc81055606)

[Figure 25 Create the second visualization 32](#_Toc81055607)

[Figure 26 The presentation of training and testing splitting 32](#_Toc81055608)

[Figure 27 The machine learning practices 33](#_Toc81055609)

[Figure 28 Prediction using the logistic regression model 33](#_Toc81055610)

[Figure 29 The evaluation practices into the experiment 33](#_Toc81055611)

[Figure 30 Collecting specific data from the testing dataset 34](#_Toc81055612)

[Figure 31 The mentioning of the tweet data as well as the connected calculation of sentimental values 34](#_Toc81055613)

[Figure 32 Initial stage of implementing the classifier 35](#_Toc81055614)

[Figure 33 The accuracy scores 35](#_Toc81055615)

[Figure 34 The process of initiating the LSTM 35](#_Toc81055616)

[Figure 35 The calculation of the padding code 36](#_Toc81055617)

[Figure 36 Maximum value evaluated 36](#_Toc81055618)

[Figure 37 The creation of the models 36](#_Toc81055619)

[Figure 38 NumPy library functions and evaluation procedures 37](#_Toc81055620)

[Figure 39 The evaluation of the data 37](#_Toc81055621)

[Figure 40 Generation of the confusion matrix 37](#_Toc81055622)

[Figure 41 The calculation of the accuracy score 37](#_Toc81055623)

[Figure 42 Data from the dataset 38](#_Toc81055624)

[Figure 43 The length of the selected dataset 38](#_Toc81055625)

[Figure 44 The extraction of all the target values 39](#_Toc81055626)

[Figure 45 Showing processed data 39](#_Toc81055627)

[Figure 46 The values associated with the sentimental connected 40](#_Toc81055628)

[Figure 47 Graphical representation is generated 40](#_Toc81055629)

[Figure 48 Graph of positive words associated with the frequency 41](#_Toc81055630)

[Figure 49 The classification practices with different models 41](#_Toc81055631)

[Figure 50 The accuracy score of the logistic regression 42](#_Toc81055632)

[Figure 51 The accuracy score of the Random Forest Classifier 42](#_Toc81055633)

[Figure 52 Confusion matrix generated 42](#_Toc81055634)

[Figure 53 Accuracy involved with the process of LSTM 42](#_Toc81055635)

[Figure 54 Import of the Pandas library 51](#_Toc81055636)

[Figure 55 Information about the size of the dataset 51](#_Toc81055637)

[Figure 56 Collection of feature data 51](#_Toc81055638)

[Figure 57 The development of the logic that this dataset 51](#_Toc81055639)

[Figure 58 Selection of features 51](#_Toc81055640)

[Figure 59 Data cleaning 52](#_Toc81055641)

[Figure 60 Removal of all kinds of URLs 52](#_Toc81055642)

[Figure 61 Removal of all the numeric data 52](#_Toc81055643)

[Figure 62 The construction of the frame for the data 52](#_Toc81055644)

[Figure 63 The calculation of all the positive and negative words 53](#_Toc81055645)

[Figure 64 The development of the word cloud using the data 53](#_Toc81055646)

[Figure 65 The seaborn library for the presentation of data collected from the dataset 53](#_Toc81055647)

[Figure 66 The library of wordcloud which provides the designing of the image 53](#_Toc81055648)

[Figure 67 Create the second visualization 54](#_Toc81055649)

[Figure 68 The presentation of training and testing splitting 54](#_Toc81055650)

[Figure 69 The machine learning practices 54](#_Toc81055651)

[Figure 70 Prediction using the logistic regression model 54](#_Toc81055652)

[Figure 71 The evaluation practices into the experiment 54](#_Toc81055653)

[Figure 72 Collecting specific data from the testing dataset 55](#_Toc81055654)

[Figure 73 The mentioning of the tweet data as well as the connected calculation of sentimental values 55](#_Toc81055655)

[Figure 74 Initial stage of implementing the classifier 55](#_Toc81055656)

[Figure 75 The accuracy scores 55](#_Toc81055657)

[Figure 76 The process of initiating the LSTM 56](#_Toc81055658)

[Figure 77 The calculation of the padding code 56](#_Toc81055659)

[Figure 78 Maximum value evaluated 56](#_Toc81055660)

[Figure 79 The creation of the models 56](#_Toc81055661)

[Figure 80 NumPy library functions and evaluation procedures 57](#_Toc81055662)

[Figure 81 The evaluation of the data 57](#_Toc81055663)

[Figure 82 Generation of the confusion matrix 57](#_Toc81055664)

[Figure 83 The calculation of the accuracy score 57](#_Toc81055665)

[Figure 84 Data from the dataset 57](#_Toc81055666)

[Figure 85 The length of the selected dataset 58](#_Toc81055667)

[Figure 86 The extraction of all the target values 58](#_Toc81055668)

[Figure 87 Showing processed data 58](#_Toc81055669)

[Figure 88 The values associated with the sentimental connected 59](#_Toc81055670)

[Figure 89 Graphical representation is generated 59](#_Toc81055671)

[Figure 90 Graph of positive words associated with the frequency 60](#_Toc81055672)

[Figure 91 The classification practices with different models 60](#_Toc81055673)

[Figure 92 The accuracy score of the logistic regression 60](#_Toc81055674)

[Figure 93 The accuracy score of the Random Forest Classifier 60](#_Toc81055675)

[Figure 94 Confusion matrix generated 61](#_Toc81055676)

[Figure 95 Accuracy involved with the process of LSTM 61](#_Toc81055677)

# List of Tables

[Table 1 Comparison table 19](#_Toc81055678)

# Chapter 1 Introduction

## Background information

In the past few decades, the communication technology developed, a lot of credit goes to the internet, which brings the peoples together closer. The internet help to overcome the slow communication which uses in the past through the help of letter, telegram which is a too slow process of the communication all the problems are mitigated by introducing the social media platforms in the human life cycle for the communication. Social media is now emerging in every human life sans have a diverse impact on the daily life of a person. **The sentimental analysis is a technique to detect the opinion of an individual person or the group in the text form**. The sentimental analysis is the trend of the many research and social media play an important role to collect the information through the help of it by using the text data mining and the information retriever (Mutinda, Mwangi and Okeyo, 2021). By using social media platforms, many users express their personal views and express their thoughts using Facebook, Twitter, etc. and make the relationship by communicating with others by discussing various topics like politics, social issues and their personal problems. Sentimental analysis is a process of finding the negative and the positive viewpoints in written form. Many organization use this technique to determine the sentimental statements from the social data, measure their brand values and understand the customer requirements, though the customer experiences their views and the felling for the organization by using the social media platform more confidently and freely which play an essential tool for the organization to monitor and understand the sentimental statements. Sentimental analysis is the most common text classification approach, which analyzes the incoming messages and advises the underlying views (Sciandra, 2020).

According to the world health organization, the depression is the world's fourth-largest diseases which the millions of the peoples are suffering the word entirely due to depressions millions of peoples end there life there is very low medical support available for the treatment for the depression due to difficulty in diagnosing mental problems. Depression is a mental sentiment which is also called clinical depression, which is instability in the mental health of a person which characterized the mood, loss in interest, negotiations which make them feel guilt which affects the feelings, thinking and handling the daily schedules. The social network sites are the community in which various types of people make a network of the different people's different organizations and express their feelings in the form of the text through which the advanced research is done by focusing on the data of the confirmed social media observation which increase the confidence of the collective data values to analyze the and monitoring the data to mitigate the problems related to the mental health, and they help to provide the accurate treatment to the patient. There is various type of the sentimental analysis available which are mention and explained blow (Mussiraliyeva, Bolatbek, Omarov and Bagitova, 2020).

* **Emotion detection:** In this type of sentimental analysis, the main focus is to detect the emotions, for example, happiness, anger, aggravation, annoyance, unhappiness and many more.
* **Aspect-based sentimental analysis:** In this sentiment analysis, the statement is analyzed, which is the text-based approach or in the written form. For example, an organization performs a feedback form to their customer to write their view about the organization's performance. The customer gives their review about which can be in any way like positive, negative or neutral. All this type of reviews is aspect-based sentimental analysis, which helps the organization and the business organization to enhance the customer experience, which makes them the first priority (Coban and Ozel, 2018).
* **Multilingual sentimental analysis:** Multilingual sentiment analysis is hard and can be difficult to retrieve. In this approach, there is various type of the pre-processing and resources are needed in which there are some resources easily available on the internet social media, but some need to be created. There are many benefits available that are more important because they help in detecting to mitigate the problem related to the health issues like depression, help organizations to understand the whole feedback of their customers. It is estimated that more than 90% of the data are unbalanced around the world, which creates in every second through social media like Facebook, twitters, emails, which make it hard to understand and analyze the sentiments in a well-mannered and timely (Fernandez and Alani, 2021).

## Problem statement

The main problem that has been noticed on social media is of detection of social media sentiments. Various people upload the status on their social media handles, illustrating whether they are happy, sad, crying, or some other emotion. Even some of them do in such a sarcastic way that one could not determine their actual feelings. Thus, this area of the field has gained much more importance. The analysis of sentiments is done for determining the feelings of a person so that appropriate actions could be taken (Kathuria et al., 2021).

## Research questions

Research in the context of sentimental analysis, machine learning has gained much more importance. For accomplishing the task, machine learning techniques are used. The below-listed research questions will help us in identifying the importance of research and highlight the main methodology used for completing the research:

* Why is the dataset of Twitter taken for this research? And what attributes are extracted from the selected dataset?
* Which models of machine learning are used in classifying social media sentiments into different categories?
* Explain the data processing and its visualization in the context of machine learning.
* How can you say that the accuracy of the LSTM method is greater than logistic regression and random forest classifiers (Sakketou and Ampazis, 2020)?

## Research aim

The aim of this research is to propose the machine learning techniques approach is used for the sentimental analysis is done on social media and social networking sites. On social media sites, the data are freely available because they spread their data conveniently on the web pages. These types of data enhance the enrolment of the new researchers to pitch them in the field of sentimental analysis. Machine learning techniques allow the system to learn more new tasks without being specifically programmed to perform them. The machine learning approach train the sentimental analysis to read the written sentiments data to analyze sentiments which are available on the social media to read ahead of definitions to understand the things, content, expressions, views, and undamaged words. Many people express their feelings and views on the internet on social media platforms in discussion blogs. The large multinational institute hires the researchers to analyze these data to enhance their productivity and services. The main focus of the organization is to analyze and determine the sentiments from the feedback forms which are written by the customers. The machine learning techniques help them to analyze the data in a more enhanced way by using the different approaches and algorithms into its system of machine learning, which increases the accuracy of the performed task (Abd Elaziz et al., 2021).

## Research objectives

The main objectives of the research are given below:

* To propose a machine learning model that can help in determining the type of sentiments by simply analyzing the social media posts.
* To classify the sentiments into different categories so that the emotions of individuals can be determined.
* To understand the feedback of customers after purchasing the product. This approach is used by big businesses such as supermarkets and more.
* To design a model that achieves higher accuracy.

