University of Essex

Department of Mathematical Sciences

MA981: DISSERTATION

Sentimental Analysis of Social Media With Python

Ronak Chudasama

Registration Number: 2112224

Supervisor: Dr Osama Mahmoud

November 25, 2022

Colchester

TABLE OF CONTENTS

Section 01: Introduction

1.1 Introduction

The sentiment analysis is known as a technique that can assist with predicting a huge amount of data. This also refers to the opinion mining so that a technique of NLP recognizes the insight of the texts that are analyzed. With the help of this technique it is possible to determine the popular categories of a specific service, product or even an idea. The approaches of data mining are mainly applied here with the help of AI and ML then analysis of sentiment can be performed. Here in this project mainly the social media data will be analyzed and overall information will be extracted from the big data.

1.2 Aim & Objectives

The aim of this project is to perform a sentiment analysis on social media information so that decision making approaches can be identified. With the help of python the prediction can be performed easily. The social media data will be gathered in the first place and after that the huge amount of data will be pre-processed to apply algorithms. These analysis approaches help the data to divide into three categories and they are known as training, test and validation data. The algorithms of classification will be applied to analyze the social media data at the end. This is the overall aim of the project.

- To gather data related to social media for applying the technique of sentimental analysis.
- To prepare the raw data for further processing of data mining techniques.
- To divide the data into training and testing.
- To apply a suitable sentiment algorithm to receive the final outcome of the project.

1.3 Research Question

The research question will help the researcher to reach a better aim at the end of the project.

- How will the data be collected to apply the technique of sentiment analysis?
- How will the preparation of raw data be conducted so that data mining techniques can be applied upon?
- How is the splitting of data performed into training and testing?

- What will be the algorithms mainly applied to meet the outcome of the project?

1.4 Research Significance

The significance of research is mainly conducted to understand how much importance a research might hold. The long period of research and analysis should consist of an appropriate significance at the end. Otherwise the researchers will not be able to receive the confidence of conducting such research. For every researcher it is important to know what will be the ultimate research deliverables so that they can be more interested in performing the research. In this project a machine learning analysis is being applied to big data and for that purpose it is important for the researcher to know the ultimate outcome and meaning of the research (Pérez *et al.* 2021). How the project can be applied to real life and how people can take advantage of the analysis. If the social media data can be analyzed with the help of proper sentiment techniques then the marketing strategy can be fully grasped. The main reason for social media at first was to keep touch with family and friends but in recent days social media is also becoming a platform for business where different promotions are also taking place. The different opinions of the consumers also can be found on social media these days. The opinion of the customer holds a significant power these days and that is why it is crucial to apply different prediction techniques so that the process of decision making becomes easier to handle. The service, product or idea can be easily found with the help of sentiment analysis. This is the reason why in the present days it is known as a real time data prediction tool. With the help of natural language processing the effect of a certain category can be shown and that provides a better visualization of information that individuals can apply to benefit their own business on social media.

1.5 Research Framework

The research framework is known as the proper structure of a project so that in which way a data is being analyzed can be found simply. In this project the structure will be followed as a general one. First the introduction of the project will be described where the aim and objectives of the project will be presented. After the aim and objectives the main goal is to state the research question so that the technique of achieving the deliverable can be found easily. After this the significance of the research will be shown where the overall importance of performing the project will be presented for the researchers. After the introduction the next chapter will be literature

review where different past approaches of performing the sentiment analysis will be discussed. It is important to look into the past research as that informs the coming obstacles that the present researcher might face in the time of analysis (Bhavsar *et al.* 2019). After looking into the previous approaches now the researcher will finally proceed with the methodology of performing the sentiment analysis of social media data. The methodology will present every little detail of the approaches that the researcher will follow in the time of analysis. After this the analysis will be discussed to understand how the researcher is going to achieve the goal and in this case the deliverables will be based on performing a sentiment analysis. The algorithms will be shown how they are applied to the data to achieve the outcome of the analysis. After this the outcome of the analysis will be discussed and a thorough discussion of the project will be shown as well. Lastly a conclusion and recommendation will be presented describing all the future precautions that the researchers might recommend to perform such analysis.

1.6 Conclusion

The overall project significance will help to understand how much importance this particular research holds. This project is going to be performed in such a way so that the deliverables can be achieved with simplicity. These techniques of sentiment analysis will present how the overall outcome of the research will be. Next the literature review will be presented to observe the past approaches of similar implementation.

Section 02: Literature Review

2.1 Introduction

There is a lot of information floating around the social media and a lot of updates are given during a second in the different social media platforms and analyzing these messages is not a simple task and requires complex techniques and methodology to process these data. The size of the data is very huge and cannot be analyzed using the common techniques is not possible, there needs to be some technology and technology needed to be employed for solving the complex problems and analyzing the emotions and sentiment of the social media user. It is required to carry out extensive research for collecting various types of information like understanding the different characteristics

and messages that are shared on the different platforms. The research will assist in finding different measures and techniques that can help in analyzing the data that has been collected through various social media and different contexts regarding the different types of opinions and emotions. The insights that are generated through the analysis are very much helpful for different organizations and brands to understand their performance. It will help understand the people's opinions and attitudes toward various topics which are beneficial for the company for the development of the different types of strategies and planning for targeting the actual consumers or customers. For finding out the various information regarding the sentiment analysis and the method using which a successful analysis can be done various research paper and technical articles were evaluated and in this section of the chapter those articles were selected and reviewed which provided the techniques and various methods for solving the problem and understanding its importance. The authors view will be shared in the review and their proposal and opinion will be discussed explaining their importance. The chapter will also analyses and find out the limitation that has been found out during the literature review which will help in understanding the gaps and solving the problem in this project.

2.2 Empirical Study

According to **Haliyana Khalid *et. al*. 2019**, Sentiment analysis is a method for extracting, converting, and interpreting opinions from text and categorizing them as positive, negative, or natural sentiment. It uses "Natural Language Processing" (NLP). To better understand their customers and decide how to improve their products or services, the majority of previous studies used sentiment analysis to analyses reviews of movies or products. Subsequently at the beginning of the previous decade, academics have been studying sentiment analysis; the majority of these publications first appeared after 2004 and have since exploded in number (Manguri *et al.* 2020). The three stages of sentiment classification are: sentence level, document level, and feature level. The objective is to categorize the viewpoint into positive and negative sentiment based on the language, document, or feature. Machine learning and "lexicon-based approaches" are one of the two main techniques for sentiment analysis that have been identified. In contrast to the lexicon-based technique, which counts the positive and negative words that are associated with the data, machine learning approaches use analytics and recognize sentiment from data. Researchers have been creating a fresh, reliable model for sentiment analysis. However, it is difficult to create a

model when the majority of it is created for the English language. On the other hand, a recent study demonstrates that sentiment analysis model design exists in various languages. Regarding the use of sentiment classification, it has reportedly been done in contexts related to marketing, business politics, and public action. The author describes that the application of this analysis is not limited to only a single portfolio but has a wide variety of ranges. The application of this analysis can be done in the sector of business, marketing as politics uses this analysis for understanding their voter's opinions as well as it had a huge contribution in the health sector.

According to **Zulfadzli *et al.* 2019**, the analysis has vast application in the different fields where there is a requirement of decision making and strategy development. The sentiment analysis can be applied to the various events happening in the world, the events like activities, sports, or kind of natural or manmade calamities. It can help in understanding the people opinion towards various topic happening all around the world e.g., opinion toward ISIS for the people living in western countries. Sentiment analysis enables spreading knowledge about data security and the risk of security breaches (Diyasa *et al.* 2021). Additionally, it serves as a template for how businesses should handle security breaches in order to influence public opinion. Additionally, sentiment analysis was done on the "social media " to evaluate the employment sentiment score and unemployment rate. We can see how sentiment analysis is used in healthcare, where a framework for providing sentiment as a service has been developed, and where it is used to locate disease outbreaks. In addition, sentiment analysis helps plan an effective rescue effort by identifying people's emotional needs during a disaster. Additionally, by watching and analyzing emotions in text, sentiment analysis enables determining a person's level of depression. In his published research paper while explaining about the application of the sentiment analysis in the different fields, he also mentions it in politics and how politicians use the sentiment analysis to understand their voters as well as their opinions (Pokharel B.P. 2020). In the conducted research it has been found that using the twitter data more reliable results and insights are found and it has been stated that there is correlation of 94% is found from the polling data and has become one of the primary choices for understanding and utilizing the "complicated polling techniques".

