

UNIVERSITY OF BUEA

P.O. Box 63,
Buea, South West Region
CAMEROON

Fax: (237) 3332 22 72

Tel: (237) 3332 21 34/3332 26 90



REPUBLIC OF CAMEROON

PEACE-WORK-FATHERLAND

Department of Computer Engineering
Faculty of Engineering and Technology

Cover page

PROJECT TITLE

WHATSAPP FEELING ANALYSIS USING BIG DATA

A dissertation submitted to the Department of Computer Engineering, Faculty of Engineering and Technology, University of Buea, in Partial Fulfilment of the Requirements for the Award of Bachelor of Engineering (B.Eng.) Degree in Computer Engineering

TAWAH PEGGY CHE NICO
Matriculation Number: FE16A088

Supervised by:

Mr. Sop DEFFO Lionel Landry

Academic Year: 2019/2020



UNIVERSITY OF BUEA

**REPUBLIC OF CAMEROON
PEACE WORK FATHERLAND**

P.O. BOX 63

Buea SouthWest Region

CAMEROON

Tel(237 3332 21 34/ 3332 26 90

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING

WHATSAPP FEELING ANALYSIS USING BIG DATA

*Dissertation submitted in partial fulfilment of the Requirements for the award of
Bachelor of Engineering (B.Eng.) Degree in Computer Engineering.*

By

Tawah Peggy Che Nico

Matriculation Number: FE16A088

Option: Software Engineering

Supervised by:

Mr SOP DEFFO Lionel Landry

Academic year 2019/2020

CERTIFICATION OF ORIGINALITY

We the undersigned, hereby certify that this dissertation entitled “**What Sapp Feeling Analysis and Big Data in Telecommunication**” presented by **TAWAH PEGGY CHE NICO**, Matriculation number **FE16A088** has been carried out by her in the Department of Computer Engineering, Faculty of Engineering and Technology, University of Buea under the supervision of **Mr Sop Deffo**.

This dissertation is authentic and represents the fruits of her own research and efforts.

Date.....

Student

__TAWAH PEGGY GHE NICO__

Supervisor

__Mr SOP DEFFO Lionel Landry__

Head of Department

.....

DEDICATION

This piece of work is dedicated to my beloved sweet loving MOTHER Mrs QUINTA NGWE ASANJI who has done everything in her will to see that success becomes my potion in Jesus Name.

ACKNOWLEDGEMENT

Firstly, I thank God Almighty for this wonderful inspiration given to me to accomplish such great task. Despite the challenges, he made it possible for things to work out smoothly. I thank every individual who helped me in accomplishing this project. I also want to thank and acknowledge my family as a whole and my mum especially for seeing to it that my success is guaranteed. In deed my stay in the University of Buea for the Bachelor in Engineering has come to an end. I thank the faculty of engineering and technology for given me this project and also for assisting me in doing it by teaching me the way to go about. And hence obtaining a Bachelor Degree. Great thanks also go to my supervisor Mr Sop Deffo for being with me through the internship process and also through the final year project research and beyond all his advice he has been given to me and making it easier for me to write ma project report. All those who contributed right from my start of the degree program till now that I'm finishing, I want to take the opportunity to thank you all very much for your loyalty. The elder brother Emi, even though not having, but struggle all means to support me till I finish the school years. I still extend much thank and regards to Go-Groups company LTD and staff for making me understanding several things. The internship process help me a lot to accomplish this final year report and implementation with no stress. Thanks to Eng. Tafang Joshua, Eng. Bide Nelson Noyo, Eng. Mantoh Nasah, Eng. Noel Magaza and Eng. Takungang Dieudonne, Mr Bruno Kechaman and Mr Nji Kaesey Adams Annuh without you all there would have been nothing to report about. May the Almighty God be with you all and continue to bless you in your Endeavour.

Abstract

Getting to understand what people actually think about your company or business is very important and resourceful and so sentiment analysis comes in to play a great role. Sentiment analysis is actually getting the sentiment from a tweet or chat by checking verbally each word and getting if the word is positive, negative or neutral. This document is a report of my final year project's analysis, design and implementation. It aims at describing the process of carrying out sentiment analysis on a huge dataset gotten from twitter through the twitter API by streaming tweets using the tweepy library. The data obtained is then processed with the focus on the text, and then the sentiment of each tweet is determined. The result of the sentiment analysis is then gotten in the form of graphs and charts from which decisions and predictions can be made based on the subject of interest using python for back end and training of the model while angular was used for the front end. The methodology used is the agile methodology with a small part from the prototype model as shown in chapter 3. the universal modelling language was used for design using the community version of visual paradigm for drawing. Significant work was done to improve and advance the field of sentiment analysis using machine learning and a supervised learning. Leaving me with an accuracy of 73% and a good responsive application