## Project plan

Figure 1 Work breakdown structure of the project

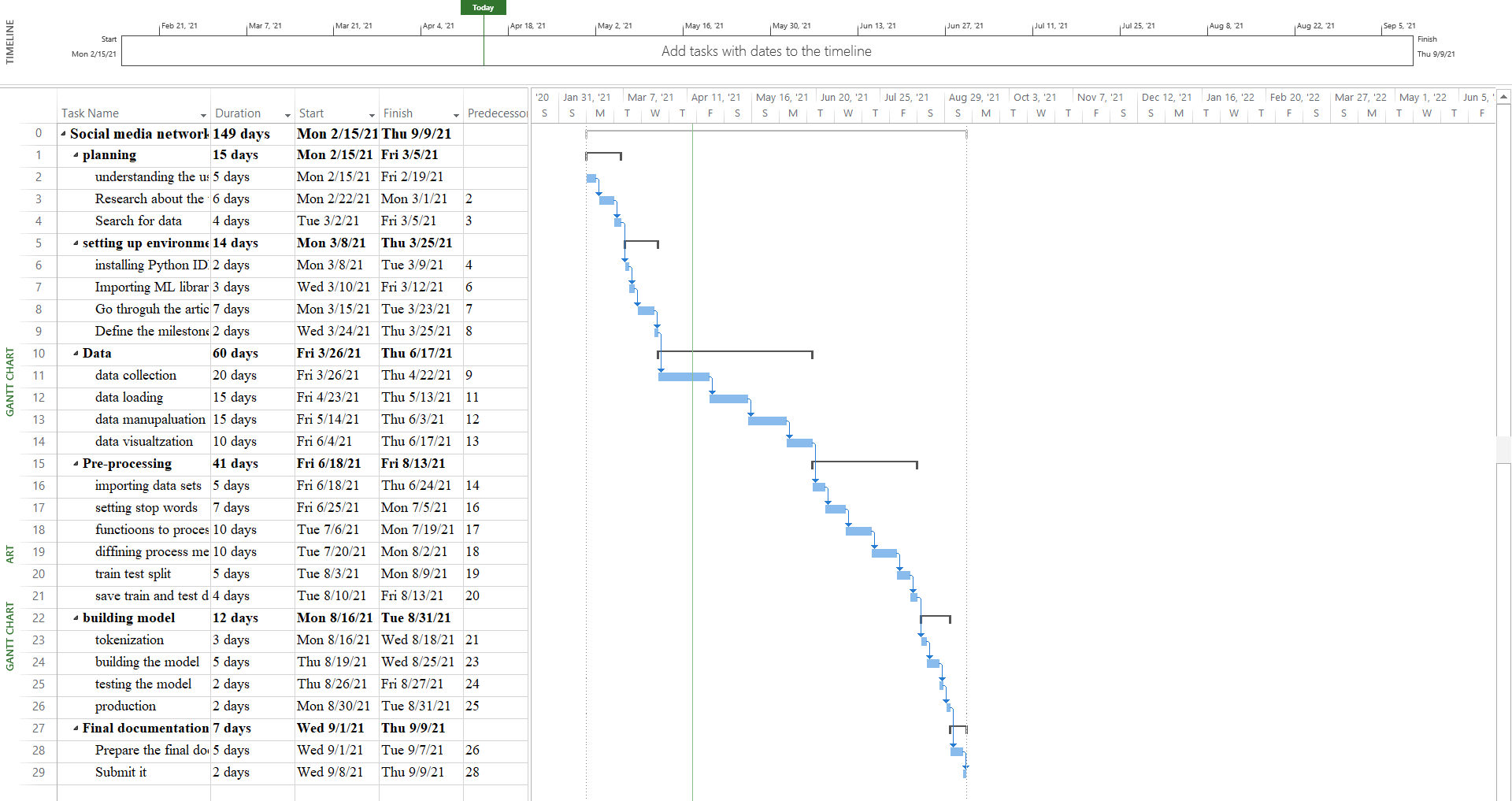


Figure 2 Gantt chart

## Overview

**Chapter 1:** In the very first chapter of the thesis report, we have discussed the background information, problem statement, research aim, research objectives, research questions, and project plan that will be followed while completing the research.

**Chapter 2:** The next chapter is about the literature review in which different research articles are selected on the basis of a given topic, and reviewed highlighting the main problem, proposed method, and results. Moreover, at last, a comparison table is also made for a clear understanding.

**Chapter 3:** The next chapter includes the models, classifiers, and algorithms that are used in the research.

**Chapter 4:** The model of machine learning that we have implemented is shown here. This chapter represents the coding section that is implemented. It highlights the data pre-processing, data visualization, and cleaning.

**Chapter 5:** This chapter represents the results that are obtained from the above implementation along with the achieved accuracy.

**Chapter 6:** The discussion contains the comparison of different models that are used in the research.

**Chapter 7:** Conclusion and future work is the last chapter representing the summary of the overall research and how the research can be carried further in the future.

# Chapter 2 Literature review

## SVM based model

In this section of the literature review, the sentimental analysis is done on social networking sites which is an online society in which people communicate across the world and express their views on different types of topics and share their personal feelings, ideas their opinion with each other. The sentimental analysis is done on social network sites to measure the depression index. In which the multi-kernel SVM-based model is proposed to recognize the sentiments like the depression of the people and the divide into the three categories of type like users microblog transcript, user profile and user actions are collected to describe the user's mental situation. In which the author proposed a data mining application based on the categorization techniques; therefore decision tree is used to forecast the upcoming candidate for the depression measurement (Gan, Su and Li, 2020). Also, in the SVM model, the hybrid machine learning algorithm technique is used to choose the feature from the training data locations and transfer all the features to the artificial neural network to do the sentimental documentation categorization. The SVM model, the hybrid wrapper and administered learning category approaches are implemented to detect the negligible features subsets for the text sentiments categorizations. The main object of the text sentiment categorization is to predicting and analyzing the opinions sentiments linked to the division sentiments. As per the approach, the user uses different types of machine learning categories for the sentiment analysis categories because a single approach is not beneficial to analyze the different types of data sets. As per the proposed method for the sentimental analysis, the results show that excellent results as compared to the other approaches like nave bytes and highest entropy categorize in which the accuracy results of the SVM are 91% and 83% Naïve Bayes and 80% in highest entropy (Hassan et al., 2017).

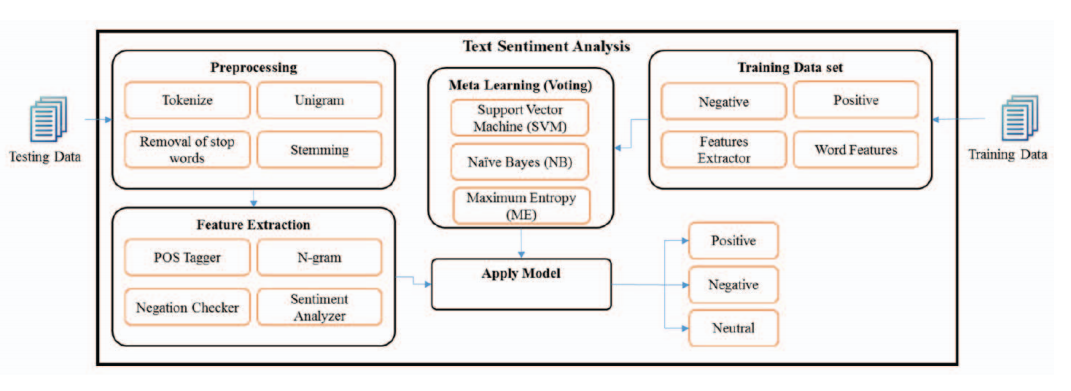
****

Figure 3 Depression measurement is analyzed through sentimental analysis

There are many reasons available on social media on which the public shares their expression from all the topics. The uncertainty in the political, cultural and various social issues are diversely spread between the peoples across worldwide, which represent by their sentiments on social networking. The major language which uses the English language, which is one of the most common languages which people use to express their expressions on the social media future more other native language is also used by the public as per their national language. Thus, it is also necessary to include the sights in such verbal communications besides through broadly second-hand languages for informative improved insight as of the data (Potamias, Siolas and Stafylopatis, 2019). The main focus is to analyze the social media sentiments is done on multilingual written data to find the concentration of the sentiments radicalism. The research is based on the SVM model in which the research categorizes the unmannered textual views in any four classifications on the basis of the radicalism text level, which are neutral, restrained, extreme low, extreme high. In the initial phase, the multilingual lexicon is created with the concentration weights. In which domain experts validate the lexicon in which the accuracy is 88% per is validated. Moreover, the algorithm of linear support victor and multinomial naive Bayes is implemented for the sentimental classification. The accuracy of the linear support vector is 82% is accumulated on the fundamental multilingual datasets, which is very good results are carried out for the sentimental analysis on the social media textual data (Asif et al., 2020).

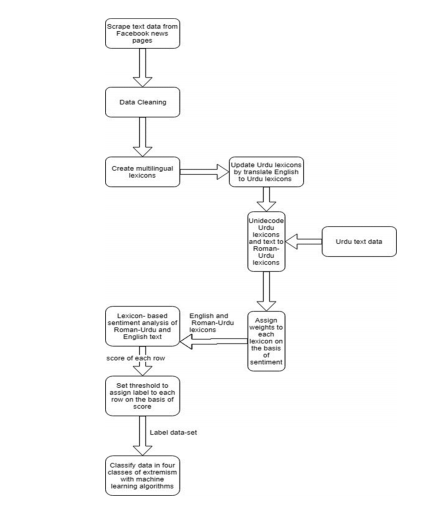


Figure 4 Multilingual textual data framework

## Naive Bayes

In recent years opinion mining is getting hiked to know the opinion to classify the review in a tow point scale in the context of positive and negative. Getting review and doing the sentimental analysis is getting hiked into the tourism sector. Due to the Arabic springs in Tunisia tourism sector is forbidden. So that there are huge comments are published by the tourist, and they give their huge opinion which causes an increase in the comment data. Due to the increase in the volume of the comment, the different social media is suffering from the huge data such as Facebook, Twitter. Due to all this consequence related to the tourism getting knowing the opinion is getting important and play an important role to understand the situation and make it stable. In this case scenario, the author focuses on the tourist point of view by using Twitter data after seeing the revolutionary effect on Tunisia tourism.

So to execute the sentimental analysis, the author develops a sentimental lexicon that is based on the emoticons as noisy labels and sentiWordnet to develop the lexical scales to categorize to investigate the review of the tourist in the more optimistically way to present the effective learning. So to perform the sentimental analysis, the researcher uses the different ML classifiers such as Naïve Bayes, SVM and entropy classification to achieve the high accuracy results by using different algorithms. The research is conduct by taking the user review based on the user's nationality and the duration of the review, which the user posted after the incident. The basic aim of the research is to conduct and identified the various points of view about the tourist in the context of safety. The research is encouraged by tracking the changes in people's sentiments on a specific topic and investigate the trend of the topic that increases the interest to analyze the present rather than analyzing the past for the betterment of the present. The performed research shows the effective outcome results (Chaabani, Toujani and Akaichi, 2017).