According to Pérez 2021, various businesses involved in social media also use the feedback that is received through the social media for the sentiment analysis of their consumers and it is proved

to be an important factor for the improvement of business as required actions can be taken by understanding these feedbacks from the social media. The views of the customers can be easily analyzed and can understand the experience with products of the company that it's providing and various decisions are driven through the analysis that is carried out. When an opinion is specially extracted from a text document it is important to gather an amount of knowledge that can be applied to understand a particular approach of analyzing the texts (Gunawan *et al.* 2020). As the data is related to a vast amount of social knowledge it is important to understand if the different perspectives are analyzed or not. The issue that the researchers have found is utilizing the tool that is known as a mining tool. In social media there might be many other languages present, to process all the different languages it is important to first use a translator that will convert the overall languages to a single language that will be understood by a machine learning technique. This issue needs to be addressed using a different toolkit presented by python for the analysis of sentimental. There are different libraries available in NLP that will show how effectively a sentiment analysis can be applied (Basarslan *et al.* 2020). These known techniques are known for solving many other machine learning issues these days. Finding emotions from a particular text is not simple and with the help of NLP it is possible to extract the feeling that is hidden behind the expression. There are different NLP techniques that might assist with the process of sentiment analysis.

According to Bhavsar 2019, twitter is known as a website that is popular in the platform of social media and the huge amount of data can also be analyzed with the help of sentiment analysis. This helps to assist the individuals with expressing their feelings and thoughts based on several subjects for the past decades. The market analysis can be performed with the help of different techniques of ML. There are many other insights of big data that can be extracted from the analysis of machine learning. The accuracy of different algorithms might help to further predict the emotion of languages (Siddharth *et al.* 2018). The improvement of prediction can be developed when several techniques of ML are followed. The web technologies are increasing to another level thus it is crucial to understand how they are being applied to different sources. The model that is extracted from the output will help to see how the raw information is extracted in an accurate format. The categories are divided into two expressions and they are known as positive and negative. The sentiment analysis is known for solving many issues and it is based on analyzing different expressions. The elimination of different features can be appropriately done with the help of

machine learning approaches. The system of sentiment prototype will help to analyses the social network information. There are different techniques and approaches to how the extraction will be performed (Anand *et al.* 2020). The Python will show the overall prediction of data science so that related applications can be extracted out of the text that are being analyzed. The classification of several approaches will help to show how the prototypes are being designed to understand the expression. The accuracy of the model is the ultimate goal that the researcher wants to achieve at the end of the research so that algorithm implementation can be conducted.

According to Manguri 2020, from the past research the overall expansion of social media can be observed based on the internet that is rapidly spreading its opportunities. The use of commercial segments and also the proper accessibility of information can be found from a proper academic research done by the author. The social information that is found on the internet consists of several events that appear in the daily lifestyle. The data of twitter can be based on different reviews of movies, songs or an event that was recently known as a trending topic on the internet. With the help of python programming the several libraries the language of the social media will be analyzed and based on that the measurement of sentiment analysis will be performed (Shekhawat B.S. 2019). The gathered data from twitter is mainly to understand what different expressions are provided by different groups of people so that it can be effectively applied to solve different parts of expressions shown in the social media. The field of natural processing language is known as sentiment analysis. Social media is known as a platform where different opinions are submitted by different individuals so that they can be analyzed and based on them a proper decision making approach can be found. This is the overall understanding of sentiment analysis in the field of machine learning prediction. The sentiment analysis is known as an emotion that can be found inside a phrase. With the help of this analysis analyzing customer feedback or any other survey response can easily be analyzed. Tracking social media and experiencing customer trends are possible to perform with the help of sentiment analysis and its different techniques. The feedback and its usefulness is possible to analyses with the help of sentiment analysis. This will show the overall guidance on how the sentiment analysis works.

According to Diyasa 2021, sentiment analysis of social media refers to the process when one understands applying information about what people think or feel about one's products, brand

image and the services that they offer. In this current situation, this has even gone above the vanity metrics such as the mentioning, liking and other activities. Entertainment of social media analysis refers to gathering and evaluating data in the posts which people use to share about one's company, its products and services on the digital media platform. This analysis is considered to be very necessary regarding the well-being situation of the organization. This helps to understand the organization what people think, feel and write about their companies and its services. It will define one of the major factors like whether the public or the market is happy or unhappy with a product that a company offers (Singh G. 2020). If a certain product is preferred by a large amount of audience, then using these social media posts, the companies will be able to access how many consumers and for which locations they are demanding their products and services; on the basis of which they will be able to do the needful like opening some extra branches where the market need is more than the other places. These patterns of preference or the sentiments of people are defined by using the programming language python. This use of artificial intelligence like the use of python will help the organizations to recognize the preference of the consumers. The present obstacles can be analyzed with the help of sentiment analysis.

2.3 Research Gap

There are some limitations that sentiment analysis has and they cannot be ignored as in the time of analysis these gaps should be the most powerful concern for a researcher. The first factor that the researchers have faced is based on the problem of Tone where it becomes difficult to interpret and understand the overall words that are written is fine but figuring out the expression becomes difficult (Chandio *et al.* 2019). The huge amount of data is impossible to understand and the process is quite difficult because the response consists of both the expression called the positive and the negative. The subjective sentiments are difficult to find out when a particular brand is looking for it. With the help of accurate tone, finding out the overall expression is not that easy for the researchers. Second comes to the polarity where certain words like hate and love or negative and positive presents with the help of polarity. These words are understandable by the monitor but there are some words like average or not that are bad. These are not a 100% yes or no and in that case these are the expressions that cannot be analyzed with the help of sentiment analysis. The sarcasm and irony that is used by people in the conversation or in social media memes. The legitimate context of a sentence is very much difficult to understand when several tools are being

used to detect the response. Next comes to the emoji that are present inside the social media expressions that are posted by humans. The NLP is mainly created to be a specific language specific expression. When NLP has the ability to extract emotions from an image and in that case emoji are known as languages in the first place. There are many emotions that might work like an expression analysis and in that case emoji are deleted when the NLP is used for analyzing.

2.3 Conceptual Framework

The conceptual framework is known as a framework that consists of two types of variables and they are known as dependent variable and independent variable. There are some steps that have to be analyzed when a framework is being created. The first step is to choose a proper research question and that helps to identify what are the ultimate deliverables that the researchers want to achieve and find out to receive an appropriate focus and this step is essential (Sharma *et al.* 2020). It is important to create the framework before the process of data collection begins. The variables that will be measured need to be properly kept in track before the implementation will start for the sentiment analysis. The next step is to identify the dependent and independent variable of the research so that easily the relations can be figured out of the overall present variables.

Before proceeding with the dependent and independent variables it is important to find out what these variables actually mean. The independent variables are mainly to signify the case of a research and the before effect of the outcome. And the dependent variables are the effect of the independent variables. And this is the reason why independent variables are not dependent upon the entire research. But when it comes to the dependent variables they are totally dependent on the research (Karcioğlu *et al.* 2019). These are the reasons why before choosing the dependent variables it is important to look for every aspect of the research because they cannot be changed in the future. The independent variables are changeable based on the research needs. If an analysis can be more simple while changing an independent variable.