Table of Contents

Cover page.....	i
.....	ii
Title page	ii
CERTIFICATION OF ORIGINALITY	iii
DEDICATION.....	iv
ACKNOWLEDGEMENT.....	v
Abstract	vi
LIST OF FIGURES	xi
1.2.1 Introduction	1
1.2.2 WHAT THEN IS SENTIMENT ANALYSIS?	1
1.2.3 HOW SENTIMENT ANALYSIS WORK	2
2.2.4 BENEFITS OF SENTIMENT ANALYSIS	2
2.2.6 TYPES OF SENTIMENT ANALYSIS.....	3
2.2.5 SENTIMENT ANALYSIS CHALLENGES	4
2.2.6 APPLICATION OF SENTIMENT ANALYSIS WITH RESPECT TO SOCIAL MEDIA	5
2.2.7 CLASSIFICATION OF SENTIMENT ANALYSIS ALGORITHMS	6
NAÏVE BAYES	6
LINEAR REGRESSION	6
SUPPORT VECTOR MACHINES	6
DEEP LEARNING.....	6
3. BIG DATA.....	6
3.2 DEFINITIONS RELATED TO BIG DATA.....	7
3.3 CHARACTERISTICS OF BIG DATA.....	7
VOLUME	7
VARIETY	7
VELOCITY.....	7

VARIABILITY	7
3.4 Types of Big Data.....	8
STRUCTURED.....	8
UNSTRUCTURED	8
SEMI-STRUCTURED	8
3.5 BENEFITS AND ADVANTAGES OF BIG DATA ANALYSIS PROCESSING ...	8
3.6 DISADVANTAGES OF BIG DATA PROCESSING	9
3.7 STAGES OF BIG DATA ANALYSIS	9
4 TWITTER.....	12
4.1 DESCRIPTION AND DEFINITION.....	12
4.1.1 DESCRIPTION	12
4.2 IMPORTANCE OF TWITTER SENTIMENT ANALYSIS	12
REAL-TIME ANALYSIS	12
SCALABILITY.....	12
CONSISTENT CRITERIA.....	12
THE TWITTER API	12
TWITTER STREAMING API:	13
STATUSES AND FILTER (FREE) API:	13
USING TWITTER API FOR SENTIMENT ANALYSIS.....	13
STEPS USED IN CARRYING OUT TWITTER API SENTIMENT ANALYSIS	13
PREPARE YOUR DATA	13
CREATE A TWITTER SENTIMENT ANALYSIS MODEL.....	13
SEARCH FOR TWEETS	13
TAG DATA TO TRAIN YOUR CLASSIFIER	14
TEST YOUR CLASSIFIER	14
PUT THE MODEL TO WORK.....	14
VISUALIZE YOUR RESULTS	14
5 DEFINITION OF KEY TERMS	14
NATURAL LANGUAGE PROCESSING (NLP).....	14
CONTENT CATEGORIZATION.	14
TOPIC DISCOVERY AND MODELLING.....	14
Contextual extraction.....	14
Sentiment analysis.....	14
TOKENIZING	15
PICKLING.....	15

TEXT CLASSIFIERS	15
TWEEPY	15
TEXTBLOB.....	15
6 PROBLEM STATEMENT.....	15
CHAPTER TWO: LITERATURE REVIEW.....	17
1. INTRODUCTION.....	17
2. GENERAL CONCEPTS ON TWITTER SENTIMENT ANALYSIS	17
Datasets.....	17
DATA RETRIEVAL.....	18
Homophily	19
Reciprocity.....	19
INFORMATION DIFFUSION	19
INFLUENCE ON TWITTER.....	19
LEVELS OF SENTIMENT ANALYSIS	19
MACHINE LEARNING APPROACHES	20
UNSUPERVISED LEARNING	21
SUPERVISED LEARNING:.....	21
3. RELATED WORKS	21
Afroze Ibrahim Baqapuri, 2012. (Twitter Sentiment Analysis).....	29
2.4 Partial conclusion	32
CHAPTER THREE: ANALYSIS AND DESIGN	33
3.1 Introduction	33
3.2 Proposed Methodology	33
3.2.1 Definition of agile methodology	33
3.2.2The General Principles of the Agile Method	33
History of Agile Method.....	34
3.3.3. Companies that Use the Agile Method	34
3.3.4. Benefits of Using the Agile Method.....	35
3.3.5. CRITICISM OF AGILE DEVELOPMENT	35
3.3.6. DIFFERENCE BETWEEN AGILE AND TRADITIONAL (WATERFALL OR SPIRAL) DEVELOPMENT	36
3.3.7. AGILE METHODOLOGY GLOSSARY	37
3.4 BRIEF DESCRIPTION OF THE PROTOTYPE MODEL.....	38
3.5 System Design and Architecture	39
3.5.1. Use Case Diagram	39