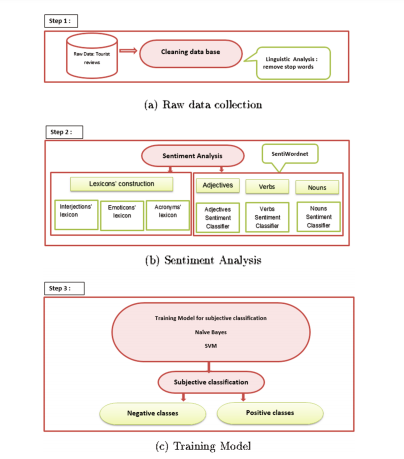


Figure 5 Architecture of the system

In the past, some years, as digital social media platforms taking a special place in human life, people starting their expression, emotions and opinion more rather than socially interacting physically. Social media has taken a special attraction to get investigated in the context of the sentimental analysis that increases the rapid growth to explore the opinion mining which is available on the different social media platforms thorough musing different approaches of machine learning that help in to provide the polarity measurement. Rather than using machine language techniques and tools for sentimental analysis during the election, there is a huge need required to use the state-of-the-art approach. The researcher has done the research and contributes its service to deal with the challenges the author focus on to use the hybrid approaches in the context to the sentimental analysis along with the machine learning techniques to identify the best approach with the results of the increased outcome to get the better sentimental analysis learning during the time of the elections. So as per the researcher, the model is proposed, which is based on the lexicon-based sentiment analysis (sentiments alignment is of words, sentence and phrases are done are measured in the form of the documents). On the basis of the dictionary, the polarity of the lexicon method is measured. The approach is used by undertaking the naïve Bayes and SVM to measure the lexicon-based model that is measured on the basis of the dictionary. The work is done by focusing on providing the comparison between the sentiment lexicons "TextBlobs, W-WSD, and SentiWordNet" so that the best option can be used for the sentimental analysis. Ads per the article, the two different ML algorithms and three different lexicons are used together on behalf of this the results are measured. The results show that TextBlob shows better outcomes (Hasan, Moin, Karim and Shamshirband, 2018).

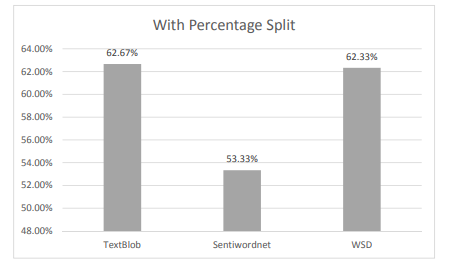


Figure 6 Results of the proposed approach

## Deep learning

Social media is one of the biggest and popular weapons, which is become now getting popular for companies to interact with consumers. As the consumers and company demands are getting rapidly switching, social media-based decision support system are rapidly develop for the arena of marketing, but there is always limitation is available regarding innovation-centric studies and product development. This research is done by proposing by utilizes an opinion retrieval theme with the combination of the sentimental analysis to provide the pillar to the decision-making process for produced analysis and creation. For the achievement of the goal by the author, it set up an **end-to-end social media-based opinion retrieval system through ML and natural language processor techniques**. In this experiment, the Google glasses are used to aim the commercial targets rather than the effective technological offering. In the approach, a framework is designed to train the task related to the opinion task and sentimental assumptions by using the **multi-task deep neural network**. The sentiment labels are split into two categories in that the tweets are divided based on the suggestion and opinion. The approach works based on the negative and positive aspects. If the data is in the negative form, then it is considered a product-related issue, and the tweet is positive, then it is used as the innovation of the new product or development of the new product. All the data and the result are projected in the form of the visualization in which the group of the keywords is eliminated from each sentiment label category. As per the analysis and according to the author, this study not just contributes to its works, but this study also offers a practical contribution to the future innovation of smart glasses (Gozuacik, Sakar and Ozcan, 2021).

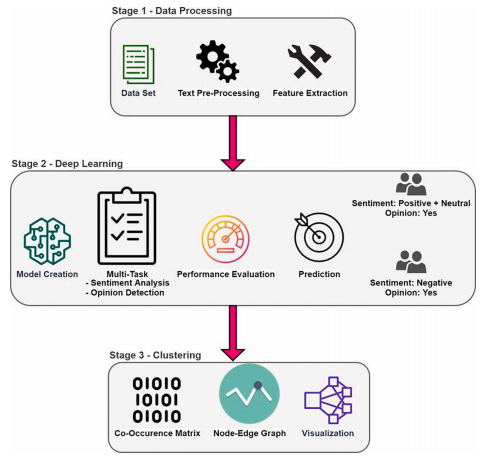


Figure 7 Architecture of end to end opinion retrieval system

The social media platform is one of the easiest ways in which the people around the world interact with each other around the world in the form of voice, text, voice notes, videos, images and in speech form, but textual and speeches are the most common way of the interaction which are widely used by the different users on the social media platform such as WhatsApp, Instagram, Facebook and twitters. By using the social media platforms by the different business companies, news channels, marketing sales etc. Using all the popular social media platforms for product quality management and for investigation purposes. The data is received from the different types of sources and mindsets that cause a large amount of the data in the form of textual form. Just because of receiving the huge textual data, categorization of the text becomes a crucial task. To solve this issue in this current DNN and NLP are widely adopted. As per this solution, the implementation of the CNN and Word2Vec approaches is used for better text classification. As per the analysis of the article, the proposed approach had done a great job by filtering the data and develop the word vector from the pre-trained model and uses the CNN in an effective manner to extract the features for the short textual classification (Sharma, Chaurasia and Srivastava, 2020).

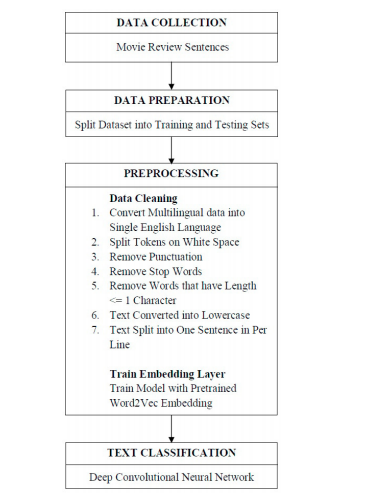


Figure 8 Proposed framework model

## Logistic regression

Sentimental analysis is now becoming a center of research for any organization due to its various type of fundamental benefits through which many organizational benefits are taken. The sentimental analysis provides various types of offers and abilities in real-time in the context of the feelings related to any product or event. Sentimental analysis is a huge process that is combined with a different type of procedure process, but one of the most common phrases of sentimental analysis is pre-processing of the texting that is performed on the Twitter data. The research that is already performed on the sentimental analysis majorly concentrated on the new sentimental feature of extraction rather than focusing on the pre-processing text. In this research, the research has done the discussion on the impact of the text pre-processing approaches on the sentimental classification presentation in 2 different sorts of the classification task, and total up the productivity of the 6 pre-processing approaches through adopting the 2 feature extraction and five datasets of twitters. While performing the experiment, it is reported that the productivity of the F1 measure improves effectively in sentimental analysis classification while using the pre-processing methodology through increasing the abbreviations and eliminating the denial, but it all hardly replace when eliminating the URLS and eliminating the various stop points. While performing the simulation, it is reported that RF and NB are more conscious than LR and SVM while Appling the various type of the pro processing method (Jianqiang and Xiaolin, 2017).

In the current era of digitalization, the amount of data is increasing rapidly. One of the biggest aspects is that data can be in any form, such as in textual, image, pictures, audio, and videos. One of the biggest factors behind the generation of the huge amount of data is social media platforms on which every end-user represents their own ideas, thoughts, and reviews, feelings for a particular product and for an event. So the author proposes an experiment in which the author collects the data from the API twitter in which the data is of every individual person of the Jakarta governor election are taken. After that, the collected data is processed for the future text pre-processing phase. After completing the pore processing process, the twitters are listed by feature extracted, and then the list is converted into the feature vector in the form of binary and then change into the Tf-idf approach. The whole data sets are carried into two different categories that are testing and testing data sets in which the training data sets are labeled manually, and for the examination of the algorithm, K-Fold cross-validation is used for checking the outcome. After simulating the experiment, the results are noted, and it is reported that the average result is 74% with the combination of the data sets in 90:10 (Ramadhan, Astri Novianty and Casi Setianingsih, 2017).

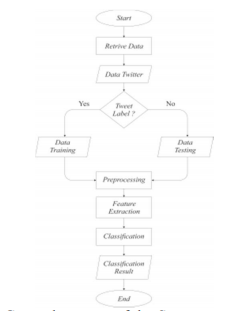


Figure 9 Flow of the proposed system

## Random forest classifier

Sentimental analysis is a computational analysis of a user's reviews, personal thoughts, behavior and emotion for a particular product or a thing. Sentimental analysis categorizes the text according to the text division in the form of positive, neutral or negative. In the current environment of the research, then major consultation is done on the lexical and ML-based approaches to perform the experiment on the social media posts. Performing the social media sentimental analysis has various types of dimension problems. Therefore the high dimension of the nature of the data needs a certain pre-processing and feature extraction that is used for increasing the accuracy of the classification. This research is conducted by the author to show the comprehensive overview of the sentimental analysis tools and techniques based on the already done researchers and investigate the ML algorithms and the classifiers, and feature extraction techniques such as HASS tags, POS in the context to the sentimental analysis over the social media user’s opinion data sets. The framework is proposed in which the data sets are taken from Twitter, and the inspection and pre-processing are done, which gain the crossing truths about the efficiency and shortage of the sentimental approaches. After proposing the approach, the examination is done in which it is reported that POS is the best appropriate for the feature extraction because it is capable of finding the dependency in between the label classes and feature values. On the other hand, the HASS tagging is the best suitable feature extraction that provides the best outcome result with the random forest algorithms (Singh, Tomar and Sangaiah, 2018).

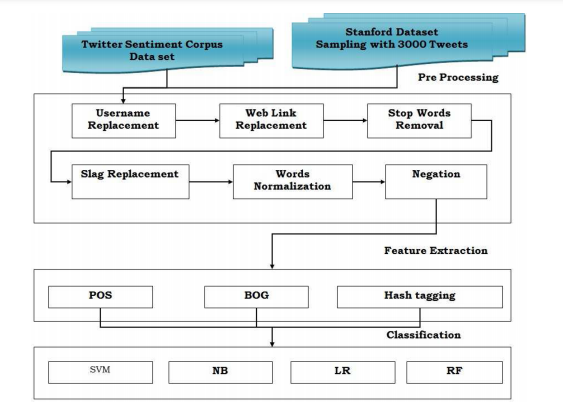


Figure 10 Pre-processing and feature extraction framework

In the existing time, the sentimental analysis is not just limited to getting the opinion and review of the user through using the social media data for the organizational, business and for many other purposes. Now In this current generation of the financial environment, cryptocurrencies are beaming increasing and become an important part of the financial market due to its special capability of low data barriers of entrance and high data availability of the cryptocurrencies, which make it a perfect domain of the research. Cryptocurrencies play an important role which attracts special attention to analyze the statistic of the financial market through the approach of the sentimental analysis and machine learning approaches to challenge the task of stock market assumption/prediction. As per the researcher, all the researches which are done in the past are majorly done on bitcoins. So by focusing on the previous researches, the author proposed a model that is based on the common ML approaches of tools and techniques on the available social media information to predict the price movement of the various type of cryptocurrencies that are Ethereum, Bitcoin, Litecoin and ripper to analyze the statistic of the market movement. So after proposing the approach, the result is compared for the operation of SVM and neural network and RF through using the element as the input feature from the data of the market and twitters. By performing the experiment, the research shows that sentimental analysis and ML is the best possible way to predict the cryptocurrencies market. In contrast, the data which is taken from Twitter is one of the best ways that can be used for prediction purposes (Valencia, Gómez-Espinosa and Valdés-Aguirre, 2019).

## Long short term memory

As long as the internet is reaching its reach around the world parallel the social media platform such as Facebook, Instagram is also reaching its reach in every economic class that creates the multilingual sentimental analysis that is becoming very important to the analysis. (Konate and Du, 2018) conduct the research initiative by using the code mixed Bambra-French Facebook group. So the author creates a framework based on four Long short term memory methodologies along with the 2 CNN and also uses the 6 other models for the experiment purpose on the gathered datasets. The Bambara text, which is used for the sentimental analysis, is totally unique, so to eliminate the problem related to the text written, the dictionary is used in which the different types of characters and word index are mentioned to create the text characters and the word placement in the place of the pre-training words vectors. In the research, the effect of review span in the approach and efficiency in between them are compared. So after simulating the approach. It is reported that the one-layer CNN model plays an effective outcome by achieving an accuracy of 83.23% (Liu, 2018).