In this project the sentiment analysis will be implemented where the dependent variable and independent variables are important to find out. The independent variables in this case are the libraries and algorithms. And the dependent variable in this case will be the final models that will be presented at the end of the research (Tao *et al.* 2019). The overall research will help to find out

how the analysis of sentiment can be applied to predict the real life data and the future prediction will be achieved out of the analysis. Later in this project the methodology of the sentiment analysis will be presented.

2.4 Conclusion

The overall conclusion of the literature review will help the researcher to achieve a clearer outcome. The analysis that will be performed here is known as sentiment analysis, and in this analysis the overall social media data will be examined to extract the emotions out of it. The importance of literature review will present the overall importance of the researchers. The project will show how effectively the sentiment analysis will be performed based on the dataset that will be chosen for the research. This will show the overall effectiveness of the analysis on how the big data will be gathered and the expression will be extracted. The main objective of the project is to find out the meaning of the analysis. The social media data needs to be analyzed so that the customer prediction of business can be analyzed with accuracy. The present researches will be established with more accuracy when the past researches are well observed with more accurate sentiment analysis. The gaps are mainly to understand the obstacles that the researcher can face in the coming future. In the next research the overall method of the research will be conducted to understand how the overall analysis will be performed.

Section 03: Methodology

3.1 Introduction

The description of methodology is mainly to present the procedure where the whole research will be implemented. There are different techniques that might be adopted in the time of research so that the performance of research enhances. This part of the research will mainly show the overall outline of implementation that will be applied to achieve the desired outcome. The method of conducting the research will be shown and also what will be the design, approach and technique of research can be easily found out with the help of this presentation. The limitations and the overview will help to maintain the quality of the sentiment analysis. Here all the analysis steps will be described so that individuals can understand how easily the analysis can be followed. In

this sentiment analysis all the chosen procedures will help the researcher to receive the ultimate objective of the project. In this case the objective is mainly to predict the attacks and based on creating certain models that will be tested and trained to deal with the real information.

3.2 Method Outline

This particular research will be conducted with the help of a special method so that the appropriate goals can be achieved at the end. The first stage will be planning and this step will consist of identifying several strategies and approaches that will assist the researcher to meet the ultimate goal. The data collection, approaches, and methods will initiate the overall analysis. The collection of data will help to conduct the analysis and after that the desired outcome will be obtained from the research (Mitra A. 2020). The proper techniques will be analyzed by the researchers so that potential outcome can be achieved of the analysis.

3.3 Research Philosophy

When a researcher wants to achieve an objective then it is possible to examine different philosophies so that a strategic process is followed throughout the research. The research philosophy assists with identifying the research strategies, approaches and methods. There are multiple philosophies that might be adopted so that aim can be achieved (Gujjar *et al.* 2021). There are different philosophies like "realism, positivism, objectivism, pragmatism and constructivism". Objectivism is the philosophy that will be applied here because the project is mainly based on achieving the ultimate goals of the research. The background will show how the outcome of the research will be met so that customer analysis can be performed with the help of sentiment analysis.

3.4 Research Approach

There are multiple approaches that are present to conduct an analysis, in this case the analysis will be based upon a particular approach. There are mainly two approaches that should be kept in mind, the first one is deductive approach and the second one is the inductive approach (Devi *et al.* 2020). The deductive approaches are mainly based on the data that are already present on the internet. The analysis will be conducted based upon the existing theory. And when it comes to

inductive approaches those are mainly analyzed by the researcher. In this scenario the inductive approach is undertaken to reach the outcome of the research.

3.5 Research Design

Every significant research on a certain subject matter is associated with a proper research design. The design of a research is divided into two types such as “qualitative and quantitative research” design. Research design is an important aspect regarding the context of the research as a research design is something which helps a researcher to follow a specific way or structure for an entire project (Ahmed *et al.* 2022). The design of a research makes the research symmetric and relevant according to the topic. The first type which is the qualitative research design refers to the perception and observation-based analysis of the sources and information, when the second type of design, which is the quantitative research design, is entirely committed to give measurable and entire statistical results from the research. Basically, selecting an appropriate research design depends on the philosophy of the research. Here, in this current research, the philosophy is objectivism. This current research follows the design of quantitative design (Zahidi *et al.* 2021). Here the researcher is conducting a primary analysis and it will give some new and precise outcomes.

3.6 Research Strategy

The research strategy is known as the procedure that describes the techniques which are mainly applied by the researchers and collecting the relevant data to the research area. The project here is based on an analysis that will help to analyze a big amount of data so that they can be used to achieve an effective outcome (Nuser *et al.* 2022). When a research is being conducted there is a possibility that it includes several techniques so that it assists with the overall outcome of the research. The chosen strategy is mainly described based upon how the overall research is being conducted. If the data collection method is primary then most likely the strategy is followed using surveys. When the data collection method is secondary then the archival research is followed where the data are already present and extracting the information from the internet will help to conduct the analysis.

3.7 Research Method

There are mainly three types of methods available when a research is being conducted. They are known as mono method, multi method or mixed method. This research is mainly to understand an expression present inside a text (Ibanez *et al.* 2020). The overall analysis will be based upon analyzing the big data so that easily a point of view can be understood. The outcome of the research can be achieved using a single research method or several research methods. In this case the research method that will be followed is known as mono method. One single method will show how effectively the analysis will be performed and how the social media data can be analyzed. This will show the overall method of the research that will be followed.

3.8 Data Collection Method

Data collection method mainly refers to the procedure in which the information will be collected. The researcher is conducting a research on the detection of the frauds of credit cards and the data for this research is collected through secondary analysis (Al-Shabi M.A. 2020). The researcher has reviewed several journals, texts, papers and other internet sources on this topic which gave a relevant idea to the topic. The data collection method is very significant for the entire conduction of the research. In this research, the researcher has performed a secondary analysis through which the data for this research is gathered. There are mainly two kinds of data collection methods such as the primary collection method and secondary collection method. In the primary collection method process, the researcher operates a research while the outcomes that the project provides are entirely new and raw. But in the secondary data collection, the data with which the researcher works is already existing information (Saifullah *et al.* 2021). Though the secondary data collection method does not refer to only copying blindly from the past research but to take inspiration or ideas from those works and do the current project with efficiency.

3.9 Research Ethics

Several ethics are associated with a research while a researcher conducts a research. Numerous steps are involved in the project which have to follow the values and standards of operating a research (Obeid *et al.* 2020). This research has maintained all the research ethics as a project requires to get fulfilled ethically. The data that the researcher has gathered are genuine and the

researcher has mentioned all the names of the authors as well who has performed the research. Each of the participants or the authors have the consent on the basis of which the data has been collected for the research. While maintaining all the ethics of the research it is possible to enhance the quality of the analysis and apply them on real life data.

3.10 Research Limitations

The limitation of research is mainly based upon the disadvantages of research that might restrict the research from being successful. The limitations are mainly to understand how the correct objective will be achieved while following the procedure of research (Konstantinov *et al.* 2020). The maximum amount of limitations are based upon how far the research will illustrate and in which level it will stop. As this overall implementation will be conducted based upon software analysis it is important to properly upgrade the technologies that will be followed in the time of analysis. Also in the time of data collection the overall information that will be gathered is based upon a proper selection of variables and relevancy. If the data is not relevant then the outcome of the research will not be successful.

3.11 Conclusion

The overall analysis will be conducted based upon the methodology that is mentioned here and how the project. The overall methodology starts from the steps of collecting the data and following a particular method and research design so that quality of the research can be increased to the next level. In this case the data will be related to credit card attacks and how the time and frequency of the attack will be related to data that will help to conduct the analysis of sentiment. These overall procedures and techniques of achieving the outcome will provide the researcher with confidence so that they can easily achieve the deliverables at the end of the project.