## Benefits of these techniques

* **Categorization data at scale:** Manually categorization and processing the huge data of the big organization is not possible. The sentimental analysis helps the organization to process the huge data in an effective and efficient manner with less burden of expenses.
* **Real-time analysis:** Critical issues can be analyzed in real-time. Sentiment analysis helps instantly find these types of conditions, so through which the organization can take action immediately, for example, an organization which provides the delivery services of food and customer is satisfied with service a feel frustrated, so the real-time sentimental analysis help to mitigate the problem of the customer in real-time and enhance the customer experience.
* **Reliable criteria:** It is measured that 60-65% of the public agree approximately the time when agreeing on the sentiment of a fixed text. The sentiments which are highly prejudiced as well as inspired with the overall knowledge believes along with the ideas are tagging text. The centralized sentiments analysis system helps the companies to apply the same criteria an all type of there data which help them to gain enhanced insights and improve the accuracy (De Bruyn, 2020).

## Comparison table

Table 1 Comparison table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S. N. | Citation | Problem | Technique | Technology | Algorithm | Results |
| 1. | (Hassan et al., 2017) | Due to huge growth in social media platforms like Facebook, Instagram etc., there is tremendous involvement found in the population and use these platforms for different purposes, but there is one such service they need to be diagnosed with is related to depression. | Natural language processing | Machine learning | Support vector machine | The accuracy results of the SVM are 91% and 83% I naive bias and 80% in the highest entropy. |
| 2. | (Asif et al., 2020) | The main focus is to analyze the social media sentiments is done on multilingual written data to find the concentration of the sentiments radicalism. | SVM model | Machine learning technology | Linear support victor and multinomial Naive Bayes. | The accuracy of the linear support vector is 82%. |
| 3. | (Chaabani, Toujani and Akaichi, 2017) | Controlling situations over the different languages engaged into social media to control the sentimental issues or rivals. | NP-Complete | Text mining | Naive Bayes, Maximum entropy and SVM. | Naïve Bayes at 80% accuracy |
| 4. | (Hasan, Moin, Karim and Shamshirband, 2018) | The rapid growth to explore the opinion mining which is available on the different social media platforms. | Machine-learning techniques | E lexicon-based approach | K nearest neighbors (KNN) and Naïve Bayes. | W-WSD has an accuracy of about 79%. |
| 5. | (Gozuacik, Sakar and Ozcan, 2021) | The consumers and company demands are getting rapidly switching, social media-based decision support system are rapidly develop for the arena of marketing, but there is always limitation is available regarding innovation-centric studies and product development. | Natural language processing | Deep learning | Multi-task deep neural network. | The highest accuracy is gathered at 85%. |
| 6. | (Sharma, Chaurasia and Srivastava, 2020) | Using all the popular social media platforms for product quality management and for investigation purposes. | Machine-learning | Data mining approach | Naive Bayes and Levenshtein algorithms. | The proposed model is providing 99.07% accuracy. |
| 7. | (Jianqiang and Xiaolin, 2016) | The impact of the text pre-processing approaches on the sentimental classification presentation in 2 different sorts of the classification task, and total up the productivity of the 6 pre-processing approaches through adopting the 2 feature extraction and five datasets of twitters. | Pre-processing method. | Machine learning-based approaches. | Logistic regression | It is reported that the productivity of the F1 measure improves effectively in sentimental analysis classification while using the pre-processing methodology through increasing the abbreviations and eliminating the denial, but it all hardly replace when eliminating the URLS and eliminating the various stop points. |
| 8. | (Ramadhan, Astri Novianty and Casi Setianingsih, 2017) | Collects the data from the API twitter in which the data is of every individual person of the Jakarta governor election are taken. | K-Fold Cross Validation is used to test algorithm performance. | Machine learning | Naive Bayes, Maximum Entropy, Support Vector Machine. | The average result is 74% with the combination of the data sets in 90:10 |
| 9. | (Singh, Tomar and Sangaiah, 2018) | The comprehensive overview of the sentimental analysis tools and techniques based on the already done researchers and investigate the ML algorithms and the classifiers, and feature extraction techniques such as HASS tags, POS in the context to the sentimental analysis over the social media user’s opinion data sets. | A social analytic method. | Artificial intelligence | SVM, Navies Bayes, Linear Regression and Random Forest. | The accuracy score of the applied model is around 80%. |
| 10. | (Valencia, Gómez-Espinosa and Valdés-Aguirre, 2019) | Cryptocurrencies play an important role which attracts special attention to analyze the statistic of the financial market through the approach of the sentimental analysis and machine learning approaches to challenge the task of stock market assumption/prediction. | Artificial intelligence and machine learning techniques. | Neural networks (NNs), support vector machines (SVMs) and random forests (RFs). | Random Forest Classifier | The best model was an MLP obtained scores of 0.72 accuracies and 0.74 precision. |
| 11. | (Konate and Du, 2018) | The sentimental analysis is totally unique, so to eliminate the problem related to the text written, the dictionary is used in which the different types of characters and word index are mentioned to create the text characters and the word placement in the place of the pre-training words vectors. | Deep learning techniques | Convolutional Neural Network (CNN) | Naïve Bayes and SVM | CNN model plays an effective outcome by achieving an accuracy of 83.23%. |

# Chapter 3 Methodology

Sentimental analysis is referred to the identification and classification of the sentiment of the various type of people around the world that are expressed in the form of text sources. This report all the presenting the extreme sentiments of the public for a particular event. Product or on personal feelings through using the post, comments, reviews written on Twitter. Twitter is the type of social media platform which creates a large amount of data in the context of the sentiments for the analysis purpose. Social media platform (Twitter). The data is taken for Twitter that is collected to help in understanding the opinion of the users on the various type of topics. So to perform the work, an automated sentimental model is proposed in the flow to the research perception. There is various type of the problem available in the sentimental analysis due to the availability of the various type of the un-useful data and character is available with the useful data which create various type of the noise and create difficulty to implement the data into the model (Singh and Kumari, 2016).

In this research, part of the different techniques and algorithms are used for sentimental analysis this research which is mentioned below (Sharif et al., 2019).

* **Naive Bayes:** In simple words, the naive Bayes is a simple combination of probabilistic algorithms use for the sentimental analysis in which it was assigning a probability to a given word, and a phrase is considered as positive or negative views.
* **Support vector machines (SVM):** It is one of the best models of the machine learning techniques in which it uses the algorithms which train and define the text within or use the sentimental polarity model as an advanced step on the x-axis and y-axis prediction. And then assign a hyperplane that divides the tags into two dimensions in which one side of the lane is red, and the other side is in the blue (Hajiali, 2020).
* **Deep learning:** Deep learning is one of the approaches which is also used to analyze sentimental analysis. Deep learning is a subordinate of machine learning which focuses on calculating the data, the same as the human brain through using the artificial neural network. Machine learning is the hierarchy of machine learning. In simple words, deep learning is multi-level and permits a system to automatically create a chain of human-developed processes. Deep learning allows the various type of algorithms in a system to be used optimistically to analyze the sentimental analysis. Deep learning uses a large number of data to analyze them. So the social media is one of the biggest platforms through which a large number of quantitative data can be gathered to analyze the sentiments in a more effective manner in a short period of time which is not manual possible.
* **Logistic regression:** Sentimental analysis is well known by knowing the opinion of a person through using mining and emotional AI that denoted to the use of the natural language processing language, computational linguistics etc., and so logistic regression is one of the most common approaches which is also used in this. Logistic regression is a predictive regression that holds the dependent variables in binary form and uses the data to describe and elaborate the linking between the individual dependent variable and single or individual nominal, interval, rational and interdependent variable (Kong et al., 2020).
* **Random forest classifier:** In terms of random forest practices involved with the concepts of machine learning, there are different kinds of understanding as well as conceptual perspectives needed to be gathered in the first place. This will help the overall understanding of the work of analysis to the highest convenience of understanding and comprehensive results (Hao and Dai, 2016).
* **Long short term memory:** This is also a great part of the understanding initiative involved with the approach is combined with the TF-IDF “term frequency-inverse document frequency” proposed is developed by focusing on the pre-processing phase as well as on the feature extraction phase so that the better outcome can be calculated. The tweet data sets are taken to simulate the framework in which 1048571 numbers of Twitter are stored in the row, and six columns are available in the data sets. The simulation of the proposed model results is analyzed on the basis of the F1 score, in which the results are in the form of positive and negative aspects.

In this work, the aim is settled to achieve the experiment on the sentiments of the tweets by the collected data. The framework is developed based on the ML approaches and proposed the model to perform the analyses. In the proposed model, there are three different types of ML algorithms and classifiers (**Logistic regression, long short-term memory, and random forest**) are used to simulate the results. The approach is combined with the **TF-IDF “term frequency-inverse document frequency”**. All the outcome and the accuracy is calculated on the bases of the **F1 score and accuracy**. The approach that is proposed is developed by focusing on the pre-processing phase as well as on the feature extraction phase so that a better outcome can be calculated to analyze the sentimental analysis on social media (Jianqiang, 2015).

## Data collection

The data set is take basically based on the Twitter users in which the different type of users are giving their opinion based on their point of view for a particular event, expression, feeling and many more. The data set has 1048571 numbers of twitters that are stored in the row, and six columns are available in the data sets. The collection has different kinds of details involved with the tweets, such as the date of the tweet, the serial number of the tweets, sentimental value involved with the tweet in the category of only two sentiments, such as positive or negative, the username of the users associated with the exact tweet they posted on the social media platform (Symeonidis, Effrosynidis and Arampatzis, 2018).

## Pre-processing

This is a crucial process to be performed with the data collected from the dataset. It includes different processes involved in the single processes for making the data completely feasible for the machine learning models to be implemented upon them. The procedure starts from the visualization of the data to get the overall understanding of the current data, which involved the target as well as the features dependent upon it. This is the integral practice for embracing the quality of the selected dataset to the highest level. It also removes the data available in the dataset, which is not relevant to the analytical practice. This will increase the relevance as well as productivity of the results to be gathered from the completion of all the processes of experimental practices (Jianqiang and Xueliang, 2015).

## Application of machine learning models

In the application of the models available for the analysis and the gathering the needful results at the end of the process. For the understanding of machine learning as well as there is a need to use different kinds of models. All the models available in the machine learning practices contains the specific capability. While focusing on the capabilities as well as application of the logistic regression, this could be easily understood that there is a different kind of practices involved with the addition of the prediction practices. But the major need for results in this experimentation is related to the long short-term memory, which is capable of providing the highest percentage of accuracy as well as performance in the limited time of the process (Jianqiang, 2016).

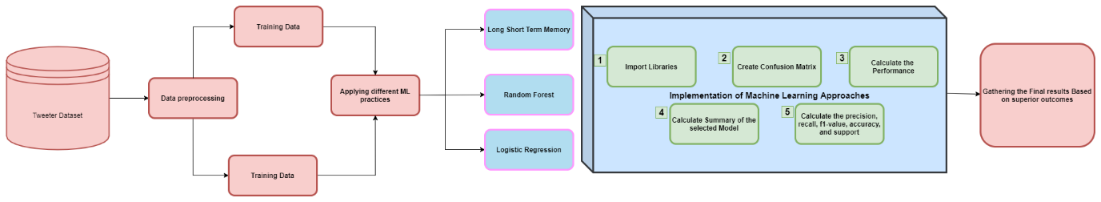


Figure 11 Final architecture design

## Performance computation

This is the part involved with each of the machine learning models to calculate the performance of the implementation in association with the testing data. This is the proof that from the used dataset or the need of the results that the applied model is effectively useful or not. From the values collected from the true positive and the true negative, adding up to the collection in which it will be divisible with the true negative and False-negative addition. This value is considered as the score of accuracy involved with the entire practice of experimental practice (Juneja and Ojha, 2017).

These are the most effective scores needed for analyzing the feasibility of the machine learning model is compatible with the selected dataset as well as the needs of the final results. This is the main reason for covering the part of machine learning, the possible accuracy of the model and considering the capability of other processes involved in the entire process of discovering some effective results and making change at a global level (Bouazizi and Ohtsuki, 2017).

# Chapter 4 Implementation

While following the details presented in the methodology associated with the planning of achieving a complete analysis of social media, different techniques use are involved. Data collection is initiated from the platform of Twitter to the preparation of the results at the completion of the experiment. The flow of the implemented code is divided into four major segments, beginning from the involvement of the experimental setup leading this practice to the involvement of the dataset to the implementation practice, then initiating towards the cleaning of the data in the segment of pre-processing covering the part to be involved with the selection of the most effective segments (Fatima, Mukhtar, Ahmad and Rajpoot, 2017). This segment prepares the data into two parts the training as well as testing. The next segment of the process is involved with the visualization of the collected data before the implementation of the classification models on the data. All these classifiers will provide the results directly or indirectly, pushing the entire experiment to the collection of the results in the form of performance as well as accuracy. So all of these mentions are the specific parts explained with the involvement of code snipped that how these processes are executed, associated with the outputs of that presented code.

## Data pre-processing

For that, the involvement of the Pandas library is required. This is considered an essential part of the pre-processing segment to be initiated with the cleaning of the dataset and making it more useful for the entire machine learning project. This is the separation phase for most of the important parts of the dataset to be examined in a different manner.



Figure 12 Import of the Pandas library

This is the initiation of the pre-processing part, in which the data related to the sentimental, ids, dateoftweet, username, and the data in the string is selected from the dataset.



Figure 13 Information about the size of the dataset

This line of code is used for gathering information about the size of the dataset, precisely talking about the length of the dataset.

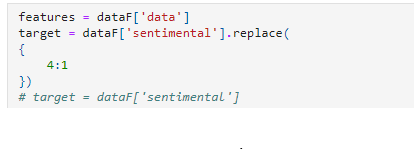


Figure 14 Collection of feature data

This is the code used for the collection of feature data needed to be extracted from the dataset. This process has a target column mentioned as sentimental.

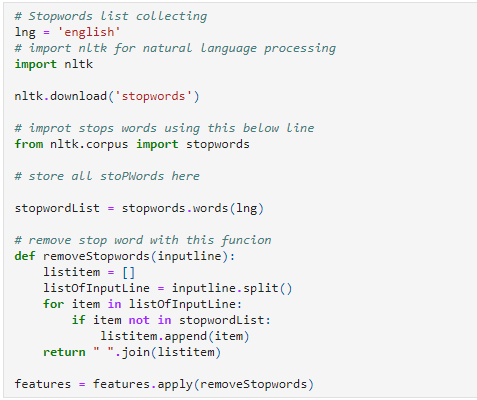


Figure 15 The development of the logic that this dataset

This is the snap of the code used here for the development of the logic that this dataset selected here consists of different kinds of values in a huge amount. So there is a necessity of cleaning it. This code consists of the selection of specific stop words that contain the word ‘english’. For that, the library of natural language processing is selected. This is capable of removing all the words initiated with the stopword mentioned in the code. The last line of the code shows the application of the feature selection.

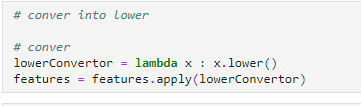


Figure 16 Selection of features

This is the part in which the selection of features is performed, containing the process of lowering the case and convert it into completely readable data.

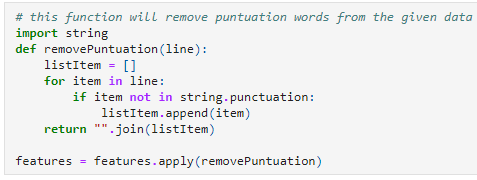


Figure 17 Data cleaning

This part of the code leads the progress of data cleaning to the next level with the removal of all the punctuations mentioned in the Twitter data.



Figure 18 Removal of all kinds of URLs

This is chosen for the removal of all kinds of URLs present in the data to make it more precise towards the analytical part.

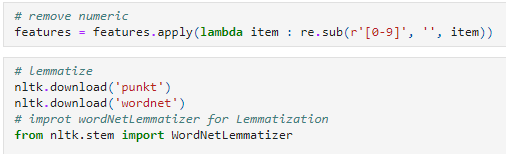


Figure 19 Removal of all the numeric data

For the removal of all the numeric data as well as the involvement of lemmatizer is performed.

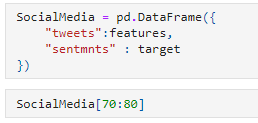


Figure 20 The construction of the frame for the data

This part contains the construction of the frame for the data to be segmented in the relation of tweets with the target of sentiments involved with it.

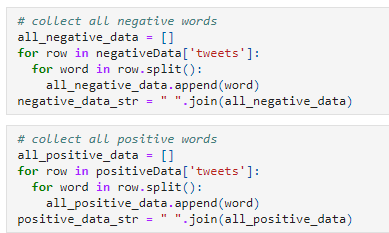


Figure 21 The calculation of all the positive and negative words

The demonstration of these two code blocks contains the calculation of all the positive words as well as the negative words involved in the entire set of data. This will help the process in generating the logical concepts with the experiment.

## Data visualization

The implementation part mentioned beyond this heading is associated with the development of the visualization practices into the code for the demonstration of data in a graphical or any other visual format.



Figure 22 The development of the word cloud using the data

This is the code used for the development of the word cloud using the data collected from the dataset and generated in the background color of white. This involved the different words collected and presented in an aesthetic manner. For the processing of this visualization, the library of matplotlib is imported as the variable name of the draw.

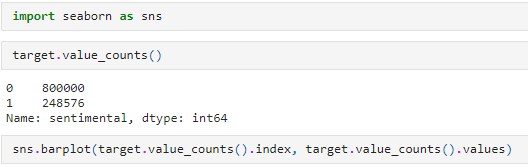


Figure 23 The seaborn library for the presentation of data collected from the dataset

This is the first visualization demonstrated in the experimental implementation. It contains the involvement of the seaborn library for the presentation of data collected from the dataset in the context of all the target values of the sentimental calculations or understanding made with the dataset.

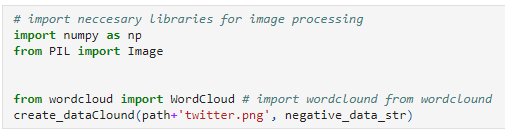


Figure 24 The library of wordcloud which provides the designing of the image

Initially, this visualization of all the words into an image the library of wordcloud which provides the designing of the image containing all the words to be displayed into it. In the first part of the visualization, the demonstration of all the negative words is gathered and presented. This includes the variance of the size of words depending upon their frequency.



Figure 25 Create the second visualization

This line of code is used here for the involvement of all the positive data to make it visible with the involvement of all the major words with large size and the least with the least size.

## Splitting of data

Before using the desired classification model on the selected dataset, there is a need to split the entire data into two segments. The first one will be called a training set, and the second one will be considered as the testing set for the machine learning practice.



Figure 26 The presentation of training and testing splitting

This is the segment of code presented above containing the presentation of training and testing splitting using the library of sklearn model selection.

## Application of machine learning models

In the path of using different classification models of machine learning, there are different challenges involved. Beginning from the analysis as well as evaluation of the performance scores of all these models in the context of training and testing datasets. This section of the experiment will demonstrate all the efforts as well as procedures involved in the implementation of different classification models to the dataset. All the practical part is presented ahead with the associated screenshots of the code.

### Logistic regression model

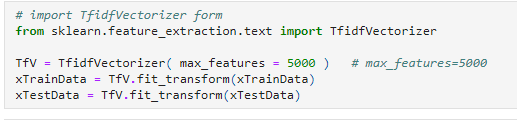


Figure 27 The machine learning practices

Before implementing any of the machine learning practices, the training and testing sets are extracted, and then selective data is used for the processing to gain a sustainable result in a considerable amount of time.

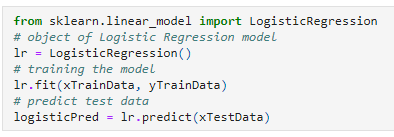


Figure 28 Prediction using the logistic regression model

With the help of a logistic regression model, there a prediction could be gathered with the involvement of the object of the logistic regression model associated with the training as well as the prediction of the test data to make a considerable result for the process.

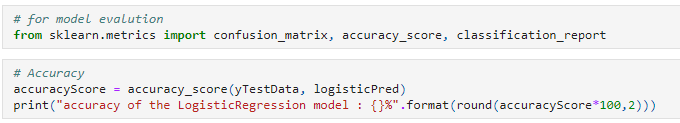


Figure 29 The evaluation practices into the experiment

This piece of code is used here for the importing of all the evaluation practices into the experiment to make sure how the data is analyzed and the correctness as well as the accuracy of the results.

In the next stage of the procedure, the involvement of the Random forest classification model is presented ahead. Before the usage of the random forest model, there is a selection of specific data from the dataset performed to make the result accurate as well as less time-consuming.

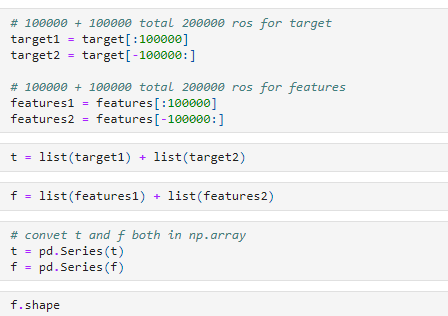


Figure 30 Collecting specific data from the testing dataset

This code presents the understanding of collecting specific data from the testing dataset to be used 200000 rows for the target set and the other 200000 rows for the features. These sets are listed in a specialized format and making them more useful than ever before.



Figure 31 The mentioning of the tweet data as well as the connected calculation of sentimental values

The handling of all the data involved with the dataset contains all the mentioning of the tweet data as well as the connected calculation of sentimental values. This also includes the collection of all the array types involved with it, such as the training data and testing data.

### ****Random forest classifier****

This implementation of a Random Forest Classifier involves the initial import of the classifier from its library of sklearn.

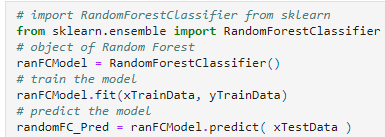


Figure 32 Initial stage of implementing the classifier

In the initial stage of implementing the classifier, there is an integral need to create an object of the model, train the model according to the requirement. And then, at last, generate the prediction scores.



Figure 33 The accuracy scores

The code mentioned above contains the understanding of the accuracy scores to be calculated for increasing the relevance of Random Forest Classification. This will provide the result in the form of Random Forest Classifier model accuracy.

### Long short term memory

This is the model of machine learning capable of covering all the small as well as big activities to be performed in the entire experiment and provide results about the overall involvement of the dataset into the relevance of the gathered results.

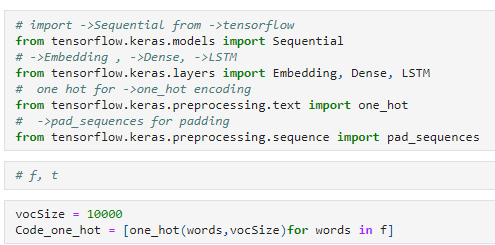


Figure 34 The process of initiating the LSTM

In the process of initiating the LSTM, there is a need to import it from its source library, which is the library of TensorFlow. In the next line of code, the other components needed for the execution of this process are collected and initialized for implementation.