Section 04: Finding / Result

```
import numpy as np # linear algebra
import pandas as pd # data processing
pd.options.mode.chained_assignment = None
import os #File location
for dirname, _, filenames in os.walk('/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

from wordcloud import WordCloud #Word visualization
import matplotlib.pyplot as plt #Plotting properties
import seaborn as sns #Plotting properties
from sklearn.feature_extraction.text import CountVectorizer #Data transformation
from sklearn.model_selection import train_test_split #Data testing
from sklearn.linear_model import LogisticRegression #Prediction Model
from sklearn.metrics import accuracy_score
import re #Regular expressions
import nltk
from nltk import word_tokenize
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

Fig 1: Importing Libraries

(Source: self-created in Google colab)

This is the beginning of the sentiment analysis where the crucial libraries will be imported. This imported library will help the further analysis to execute in the future.

```
#Validation dataset
val=pd.read_csv("twitter_validation.csv", header=None)
#Full dataset for Train-Test
train=pd.read_csv("twitter_training.csv", header=None)
```

```
train.columns=['id','information','type','text']
train.head()
```

	id	information	type	text
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...

Fig 2: Importing Datasets

(Source: self-created in Google colab)

After the importing the libraries, the next step will be about the data that will assist with performing the sentiment analysis. Here two datasets will be imported and they are known as twitter training data and validation data.

```
val.columns=['id','information','type','text']
val.head()
```

	id	information	type	text
0	3364	Facebook	Irrelevant	I mentioned on Facebook that I was struggling ...
1	352	Amazon	Neutral	BBC News - Amazon boss Jeff Bezos rejects clai...
2	8312	Microsoft	Negative	@Microsoft Why do I pay for WORD when it funct...
3	4371	CS-GO	Negative	CSGO matchmaking is so full of closet hacking,...
4	4433	Google	Neutral	Now the President is slapping Americans in the...

```
train_data=train #[(train["type"] == "Positive") | (train["type"] == "Negative")]
train_data
```

	id	information	type	text
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...

Fig 3: Creating train data

(Source: self-created in Google colab)

Showing the validation data and the present variables that will be present inside the data set. This will show all the present variables. The train data will be created here for further visualization of the data.

```
val_data=val #[(val["type"] == "Positive") | (val["type"] == "Negative")]
val_data
```

	id	information	type	text
0	3364	Facebook	Irrelevant	I mentioned on Facebook that I was struggling ...
1	352	Amazon	Neutral	BBC News - Amazon boss Jeff Bezos rejects clai...
2	8312	Microsoft	Negative	@Microsoft Why do I pay for WORD when it funct...
3	4371	CS-GO	Negative	CSGO matchmaking is so full of closet hacking,...
4	4433	Google	Neutral	Now the President is slapping Americans in the...
...	...	...	...	...
995	4891	GrandTheftAuto(GTA)	Irrelevant	☆ Toronto is the arts and culture capital of ...
996	4359	CS-GO	Irrelevant	tHIS IS ACTUALLY A GOOD MOVE TOT BRING MORE VI...
997	2652	Borderlands	Positive	Today sucked so it's time to drink wine n play...
998	8069	Microsoft	Positive	Bought a fraction of Microsoft today. Small wins.
999	6960	johnson&johnson	Neutral	Johnson & Johnson to stop selling talc baby po...

1000 rows × 4 columns

Fig 4: Creating validation data

(Source: self-created in Google colab)

The validation data will be present here and the present variables will be shown how the overall shape of the validation data.

```
#Text transformation
train_data["lower"]=train_data.text.str.lower() #lowercase
train_data["lower"]=[str(data) for data in train_data.lower] #converting all to string
train_data["lower"]=train_data.lower.apply(lambda x: re.sub('[^A-Za-z0-9 ]+', ' ', x)) #regex
val_data["lower"]=val_data.text.str.lower() #lowercase
val_data["lower"]=[str(data) for data in val_data.lower] #converting all to string
val_data["lower"]=val_data.lower.apply(lambda x: re.sub('[^A-Za-z0-9 ]+', ' ', x)) #regex
```

```
train_data.head()
```

	id	information	type	text	lower
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...	i am coming to the borders and i will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...	im getting on borderlands 2 and i will murder ...

Fig 5: Viewing train data

(Source: self-created in Google colab)

Here the training data will be presented and the transformation of the text will show how the lowercase and string value present inside the data.

```
word_cloud_text = ''.join(train_data[train_data["type"]=="Positive"].lower)
#Creation of wordcloud
wordcloud = WordCloud(
    max_font_size=100,
    max_words=100,
    background_color="black",
    scale=10,
    width=800,
    height=800
).generate(word_cloud_text)
#Figure properties
plt.figure(figsize=(10,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```


Fig 6: Word cloud creation with positive data type

(Source: self-created in Google colab)

This will be the first presentation of word cloud where the expression will be present as positive. The most used word will be shown in the output of the category positive.

Fig 7: Output of Word cloud

(Source: self-created in Google colab)

The output will show the most used word that is present in the data, here the word “Game” is shown as the most used word.

```
word_cloud_text = ''.join(train_data[train_data["type"]=="Negative"].lower)
#Creation of wordcloud
wordcloud = WordCloud(
    max_font_size=100,
    max_words=100,
    background_color="black",
    scale=10,
    width=800,
    height=800
).generate(word_cloud_text)
#Figure properties
plt.figure(figsize=(10,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

Fig 8: Word cloud creation with negative data type

(Source: self-created in Google colab)

Here the presentation of the data type negative will be shown and in this case the word cloud will also show the most used word under the category negativity.

Fig 9: Output of Word cloud

(Source: self-created in Google colab)

The negative data type that will be shown here is presented as twitter and game as well.

```
word_cloud_text = ''.join(train_data[train_data["type"]=="Irrelevant"].lower)
#Creation of wordcloud
wordcloud = WordCloud(
    max_font_size=100,
    max_words=100,
    background_color="black",
    scale=10,
    width=800,
    height=800
).generate(word_cloud_text)
#Figure properties
plt.figure(figsize=(10,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

Fig 10: Word cloud creation with irrelevant data type

(Source: self-created in Google colab)

The last word cloud will be presented as the data type of irrelevant where the most used words will be shown that comes under this particular category.

Fig 11: Output of Word cloud

(Source: self-created in Google colab)

People, love, player these are the most used words those specialty comes under this particular category of word cloud.

```
#Count information per category
plot1=train.groupby(by=["information","type"]).count().reset_index()
plot1.head()
```

	information	type	id	text	lower
0	Amazon	Irrelevant	192	186	192
1	Amazon	Negative	576	575	576
2	Amazon	Neutral	1236	1207	1236
3	Amazon	Positive	312	308	312
4	ApexLegends	Irrelevant	192	192	192

Fig 12: Counting data per category

(Source: self-created in Google colab)

Here the overall categories are presented with their information, type, id, text and lower. There is all the information present with the help of different categories.

```
#Figure of comparison per branch
plt.figure(figsize=(20,6))
sns.barplot(data=plot1,x="information",y="id",hue="type")
plt.xticks(rotation=90)
plt.xlabel("Brand")
plt.ylabel("Number of tweets")
plt.grid()
plt.title("Distribution of tweets per Branch and Type");
```

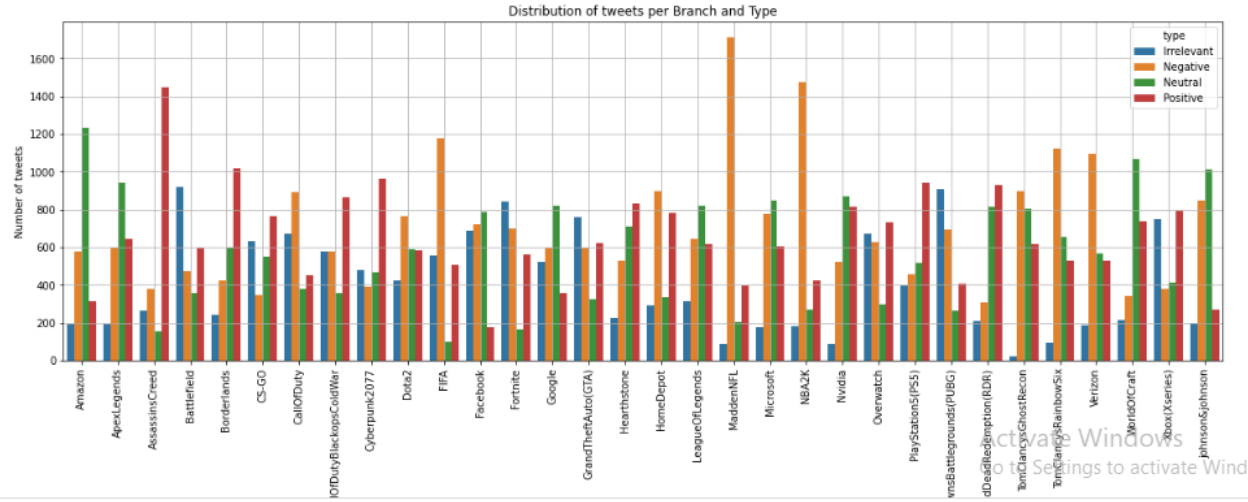

Fig 13: Comparison of branches

(Source: self-created in Google colab)

Here the overall presentation of different branches will be shown. All the different branches are shown with the help of different colors and the bar plot will show the most used and less used branch as well.