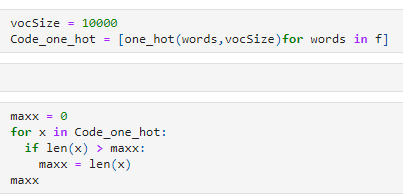


Figure 35 The calculation of the padding code

For the calculation of the padding code, the maximum value associated with the pre-processing segment is needed to be calculated.

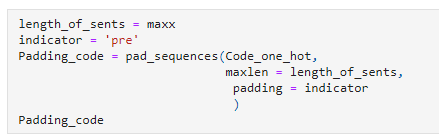


Figure 36 Maximum value evaluated

This is the part of the padding code which will be processed using the maximum value evaluated earlier.

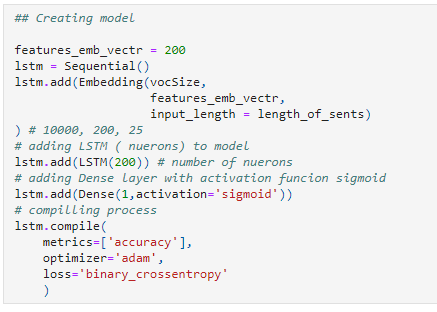


Figure 37 The creation of the models

This step contains the procedure involved with the creation of the models. These models will be used in a sequential manner for the processing of the LSTM to compile all the functions with the context of its accuracy as well as with the optimized evaluation of the results. This process is compiled into different layers, as shown in the figure.

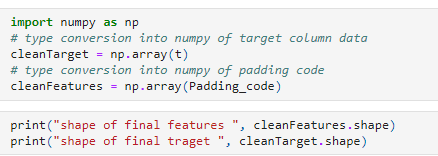


Figure 38 NumPy library functions and evaluation procedures

With the involvement of NumPy library functions and evaluation procedures, there are different calculations processed with the inclusion of padding code. This code will provide an understanding of the final features and targets of the process.

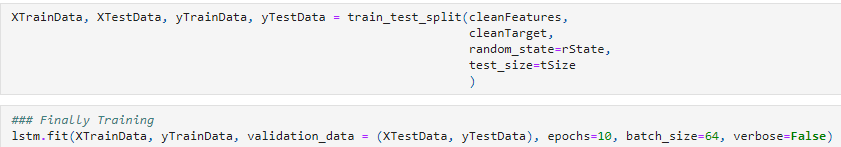


Figure 39 The evaluation of the data

The evaluation of the data, including the processing of the functions as well as the training of the data, is finally completed in this section of code.

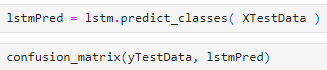


Figure 40 Generation of the confusion matrix

For the implementation as well as generation of the confusion matrix, this line of code is used.



Figure 41 The calculation of the accuracy score

# Chapter 4 Results

This part of the flow contains all the results gathered from each execution of the code mentioned, as well as explained in the above section. It includes all the segmentations as mentioned in the implementation.

## Data pre-processing

For the understanding of all the procedures involved in the pre-processing of the data, visualization of some data from the dataset is required, as mentioned below.

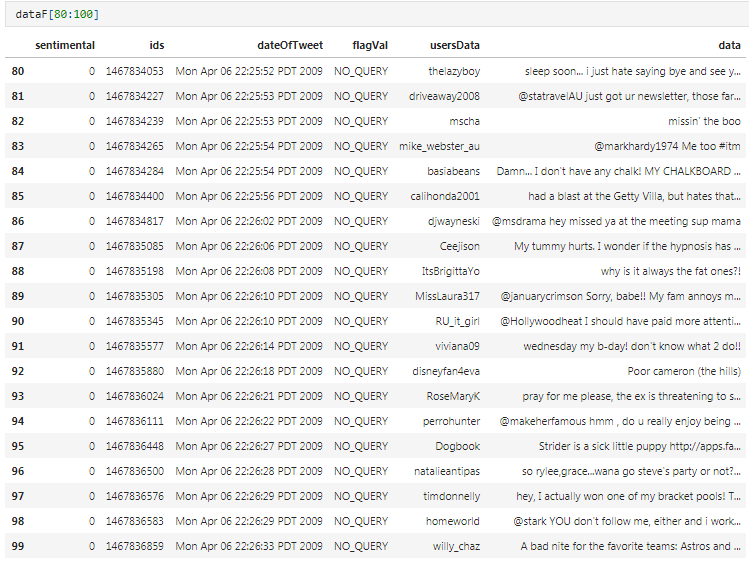


Figure 42 Data from the dataset

This screenshot contains the values categorized into six different columns. These columns are serial numbers, sentimental value, ids, dateoftweet, and the value of the flag, usersdata and the actual tweet in the form of data.

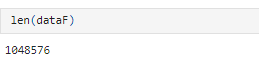


Figure 43 The length of the selected dataset

This is the length of the selected dataset presented. Around 1048576 records are present in the entire dataset.

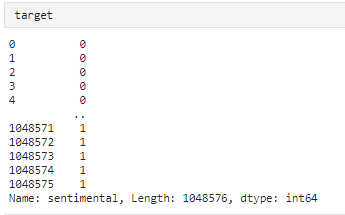


Figure 44 The extraction of all the target values

In the extraction of all the target values, this result is gathered. It includes the number of tweets associated with the sentimental value.

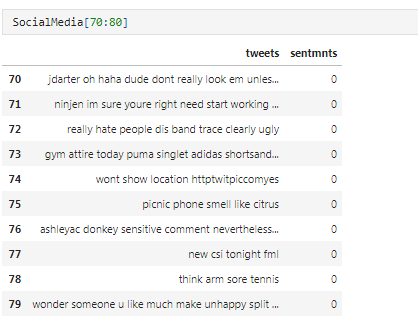


Figure 45 Showing processed data

After the implementation of all the data cleaning as well as selective attention, this data is collected as well as demonstrated in the original form in which it will be used in the machine learning classification model.

## Data visualization

In this part, all the visualizations results are gathered as well as demonstrated for the understanding of all the important data generated using the code of implementation.

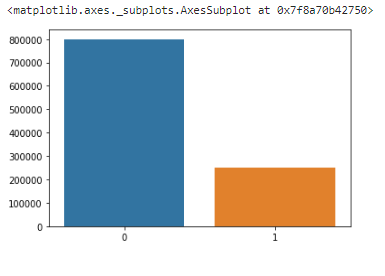


Figure 46 The values associated with the sentimental connected

This is the demonstration of the overall data collected as well as presented in the form of bars. It includes the values associated with the sentimental connected with different tweets.

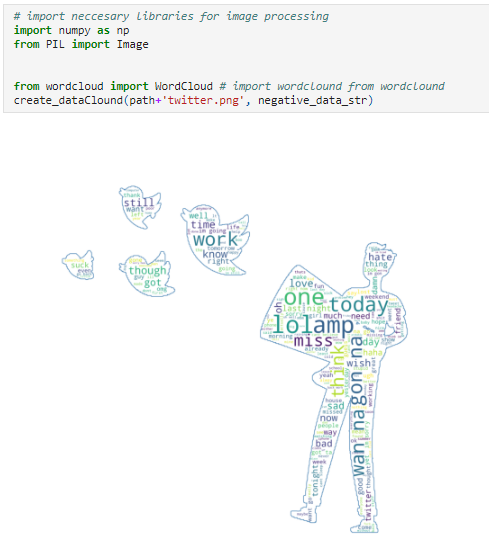


Figure 47 Graphical representation is generated

With the involvement of the wordcloud library, this graphical representation is generated with the combination of negative words associated with the frequency of these words.

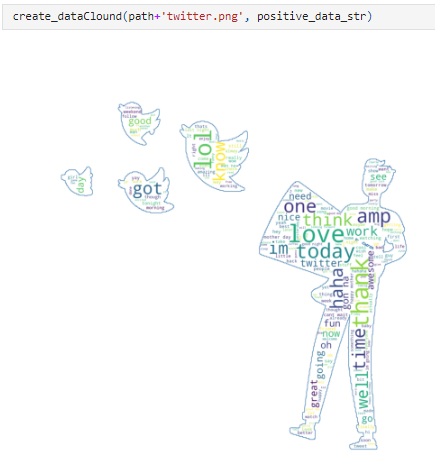


Figure 48 Graph of positive words associated with the frequency

With the involvement of the wordcloud library, this graphical representation is generated with the combination of positive words associated with the frequency of these words.

## Splitting of data

After the visualization, the process of data splitting is performed. It contains the segmentation of the data according to processing needs.

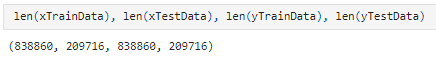


Figure 49 The classification practices with different models

Here are two sections divided to be processed with the classification practices with different models.

## Application of machine learning models

### Logistic regression model

After the implementation of all the visualization and splitting, there is the model of logistic regression.

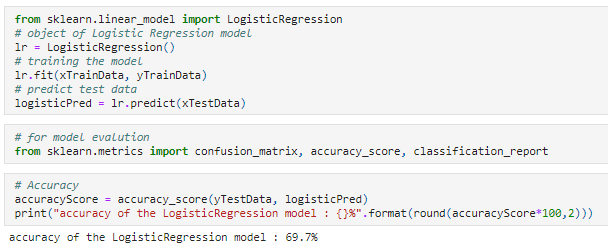


Figure 50 The accuracy score of the logistic regression

According to the accuracy score of the logistic regression, the model provided the results with 69.7%.

### ****Random forest classifier****

After the completion of the logistic regression practices, the classification of the Random Forest Classifier is implemented.

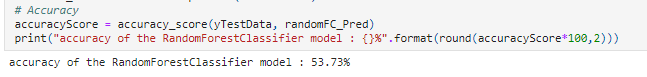


Figure 51 The accuracy score of the Random Forest Classifier

According to the accuracy score of the Random Forest Classifier, the model provided the results with 53.73%.

### Long short term memory

After the completion of the Random Forest Classifier practices, the classification of the long short-term memory is implemented.

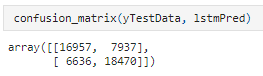


Figure 52 Confusion matrix generated

The confusion matrix generated from the data used in the LSTM processing is demonstrated above. Talking about the score of accuracy involved with the process of LSTM is around 70.85%.



Figure 53 Accuracy involved with the process of LSTM

# Chapter 5 Discussion

There are different kinds of results gathered from the knowledge gained in the process of literature review and making it more effective towards the comparative understanding between the literature-related information and the experimental practices. According to the article selected in the literature review, the results are gathered on three of the most effective perspectives the performance scores, accuracy and the precision of the practical results gathered. The scores are gathered from the literature are in the context of all the criteria of the performance analysis, such as the accuracy, precision scores as well as recall of the model implementation.

While finding the scores mentioned in the works of literature are on the basis of Twitter. There are different kinds of understanding involved with the sentimental analysis of the data collected from the platform of Twitter which is becoming the keen interest of most of the analytical practices worldwide. Beginning from the accuracy of the implementation is achieved to the 74% to make a significant mark on the race to get the highly relevant as well as the considerable result. All these results gathered here are involved with the usage of the logistic regression practices and not with the other implementation. If the comparison of the accuracy is considered with some other literature, then the LSTM will be at the top of the list of priorities. As presented in some of the most relatable explanations about the random forest, it is a highly useful concept to be implemented in this scenario of sentimental analysis. There are different kinds of practices involved with the understanding as well as application of the machine learning approaches into the implementation of the feature extraction, identifying the best algorithms to be implemented in the field of research. There are different aspects associated with the demand as well as the need for the results in a persistent manner that all the requirements of the identification of the sentiment involved with the collection of tweets.

There are some other involved of the practices included in the experimental practices of the machine learning models. From the beginning of the pre-processing segment, the involvement of specific preparations started with the selected dataset. In the field of gathering the sentimental details of the text images, videos, as well as all kinds of data sources, then the practice of analysis becomes more complex to the end of the segment, there is no such demonstration that includes these kinds of details than ever before. In the implementation of the Naïve Bayes practices to the analysis of social media-related data, then and only then the fluctuation in the results starts. The involvement of the highly impactful application of the algorithmic techniques into the work of total expertise will lead this practice of gathering the analyzed results with the highest possibility of getting the scores of accuracy mostly above the value of 75%. With the features as well as selected factors of the processing, it could be the work of art that is capable of providing the results to the highly approachable practices. This is the work towards excellence in the analytical as well as machine learning approaches and providing highly progressive learning in the field of complex processing of long short-term memory algorithms for the purposes of prediction.

To fulfill the needs of prediction from the tweet, data sets are taken to simulate the framework in which 1048571 numbers of Twitter are stored in the row, and six columns are available in the data sets. The simulation of the proposed model results is analyzed on the basis of the F1 score, in which the results are in the form of positive and negative aspects. The Bambara text, which is used for the sentimental analysis, is totally unique, so to eliminate the problem related to the text written, the dictionary is used in which the different types of characters and word index are mentioned to create the text characters and the word placement in the place of the pre-training words vectors. In the research, the effect of review span in the approach and efficiency in between them are compared. So after simulating the approach. It is reported that the one-layer CNN model plays an effective outcome by achieving an accuracy of 83.23%. After simulating the proposed approach by using the different ML algorithms, it calculated the long short-term memory algorithm provides the best results among all other algorithms.

# Chapter 6 Conclusion and future work

## Conclusion

The experimental practical, as well as the practices involved in the understanding, gathered from the initial literature review, there is a lot of learning accomplished. This kind of learning increased the level of knowledge in the field of using machine learning for the gathering of data, processing it, cleaning the data, making it usable for different classification models, which will provide the required results associated with the analysis of data. All the models involved with the machine learning practices are associated with a different kind of use case according to the need of results. In this experimentation of the machine learning models, the goal is to analyze different tweets, as well as other kinds of comments, that happened on the platform of Twitter to understand the sentimental emotion attached to them. In most of the cases where the sentimental analysis is performed, there are overall two kinds of emotions connected to most of the cases, that is, the negative emotion and the positive emotion. This emotion is considerable with the analysis of the used words by the person or just by the intention of the sentence making it significantly positive or negative.

In the overall research of machine learning practices involved with the handling of huge datasets to perform a particular task providing the results in an adequate manner is hardly manageable without the understanding of the most suitable classification models. In this research of experimental implementation of the sentimental analysis, there are three kinds of machine learning models are involved in getting the understanding about the best model to be used for the processing of the complete data. The model of logistic regression, random forest classifier, as well as the implementation of the long short term memory for the gathering of accuracy scores of it. The visualization is also used in this experimental practice to increase the interest as well as an attractive way to express the information in a small as well as effective manner. Majorly the implementation of long short-term memory provided the highest scores of accuracy in the overall practical experience of experimental practice. The score of accuracy with LSTM is coming out to be 70.85% as the highest score from all over the practical.

## Future work

In the future of machine learning implementation, there is a different kind of practices could become affordable as well as easy to be implemented in the most of the analytical practices. The understanding involved with the sentimental analysis could be enhanced with the diverse involvement of most of the models available in the machine learning concepts to make the experimental practices more precise in the direction of becoming more reliable than ever. For providing the overall performance scores better than the presented results, there is a different kind of add-on practices that could be involved in the model that supports the variety of scopes and carry the entire work of machine learning analysis to a whole new level. If there is the involvement of a support vector machine in the experiment, then the involvement of grid search into the processing of machine learning concepts could bring more accuracy to the overall scores of performance such as accuracy, precision and many more. So the involvement of any other supportive element could lead the work to some other level of accuracy scores or reliability in order to grow in an exponential manner.

# References

Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H. and Shah, J., 2020. Sentiment analysis of extremism in social media from textual information.

Bouazizi, M. and Ohtsuki, T., 2017. A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. IEEE Access, 5, pp.20617-20639.

Chaabani, Y., Toujani, R. and Akaichi, J., 2017. Sentiment Analysis Method for Tracking Touristics Reviews in Social Media Network. Intelligent Interactive Multimedia Systems and Services 2017, pp.299-310.

chniques and their interactions for Twitter sentiment analysis. Expert Systems with Applications, 110, pp.298-310.

Coban, O. and Ozel, S., 2018. An Evaluation of Character Level N-gram Termsets in Text Categorization. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP),.

De Bruyn, P., 2020. Predicting behavioral profiles of online extremists through linguistic use of social roles. Behavioral Sciences of Terrorism and Political Aggression, pp.1-25.

Fatima, I., Mukhtar, H., Ahmad, H. and Rajpoot, K., 2017. Analysis of user-generated content from online social communities to characterise and predict depression degree. Journal of Information Science, 44(5), pp.683-695.

Fernandez, M. and Alani, H., 2021. Artificial intelligence and online extremism. Predictive Policing and Artificial Intelligence, pp.132-162.

Gan, W., Su, B. and Li, Y., 2020. An Electric System Abnormal Analysis Framework based on Natural Language Processing. 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD),.

Gozuacik, N., Sakar, C. and Ozcan, S., 2021. Social media-based opinion retrieval for product analysis using multi-task deep neural networks. Expert Systems with Applications, 183, p.115388.

Hajiali, M., 2020. Big data and sentiment analysis: A comprehensive and systematic literature review. Concurrency and Computation: Practice and Experience, 32(14).

Hao, J. and Dai, H., 2016. Social media content and sentiment analysis on consumer security breaches. Journal of Financial Crime, 23(4), pp.855-869.

Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018. Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical and Computational Applications, 23(1), p.11.

Hassan, A., Hussain, J., Hussain, M., Sadiq, M. and Lee, S., 2017. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression.

Jianqiang, Z. and Xiaolin, G., 2016. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. [online] Doi.org. Available at: <http://doi.org/10.1109/ACCESS.2017.2672677> [Accessed 6 August 2021].

Jianqiang, Z. and Xueliang, C., 2015. Combining Semantic and Prior Polarity for Boosting Twitter Sentiment Analysis. 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity),.

Jianqiang, Z., 2015. Pre-processing Boosting Twitter Sentiment Analysis?. 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity),.

Jianqiang, Z., 2016. Combing Semantic and Prior Polarity Features for Boosting Twitter Sentiment Analysis Using Ensemble Learning. 2016 IEEE First International Conference on Data Science in Cyberspace (DSC),.

Juneja, P. and Ojha, U., 2017. Casting online votes: To predict offline results using sentiment analysis by machine learning classifiers. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT),.

Kathuria, R., Gautam, S., Singh, A., Singh, A. and Yadav, N., 2021. Sentiment Analysis on Images Using Convolutional Neural Network. Advances in Intelligent Systems and Computing, pp.189-201.

Konate, A. and Du, R., 2018. Sentiment Analysis of Code-Mixed Bambara-French Social Media Text Using Deep Learning Techniques. Wuhan University Journal of Natural Sciences, 23(3), pp.237-243.

Kong, L., Li, C., Ge, J., Zhang, F., Feng, Y., Li, Z. and Luo, B., 2020. Leveraging multiple features for document sentiment classification. Information Sciences, 518, pp.39-55.

Liu, B., 2018. Text sentiment analysis based on CBOW model and deep learning in big data environment. Journal of Ambient Intelligence and Humanized Computing, 11(2), pp.451-458.

Mussiraliyeva, S., Bolatbek, M., Omarov, B. and Bagitova, K., 2020. Detection of Extremist Ideation on Social Media Using Machine Learning Techniques. Computational Collective Intelligence, pp.743-752.

Mutinda, J., Mwangi, W. and Okeyo, G., 2021. Lexicon‐pointed hybrid N‐gram Features Extraction Model ( LeNFEM ) for sentence level sentiment analysis. Engineering Reports,.

Potamias, R., Siolas, G. and Stafylopatis, A., 2019. A Robust Deep Ensemble Classifier for Figurative Language Detection. Engineering Applications of Neural Networks, pp.164-175.

Ramadhan, W., Astri Novianty, S. and Casi Setianingsih, S., 2017. Sentiment analysis using multinomial logistic regression. 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC),.

Sakketou, F. and Ampazis, N., 2020. A constrained optimization algorithm for learning GloVe embeddings with semantic lexicons. Knowledge-Based Systems, 195, p.105628.

Sciandra, A., 2020. COVID-19 Outbreak through Tweeters’ Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings. 2020 IEEE Symposium on Computers and Communications (ISCC),.

Sharif, Mumtaz, Shafiq, Riaz, Ali, Husnain and Choi, 2019. An Empirical Approach for Extreme Behavior Identification through Tweets Using Machine Learning. Applied Sciences, 9(18), p.3723.

Sharma, A., Chaurasia, S. and Srivastava, D., 2020. Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec.

Singh, N., Tomar, D. and Sangaiah, A., 2018. Sentiment analysis: a review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing, 11(1), pp.97-117.

Singh, T. and Kumari, M., 2016. Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89, pp.549-554.

Symeonidis, S., Effrosynidis, D. and Arampatzis, A., 2018. A comparative evaluation of pre-processing teAbd Elaziz, M., Dahou, A., Abualigah, L., Yu, L., Alshinwan, M., Khasawneh, A. and Lu, S., 2021. Advanced metaheuristic optimization techniques in applications of deep neural networks: a review. Neural Computing and Applications,.

Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B., 2019. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. Entropy, 21(6), p.589.