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

```
#Text splitting
tokens_text = [word_tokenize(str(word)) for word in train_data.lower]
#Unique word counter
tokens_counter = [item for sublist in tokens_text for item in sublist]
print("Number of tokens: ", len(set(tokens_counter)))
```

```
Number of tokens:  30436
```

Fig 14: Analyzing the texts

(Source: self-created in Google colab)

The overall texts will be analyzed here, and for that the text splitting will be done and the unique words will be counted as token texts. The number of token is presented here as 30436.

```
tokens_text[1]
```

```
['i',
 'am',
 'coming',
 'to',
 'the',
 'borders',
 'and',
 'i',
 'will',
 'kill',
 'you',
 'all']
```

, the main English stopwords were saved on an additional variable, to be used in the following modeling.

```
#Choosing english stopwords
stopwords_nltk = nltk.corpus.stopwords
stop_words = stopwords_nltk.words('english')
stop_words[:5]
```

```
['i', 'me', 'my', 'myself', 'we']
```

Fig 15: Choosing the English stop words

(Source: self-created in Google colab)

The texts that are present inside the tokens will be shown in the first place and in the second phase the English words will be shown like I, my, me etc.

```
#Initial Bag of Words
bow_counts = CountVectorizer(
    tokenizer=word_tokenize,
    stop_words=stop_words, #English Stopwords
    ngram_range=(1, 1) #analysis of one word
)
```

```
#Train - Test splitting
reviews_train, reviews_test = train_test_split(train_data, test_size=0.2, random_state=0)
```

```
#Creation of encoding related to train dataset
X_train_bow = bow_counts.fit_transform(reviews_train.lower)
#Transformation of test dataset with train encoding
X_test_bow = bow_counts.transform(reviews_test.lower)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWarning: Your stop_words may be inconsiste
  % sorted(inconsistent)
```

```
X_test_bow
```

```
<14937x28993 sparse matrix of type '<class 'numpy.int64'>'
        with 161222 stored elements in Compressed Sparse Row format>
```

Fig 16: Splitting data for logistic regression model

(Source: self-created in Google colab)

The overall data will be split into two parts the first one is known as train and the second one is known as test. The train dataset will be created here and along with that the encoding of training will be performed as well.

```
#Labels for train and test encoding
y_train_bow = reviews_train['type']
y_test_bow = reviews_test['type']
```

```
#Total of registers per category
y_test_bow.value_counts() / y_test_bow.shape[0]
```

```
Negative      0.299190
Positive      0.282252
Neutral       0.245632
Irrelevant    0.172926
Name: type, dtype: float64
```

```
# Logistic regression
model1 = LogisticRegression(C=1, solver="liblinear",max_iter=200)
model1.fit(X_train_bow, y_train_bow)
# Prediction
test_pred = model1.predict(X_test_bow)
print("Accuracy: ", accuracy_score(y_test_bow, test_pred) * 100)
```

```
Accuracy:  81.52239405503113
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning: Liblinear failed to converge, increase the number of iterations.
  ConvergenceWarning,
```

Fig 17: Prediction accuracy of logistic regression model

(Source: self-created in Google colab)

The preparation of the model creation will begin here, with the help of splitting data the model will start to begin. The model that will be created here is logistic regression, and the prediction value will be presented as 81.52.

```
#Validation data
X_val_bow = bow_counts.transform(val_data.lower)
y_val_bow = val_data['type']
```

```
X_val_bow
```

```
<1000x28993 sparse matrix of type '<class 'numpy.int64'>'
        with 12913 stored elements in Compressed Sparse Row format>
```

```
Val_res = model1.predict(X_val_bow)
print("Accuracy: ", accuracy_score(y_val_bow, Val_res) * 100)
```

```
Accuracy:  91.7
```

Fig 18: Accuracy of Validation data

(Source: self-created in Google colab)

The validation data will be created based on the training data and also the validation data. The accuracy value will be presented as 91.7.

```
#n-gram of 4 words
bow_counts = CountVectorizer(
    tokenizer=word_tokenize,
    ngram_range=(1,4)
)
#Data labeling
X_train_bow = bow_counts.fit_transform(reviews_train.lower)
X_test_bow = bow_counts.transform(reviews_test.lower)
X_val_bow = bow_counts.transform(val_data.lower)
```

```
X_train_bow
```

```
<59745x1427378 sparse matrix of type '<class 'numpy.int64'>'
        with 4142213 stored elements in Compressed Sparse Row format>
```

Fig 19: Data labeling for model creation

(Source: self-created in Google colab)

Data labeling is known as labeling the train data and also test and validation data. The type of the elements will show how the row format will be finalized.

```
model2 = LogisticRegression(C=0.9, solver="liblinear",max_iter=200)
# Logistic regression
model2.fit(X_train_bow, y_train_bow)
# Prediction
test_pred_2 = model2.predict(X_test_bow)
print("Accuracy: ", accuracy_score(y_test_bow, test_pred_2) * 100)
```

```
Accuracy:  89.90426457789383
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning: Liblinear failed to converge, increase the number of iterations.
  ConvergenceWarning,
```

```
y_val_bow = val_data['type']
Val_pred_2 = model2.predict(X_val_bow)
print("Accuracy: ", accuracy_score(y_val_bow, Val_pred_2) * 100)
```

```
Accuracy:  98.2
```

Fig 20: Final Model accuracy value

(Source: self-created in Google colab)

The final model will be presented here and the logistic regression model will be presented with the help of an accuracy value that is presented as 98.1. This is the most accurate value and that will show how the perfection of the model is.

Section 05: Analysis / Discussion

5.1 Introduction

The project will be based on the sentiment analysis of big data, and the data will be collected from social media. Twitter is the source of the data where people will post about their opinions and that will be analyzed so that an overall analysis can be performed. The sentiment analysis is all about understanding the expression of a sentence, when it is being analyzed by a human they will be able to understand each emotion. When it comes to a machine it is difficult for a machine to understand every emotion hidden inside the sentence. Machine learning will help the sentence so that an emotion can be extracted out of them easily. In this project an analysis is implemented with the help of data that is extracted from the real life platform. The data is divided into validation and training data so that they can be easily applied to execute the analysis. Here with the help of many data preparation techniques the data will be cleaned so that the analysis can be applied with more appropriateness. In the previous section the overall analysis is being presented with relevant

screenshots of algorithm that were implemented. In this chapter the overall analysis process of the sentiment will be explained so that in the future the process becomes easier to handle.

5.2 Analysis

The analysis will first start with importing all the necessary libraries that will help the algorithms and different techniques to be implemented in the future (Khan *et al.* 2020). The library that will be implemented at the starting is numpy and it is used for all the linear algebra that will be applied in the future implementation. For processing the data that will be implemented, a library named pandas will be imported. Next the word cloud library will be imported and it will assist with the overall visualization of the data. For presenting the plots, the library called matplotlib will be imported (Biswas *et al.* 2020). The sklearn library will be imported for data transformation, the prediction model that will be created at the final phase will be known as logistic regression. The accuracy of the model will show how accurate the prediction will be at the end of the sentiment analysis. In the next phase the data will be imported and this will show the raw data that will be imported for performing the analysis of sentiment. Here the twitter data is divided into two categories and they are known as training data and validation data. The train and validation data both will be shown in the starting so that all the columns can be presented along with their respective columns. The train data will be called as train and the validation data will be presented as validation (Zahidi *et al.* 2021). After this the transformation of texts will be shown where train data and validation data both will be called with a new column known as lower. The next section will be to present the overall data visualization where different word clouds will be presented as data visualization. As this analysis is known as sentiment analysis here mainly the words will be presented as visualization. The word cloud is known as presenting the most effective worlds. The size of the words will provide an overall effect of the words that will be present inside the word cloud.

The first word cloud will show the positive expressions that will be present inside the data and the background color will be shown as black. And the different vibrant colors will show the words and their effect (Almutiry *et al.* 2021). The size of the words will show how they are differently creating an effect and the most used positive sentences will be presented as the bigger size. The other two expressions that will be presented as the word cloud are known as neutral, negative and

irrelevant. This will present the same visualization of data so that the most affected words can be extracted out of the data. When data is being analyzed it is very difficult to find out the most used words or expressions out of them (Badgaiyya *et al.* 2021). But with the help of data visualization and sentiment analysis the process becomes earlier to execute. After this overall visualization the next step will be presenting the overall count of the information that is present inside the plot. The different expressions and their types are presented while counting the overall information. This will provide a bar plot where all the different information will be presented based on their category of emotions. Next the text analysis will be performed where the different expressions of the text will be shown. First the text will be divided so that the token can be found easily out of them. The unique words that will be found on the overall data will be shown here and they will be presented as token numbers. The total number of tokens present inside the analysis are 30436. The text will also be found and this will show how effectively the unique words can be distinguished from the overall analysis. Total of five token words will be shown in the result of the analysis and this will show how easily the texts can be differentiated from the overall analysis (Cheng *et al.* 2019). Next the overall model creation will begin where the selected model in this case is known as logistic regression model. For this the first step will be to count the vector tokens so that easily the words can be analyzed and the English stop words will be selected as the words upon which the model will be created.

For model preparation the first step is to split the data into two sections and they are known as training data and testing data. The overall elements of the test data and training data will be shown in the process and their data type will also be presented so that the present elements can be found easily (Saura *et al.* 2019). The review type will be shown as train data and test data and with the help of them the overall model will be trained. The count and shape of the train data will be present in the next step and the overall type of the data expressions will be shown here so that easily the count and shape can be found. Next comes the final model presentation where the prediction value will be shown at the first step and the accuracy value will be presented as 81.52 so that test prediction can be found simply. Next the validation data will help to build the model where another accuracy score will be presented and the accuracy value will be 91.7. Next comes the preparation of the final model where the data labeling will be the first step and this will present three data labels known as train, test and validation the accuracy value of the test prediction will be 89.9 and

the y validation data will present the accuracy value of 98.2. This will finally prepare the final model. This overall analysis will help to present the final model with the best accuracy score so that the analysis can be performed easily (Saha *et al.* 2021). So that the type of the expression can be found with accuracy. This will help to present the overall sentiment analysis so that easily the expression can be found while training the model of logistic regression.

5.3 Discussion

The logistic model will show how the overall model building will be done while implementing the analysis of sentiment. The data that will be analyzed here is known as the twitter data. Different platforms and their perspectives are present inside the data and this will create an estimate impression on how the data will be analyzed (Sodhar *et al.* 2020). When it comes to analyzing the big data through eyes it is possible to understand if their expression is positive, negative or irrelevant even if it can be neutral. This can be easily analyzed if it is being performed by a human brain when it comes to a machine on how they will understand an emotion hidden within a sentence then it is not possible for it to understand the expression. This will show the overall need for model creation in this project and why a machine should be trained so that it can perform a better prediction. This will show the overall model accuracy so that it can generate a better outcome at the end. Using the methodology of trend analysis, someone may examine a text to ascertain the emotion it represents. This is accomplished by combining artificial intelligence and text analysis. A software can determine if a text's emotion is good, unfavorable, or neutral utilizing basic text analytics. It’s a potent machine intelligence method with significant business implications. One may use text analytics as a significant instrument to handle issues ranging from important aspects of any business to evaluation of a company (Madhu S. 2018). In order to help organizations run more effectively, new technologies are being developed surrounding sentiment classification. Organizations deal with a number of issues that relate to sentiment classification problems so as to achieve sentiment analysis correctness. NLP can make it challenging to evaluate sentiments or feelings since robots need to be educated to do so in the same way that the nervous system works. This comes in contrast to being aware of the subtleties of other tongues. Technology for sentiment classification is capable of addressing these problems as machine learning develops. These are the primary difficulties with sentiment analysis. If a text document includes two opposing words, an emotion analyzer should cope with contradictory punctuations. Recognized identification is a key

difficulty that computers run across. Varied settings give words multiple interpretations. Every speech spoken on the internet, this so happened, develops a distinct identity. The efficiency of communication and the use of the Web as a platform give rise to errors in spelling, punctuation, slang symbols, and punctuation (Moshkin *et al.* 2019). Uni polarity is the possibility of a text having multiple competing concepts, which might be very difficult. Customers could, for example, laud a certain item for having good attributes. On the other hand, people could also condemn it for its unappealing characteristics. Since the customer's comment can have some gaps in it, it would be challenging to do a thorough text analysis in this situation. The findings of the trend analysis could not accurately reflect the viewpoint expressed in the content and could even be deceptive (Rawat *et al.* 2021). Anyone may use denial to flip the polarization of words, adjectives, and paragraphs in everyday conversation, which is the complete antithesis of precise manner. To determine if there is a denial, many criteria can be applied. Additionally, users must be conscious of the variety of terms that might affect a denial. The majority of emotion analytical techniques use a list of phrases that contain a signal of whitespace to identify negativity. Nevertheless, the denial might vary upon the design.

Facetiousness is difficult to spot since it frequently utilizes due to its variety (Shadadi *et al.* 2022). Except the technology is made to accommodate for the likelihood of sarcastically, expression methodologies struggle to detect such connotations. Inside the discussions area of social networking websites like Twitter, Linkedin, Pinterest, and others, mockery frequently emerges. One must be knowledgeable of discourse's framework, the issue being discussed, and the setting that it is being used in in order to recognize them. Facetiousness is challenging to spot in individuals as well as in attitude advanced analytics (Gobithaasan *et al.* 2020). One method to build sarcastic utterances, it is challenging to teach an algorithm for sentiment analysis that reads sarcasm. Because phrases may have varied meanings depending on the environment in which they are spoken, emotion recognition is a challenging area of research. As a result, it might be difficult to determine how people are feeling since emotion techniques cannot determine the circumstances in which certain terms are utilized (Hanswal *et al.* 2021). Nevertheless, it could be possible to do so more easily and reliably in the future with enhanced and intelligent expression analysis algorithms. This will be the overall implementation of the sentiment analysis. This project will show how the effective implementation of sentiment analysis so that in the real world can be

applied easily. This project will help all the analysts so that they can easily train the machine to understand the emotions hiding behind them. Here in this project the main aim is to analyze the overall expressions that are present inside a human expression. This will help the researchers to guide in the ultimate direction so that they can easily detect the emotion hiding behind a review (Oyebode *et al.* 2019). Here the overall practical analysis is seen so that researchers can receive the effective use of sentiment analysis. Whether an effective technique is applied or not it is crucial to develop certain approaches so that the analysis can be done. In this case the overall project is being implemented using Python so that easily the analysis can be achieved. In this case the twitter data is being applied for understanding the emotions of it. The project will illustrate the overall application of machine learning. The python language and its practical use will be shown as well. This will help the overall analysis to align in the proper way so that easily the project can deliver a meaningful outcome.