# Appendix



Figure 54 Import of the Pandas library



Figure 55 Information about the size of the dataset

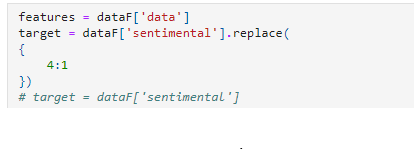


Figure 56 Collection of feature data

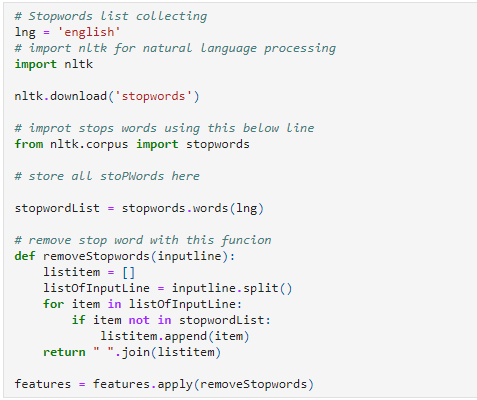


Figure 57 The development of the logic that this dataset

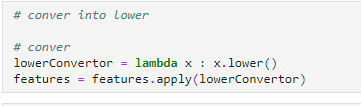


Figure 58 Selection of features

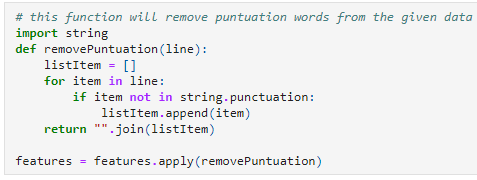


Figure 59 Data cleaning



Figure 60 Removal of all kinds of URLs

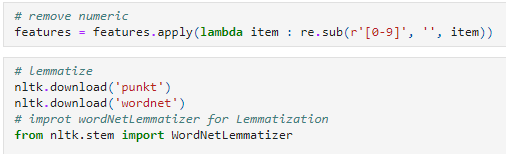


Figure 61 Removal of all the numeric data

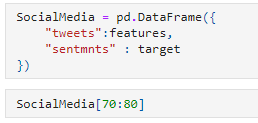


Figure 62 The construction of the frame for the data

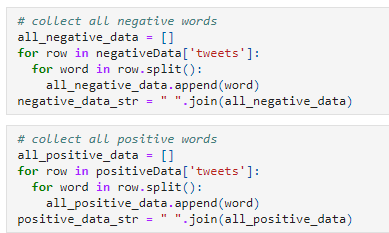


Figure 63 The calculation of all the positive and negative words



Figure 64 The development of the word cloud using the data

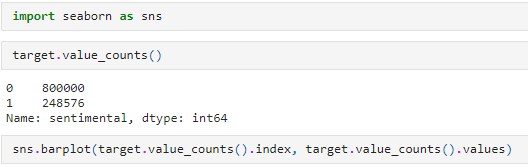


Figure 65 The seaborn library for the presentation of data collected from the dataset

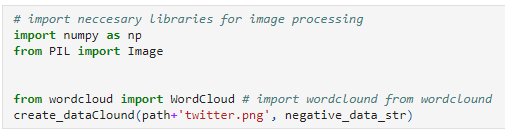


Figure 66 The library of wordcloud which provides the designing of the image



Figure 67 Create the second visualization



Figure 68 The presentation of training and testing splitting

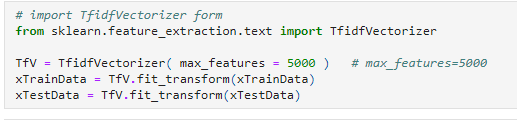


Figure 69 The machine learning practices

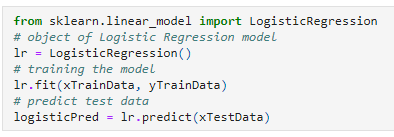


Figure 70 Prediction using the logistic regression model

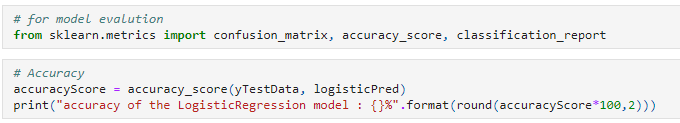


Figure 71 The evaluation practices into the experiment

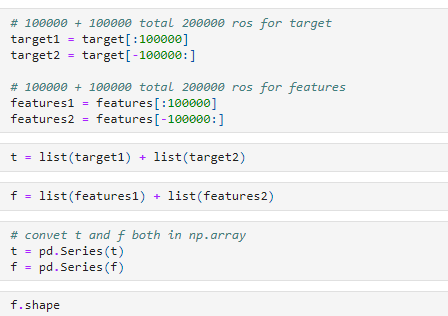


Figure 72 Collecting specific data from the testing dataset



Figure 73 The mentioning of the tweet data as well as the connected calculation of sentimental values

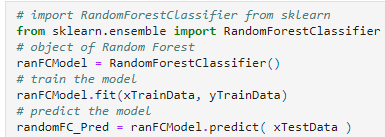


Figure 74 Initial stage of implementing the classifier



Figure 75 The accuracy scores

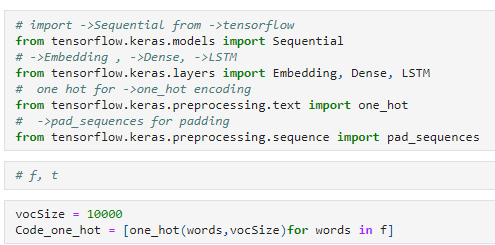


Figure 76 The process of initiating the LSTM

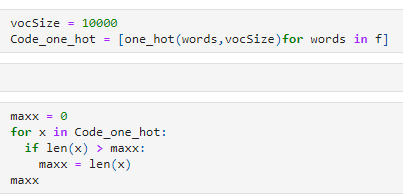


Figure 77 The calculation of the padding code

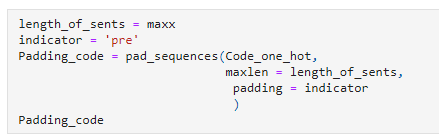


Figure 78 Maximum value evaluated

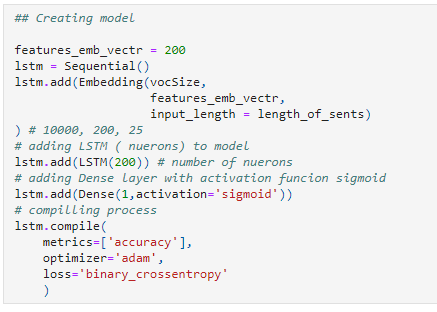


Figure 79 The creation of the models

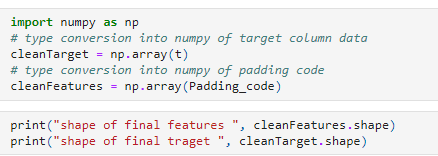


Figure 80 NumPy library functions and evaluation procedures

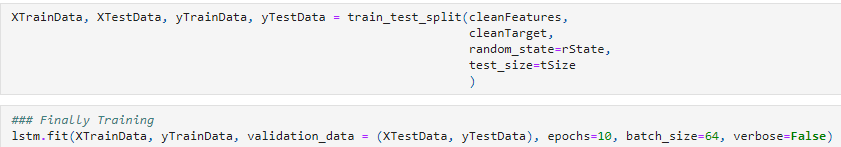


Figure 81 The evaluation of the data

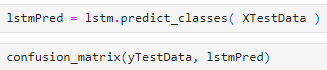


Figure 82 Generation of the confusion matrix



Figure 83 The calculation of the accuracy score

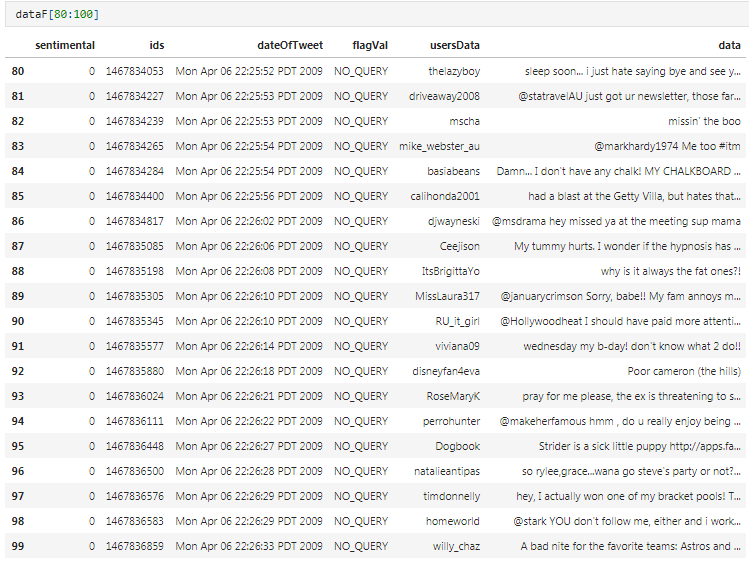


Figure 84 Data from the dataset

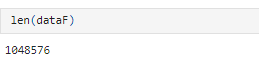


Figure 85 The length of the selected dataset

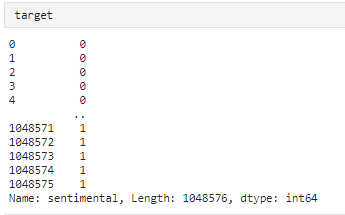


Figure 86 The extraction of all the target values

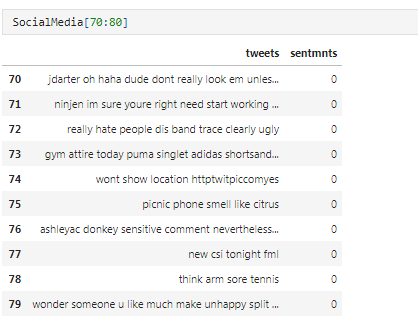


Figure 87 Showing processed data

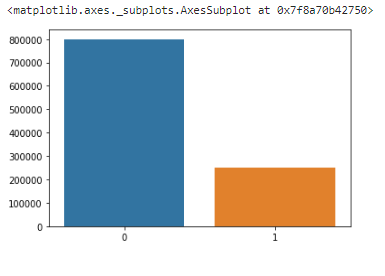


Figure 88 The values associated with the sentimental connected

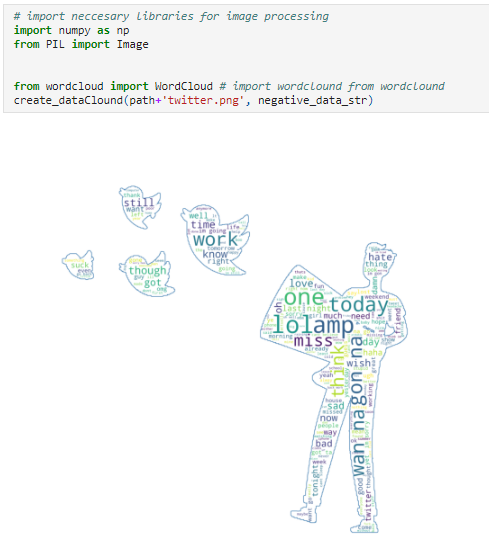


Figure 89 Graphical representation is generated

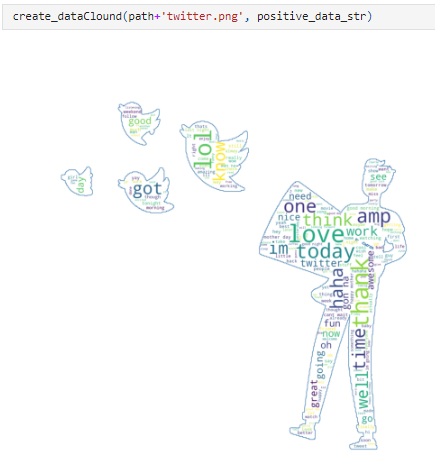


Figure 90 Graph of positive words associated with the frequency

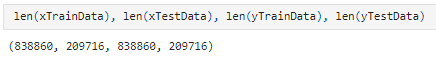


Figure 91 The classification practices with different models

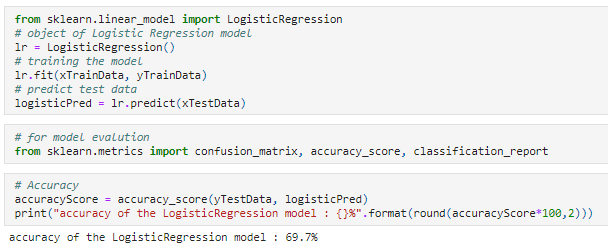


Figure 92 The accuracy score of the logistic regression

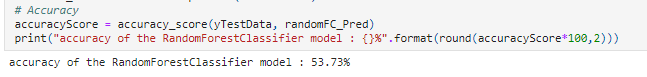


Figure 93 The accuracy score of the Random Forest Classifier

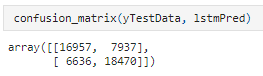


Figure 94 Confusion matrix generated



Figure 95 Accuracy involved with the process of LSTM