5.4 Conclusion

Any business or movement that uses public opinion to succeed, however that accomplishment is measured, can benefit from using emotion analysis as a strategy. Millions of individuals are active in analyzing and critiquing enterprises, enterprises, and organizations on media platforms, bloggers, and discussion boards. But those viewpoints are "recognized" and evaluated. By employing systems that not only track down all references of company goods, activities, or company, but also identify the sentiments and sentiments behind phrases spoken, others being talked are availing utilization of this vast volume of information. Organizations may respond and act appropriately by understanding the dialogues and debates being would have about the company thanks to the outcomes of sentiment classification. They are able to immediately spot any unfavorable feelings stated and transform bad client experiences into excellent ones. They could provide superior products that may tailor the advertising message people distribute in response towards the comments made by the company market segment or clients. Municipal federal agencies may learn how the community feels about them and whatever offerings they offer by reviewing and understanding remarks on Instagram and Twitter. Authorities can then utilize the information to enhance services like storage and recreation areas, neighborhood. Institutions can employ emotion analysis to look at undergraduate responses and feedback through studies they have conducted or from internet sources like any social platform. The data may then be used to

pinpoint any locations of teacher unhappiness and fix them, in addition to pinpoint and expand upon those regions in which people are exhibiting favorable feelings.

Section 06: Conclusion / Recommendation

6.1 Introduction

Machine learning is known as the most appropriate approach of ML algorithms and they are categorized in a certain way so that easily the statistical model can be built in an appropriate way. The implementation will show how the vector space can be transformed into a space. The model will be trained so that the prediction can be generated in the correct way. There are some advantages and disadvantages that machine learning sentiment analysis consumes. First the advantages will be discussed, these algorithms will be prepared in a certain way so that negation, irony or emotion can be detected easily. The algorithms will be created so that an effective word implementation can be shown and for this purpose a data will be required that is determined from the previous time. This particular analysis is faster than all the other methods known as traditional. The methods known as automated offers an outcome that is appropriate. This will show the advantages of using sentiment analysis with ML techniques. The prediction can be performed when better data is provided to the analysis. The disadvantages of this machine learning process is mainly related to the dataset that is chosen for proceeding further. Receiving the perfect classification is quite difficult when the overall data is huge. The small data also provides difficulty presenting ultimate visualization. When the data consists of many emojis or noise then they become difficult to analyze. Extracting the sentiment of the sentences also brings many troubles. When the traditional rules are followed then the process becomes cheap but when it comes to a ML process then it becomes expensive and this is the only reason why people do not tend to use the ML approach.

6.2 Recommendation

The sentiment analysis mainly assists with improving all kinds of recommendations so that they can easily be generated using the analysis of sentiment. This will show the overall effect of the sentiment analysis so that it can be used in the real industry. In this case the data that will be used

is based on the analysis of twitter (Elbagir *et al.* 2019). There people are talking about their many emotions so that easily the sentiment can be understood out of the sentences. This will show how the industry can help to boost the analysis. Reviewing the overall twitter comment will show how effectively the analysis can be performed using ML techniques. Here in this project the performance can be increased when a different model is used. In this particular analysis the model that will be chosen is known as the logistic regression model. This model will show how the common problems can be tackled using sentiment analysis. In this project the overall comments will illustrate the understanding of machine learning techniques so that the prediction can be done appropriately. Here the model will be prepared with proper validation and training data so that it can predict the sentiment of the analysis easily. This will illustrate the overall knowledge on how it can perform the overall procedure with appropriateness (Mansoor *et al.* 2020). A significant amount of data is created as well as being made available to online consumers thanks to the development and expansion of internet-based technologies. The web has developed into a forum for online education, open communication, and viewpoint exchange. Services for social media like Twitter are rapidly becoming more famous as a result of the ability for users to communicate their viewpoints on many subjects, engage in conversation with various groups, and spread the information globally. The study of emotion in Two datasets has received much attention. This study primarily concentrates on sentiment classification of different datasets, which is useful for analyzing data from tweets when sentiments are extremely unorganized, varied, and either negatively or positively (Sharma *et al.* 2018). The strategies for recommender systems that are currently in use, such as etymological roots strategies and computer vision, are surveyed, compared, and evaluated in this work along with assessment measures. They present analysis on twitter streaming data utilizing several machine learning techniques as Binary Classification, Max Probability, and SVM Classifiers. Additionally, they covered the basic difficulties and uses of sentiment classification using Twitter

.

6.3 Future Work

Future research will examine even more importance of proper analysis techniques, such as language models, sentiment analysis, and other segmentation. In this research, they present a research and simple comparison of the methodologies for information extraction, such as vocabulary and computer vision algorithms, merge and bridge methodologies, but some

assessment measures (Shelar *et al.* 2018). The outcome of the study revealed that vocabulary approaches are sometimes highly successful and take little work in sentient documents, whereas ML algorithms, like linear Regression, have the greatest accuracy and could be considered as benchmark instructional strategies. They also looked at how different characteristics affected classifiers. One can draw the conclusion that further reliable findings can be attained with better information. Greater sentiment is produced by using the word embedding model (Reyes *et al.* 2018). In effort to expand the precision of the classifier or the ability to adjust to a broad range of industries and dialects, one can concentrate on the research of mixing procedure of machine learning with opinions vocabulary approach.

6.4 Conclusion

The outcome displayed the findings from our trend analysis using Twitter. It demonstrated an increase in overall of above 4% across two text categorizations: a dichotomous, affirmative against unfavorable and a triple bond, optimistic compared to negative compared neutral. They start with a conventional numerical government word embedding system as the foundation. On explicitly annotations that represent a subset of a statistical population of a sequence of tweets, they provided an extensive collection of tests and objectives. Researchers looked at two different types of frameworks: functionality plus tree kernel features, and they show models that perform better than the word embedding background. Researchers do component evaluation for their functionality method, and the results show that the greatest good insight are those that integrate the previous polarity of terms with their components identifiers. They make the speculative conclusion that emotion recognition for social media datasets does not differentiate all that much from sentiment classification for content from all other categories.

Reference List

Journals

Ahmed, Mohammed Emtiaz, Md Rafiqul Islam Rabin, and Farah Naz Chowdhury. "COVID-19: Social media sentiment analysis on reopening." *arXiv preprint arXiv:2006.00804* (2020).

Almutiry, S. and Abdel Fattah, M., 2021. Arabic CyberBullying Detection Using Arabic Sentiment Analysis. *The Egyptian Journal of Language Engineering*, *8*(1), pp.39-50.

Al-Shabi, M.A., 2020. Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS*, *20*(1), p.1.

Anand, T., Singh, V., Bali, B., Sahoo, B.M., Shivhare, B.D. and Gupta, A.D., 2020, June. Survey paper: sentiment analysis for major government decisions. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 104-109). IEEE.

Badgaiyya, A., Shankarpale, P., Wankhade, R., Shetye, U., Gholap, K. and Pande, S., 2021. An Application of Sentiment Analysis Based on Hybrid Database of Movie Ratings. *International Research Journal of Engineering and Technology (IRJET)*, *8*, pp.655-665.

Basarslan, M.S. and Kayaalp, F., 2020. Sentiment analysis with machine learning methods on social media.

Bhavsar, H. and Manglani, R., 2019. Sentiment analysis of Twitter data using Python. *International Research Journal of Engineering and Technology (IRJET)*, *6*(3), pp.510-511.

Biswas, S., Sarkar, I., Das, P., Bose, R. and Roy, S., 2020. Examining the effects of pandemics on stock market trends through sentiment analysis. *J Xidian Univ*, *14*(6), pp.1163-76.

Chandio, M.M. and Sah, M., 2019, October. Brexit twitter sentiment analysis: Changing opinions about brexit and uk politicians. In *International Conference on Information, Communication and Computing Technology* (pp. 1-11). Springer, Cham.

Cheng, L.C. and Tsai, S.L., 2019, August. Deep learning for automated sentiment analysis of social media. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 1001-1004).

Devi, G.D. and Kamalakkannan, S., 2020. Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications. *Test Eng. Manag*, *83*(7), pp.2466-2474.

Diyasa, I.G.S.M., Mandenni, N.M.I.M., Fachrurrozi, M.I., Pradika, S.I., Manab, K.R.N. and Sasmita, N.R., 2021, May. Twitter Sentiment Analysis as an Evaluation and Service Base On

Python Textblob. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1125, No. 1, p. 012034). IOP Publishing.

Elbagir, S. and Yang, J., 2019, March. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 122, p. 16).

Elbagir, S. and Yang, J., 2019, March. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 122, p. 16).

Fitri, V.A., Andreswari, R. and Hasibuan, M.A., 2019. Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, *161*, pp.765-772.

Gobithaasan, R.U. and HAMID, N.F.S.C., 2020. Sentiment Analysis of People'S Acceptance Towards the New Malaysian Government Using Naïve Bayes Method. *Universiti Malaysia Terengganu Journal of Undergraduate Research*, *2*(3), pp.93-102.

Gujjar, J.P. and HR, P.K., 2021. Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends*, *7*(2), pp.1097-1099.

Gunawan, T.S., Abdullah, N.A.J., Kartiwi, M. and Ihsanto, E., 2020, October. Social network analysis using python data mining. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-6). IEEE.

Habibi, M.N., 2019. Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis. *International Journal of Modern Education & Computer Science*, *11*(11).

Hanswal, S.S., Pareek, A., Vyas, G. and Sharma, A., 2021. Sentiment Analysis on E-Learning Using Machine Learning Classifiers in Python. In *Rising Threats in Expert Applications and Solutions* (pp. 1-8). Springer, Singapore.

Ibanez, M.M., Rosa, R.R. and Guimaraes, L.N., 2020. Sentiment Analysis Applied to Analyze Society's Emotion in Two Different Context in Social Media. *Inteligencia Artificial*, *23*(66), pp.66-84.

Karcioğlu, A.A. and Aydin, T., 2019, April. Sentiment analysis of Turkish and english twitter feeds using Word2Vec model. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A. and Mittal, A., 2020. Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. *Critical Rev*, *7*(9), pp.2761-2774.

Konstantinov, A., Moshkin, V. and Yarushkina, N., 2020, December. Approach to the use of language models BERT and Word2vec in sentiment analysis of social network texts. In *International Scientific and Practical Conference in Control Engineering and Decision Making* (pp. 462-473). Springer, Cham.

Madhu, S., 2018. An approach to analyze suicidal tendency in blogs and tweets using Sentiment Analysis. *Int. J. Sci. Res. Comput. Sci. Eng*, *6*(4), pp.34-36.

Maheswari, S.U. and Dhenakaran, S.S., 2019. Sentiment analysis on social media big data with multiple tweet words. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN*, pp.2278-3075.

Manguri, K.H., Ramadhan, R.N. and Amin, P.R.M., 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research*, pp.54-65.

Mansoor, M., Gurumurthy, K. and Prasad, V.R., 2020. Global sentiment analysis of COVID-19 tweets over time. *arXiv preprint arXiv:2010.14234*.

Mitra, A., 2020. Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies (UCCT), 2*(03), pp.145-152.

Moshkin, V., Yarushkina, N. and Andreev, I., 2019, October. The sentiment analysis of unstructured social network data using the extended ontology SentiWordNet. In *2019 12th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 576-580). IEEE.

Nti, I.K., Adekoya, A.F. and Weyori, B.A., 2020. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. *Appl. Comput. Syst.*, *25*(1), pp.33-42.

Nuser, M.A.R.Y.A.M., Alsukhni, E.M.A.D., Saifan, A.H.M.A.D., Khasawneh, R.A.M.A. and Ukkaz, D.I.N.A., 2022. Sentiment analysis of COVID-19 vaccine with deep learning. *J Theor Appl Inf Technol*, *100*(12), pp.4513-4521.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A. and Habash, N., 2020, May. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference* (pp. 7022-7032).

Oyebode, O. and Orji, R., 2019, October. Social media and sentiment analysis: the Nigeria presidential election 2019. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0140-0146). IEEE.

Pérez, J.M., Giudici, J.C. and Luque, F., 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*.

Pokharel, B.P., 2020. Twitter sentiment analysis during covid-19 outbreak in nepal. *Available at SSRN 3624719*.

Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R.N. and Ghosh, A., 2021. Sentiment analysis at online social network for cyber-malicious post reviews using machine learning techniques. In *Computationally intelligent systems and their applications* (pp. 113-130). Springer, Singapore.

Reyes-Menendez, A., Saura, J.R. and Alvarez-Alonso, C., 2018. Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. *International journal of environmental research and public health*, *15*(11), p.2537.

Saha, A., Al Marouf, A. and Hossain, R., 2021, June. Sentiment analysis from depression-related user-generated contents from social media. In *2021 8th International Conference on Computer and Communication Engineering (ICCCE)* (pp. 259-264). IEEE.

Saifullah, S., Fauziah, Y. and Aribowo, A.S., 2021. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *arXiv preprint arXiv:2101.06353*.

Saura, J.R., Palos-Sanchez, P. and Grilo, A., 2019. Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*, *11*(3), p.917.

Shadadi, E., Kouser, S., Alamer, L. and Whig, P., 2022. Novel approach of Predicting Human Sentiment using Deep Learning. *Journal of Computer Science and Engineering (JCSE)*, *3*(2), pp.107-119.

Sharma, N., Pabreja, R., Yaqub, U., Atluri, V., Chun, S.A. and Vaidya, J., 2018, May. Web-based application for sentiment analysis of live tweets. In *Proceedings of the 19th Annual International Conference on Digital government research: Governance in the data Age* (pp. 1-2).

Sharma, S. and Bansal, M., 2020. Real-Time Sentiment Analysis Towards Machine Learning. *International Journal of Scientific & Technology Research*, *9*(2).

Shekhawat, B.S., 2019. *Sentiment Classification of Current Public Opinion on BREXIT: Naïve Bayes Classifier Model vs Python's TextBlob Approach* (Doctoral dissertation, Dublin, National College of Ireland).

Shelar, A. and Huang, C.Y., 2018, December. Sentiment analysis of twitter data. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1301-1302). IEEE.

Siddharth, S., Darsini, R. and Sujithra, M., 2018. Sentiment analysis on twitter data using machine learning algorithms in python. *Int. J. Eng. Res. Comput. Sci. Eng*, *5*(2), pp.285-290.

Singh, G., 2020. Decision Tree J48 at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text (Hinglish). *arXiv preprint arXiv:2008.11398*.

Sodhar, I.N., Jalbani, A.H., Buller, A.H., Channa, M.I. and Hakro, D.N., 2020. Sentiment analysis of Romanized Sindhi text. *Journal of Intelligent & Fuzzy Systems*, *38*(5), pp.5877-5883.

Sriram, A., Li, Y. and Hadaegh, A., 2021, August. Mining Social Media to Understand User Opinions on IoT Security and Privacy. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)* (pp. 252-257). IEEE.

Tao, Y., Zhang, F., Shi, C. and Chen, Y., 2019. Social media data-based sentiment analysis of tourists' air quality perceptions. *Sustainability*, *11*(18), p.5070.

Wang, S., Liu, D., Wang, N. and Yuan, Y., 2020. Design and implementation of an online Python teaching case library for the training of application-oriented talents. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(21), pp.217-230.

Zahidi, Y., El Younoussi, Y. and Al-Amrani, Y., 2021. A powerful comparison of deep learning frameworks for Arabic sentiment analysis. *International Journal of Electrical & Computer Engineering (2088-8708)*, *11*(1).