3.5.4 Activity Diagram	43
3.5.5. Data Flow Diagram.....	44
3.6. GLOBAL ARCHITECTURE OF THE SYSTEM.....	46
3.6.1 PLAN	46
3.6.2 DESIGN	46
3.6.3 IMPLEMENTATION.....	46
3.6.4 TESTING	46
3.6.5 RELEASING.....	46
3.6.6 FEEDBACK.....	47
3.7 DESCRIPTION OF THE RESOLUTION PROCESS	47
3.7.1 INTRODUCTION	47
3.7.2 DESCRIPTION	47
3.8. PARTIAL CONCLUSION.....	47
CHAPTER FOUR: IMPLEMENTATION, REALIZATION AND PRESENTATION OF RESULTS.....	47
4.1 Introduction	47
4.2 TOOLS AND MATERIALS USED.....	48
4.2.1 DEFINITION AND DESCRIPTION OF SOME MATERIALS USED.....	48
5 DESCRIPTION OF THE IMPLEMENTATION PROCESS.....	49
5.1. GETTING DATA FROM TWITTER STREAMING API.....	49
5.2. GETTING TWITTER API KEYS	49
5.3. Connecting to Twitter Streaming API and downloading data	49
5.4. PRESENTATION AND INTERPRETATION OF RESULTS.....	49
5.4.1. FRONTEND.....	49
5.4.2 BACKEND.....	51
5.4.3. MODEL	53
5.5. Evaluation of the solution.....	56
5.6. Partial conclusion.....	56
CHAPTER 5 CONCLUSION AND FUTURE WORK	57
5.1 Summary of findings	57
5.2 Contribution to engineering and technology.....	57
5.3 Recommendations	58
5.4 DIFFICULTIES ENCOUNTERED	58
5.5 Further works	59

LIST OF FIGURES

Figure 1:prediction and training of a sentiment analysis model.....	2
Figure 2: stages of big data analysis	10
Figure 3: describing the different phases in agile development	38
Figure 4 The prototype model	39
Figure 5 usecase diagram for twitter sentiment analysis.....	40
Figure 6 flow chat diagram for twitter sentiment analysis	41
Figure 7 Class Diagram for Sentiment Analysis using Twitter API.....	42
Figure 8 Activity diagram for twitter sentiment analysis	43
Figure 9 twitter sentiment Analysis data flow diagram	44
Figure 10 twitter life streaming sentiment analysis sequence diagram.....	45
Figure 11 twitter sentiment analysis flow chat.	44
Figure 12 front end code	50
Figure 13 front end home page	50
Figure 14 front end feedback form	51
Figure 15 backend code.....	52
Figure 16 back end result.	52
Figure 17 sentiment analysis graph.....	53
Figure 18 code for traning my model.....	54
Figure 19:trained model analysis	55
Figure 20 : testing the model	55
Figure 21: result of the model output	55
Figure 22:twitter developer account dashboard.....	56

CHAPTER 1: GENERAL INTRODUCTION

1.1 Document Definition

This document as reviewed from the beginning of the goal purpose entice of the Final Report of the Project assign or given to me to accomplish as my final year project in the faculty of engineering and technology for the academic year 2019/2020. In this document, all necessity required for the implementation of the project are stated in this document starting from the requirement document, design and implementation and followed by the desired result.

1.2 Background and context of study

1.2.1 Introduction

It's estimated that 80% of the world's data is unstructured, in other words it's unorganized. And Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analysing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs .Huge volumes of text data (emails, support tickets, chats, social media conversations, surveys, articles, documents, etc.), is created every day but it's hard to analyse, understand, and sort through, not to mention time-consuming and expensive. Structuring and understanding this huge data and getting the sense out of it can be very useful and this is where sentiment analysis using big data comes to play.

1.2.2 WHAT THEN IS SENTIMENT ANALYSIS?

It is a natural language processing (NLP) and information extraction task that aims to obtain writer's feelings expressed in negative or positive comments, questions and requests by analysing a large number of documents.

Sentiment analysis deals with identifying and classifying opinions or sentiments which are present in a source text. Social media is generating a huge amount of sentiment rich data in the form of tweets, status updates, reviews and blog posts etc.

Sentiment analysis of these users' generated data is very useful in knowing the opinions of the crowd.

1.2.3 HOW SENTIMENT ANALYSIS WORK

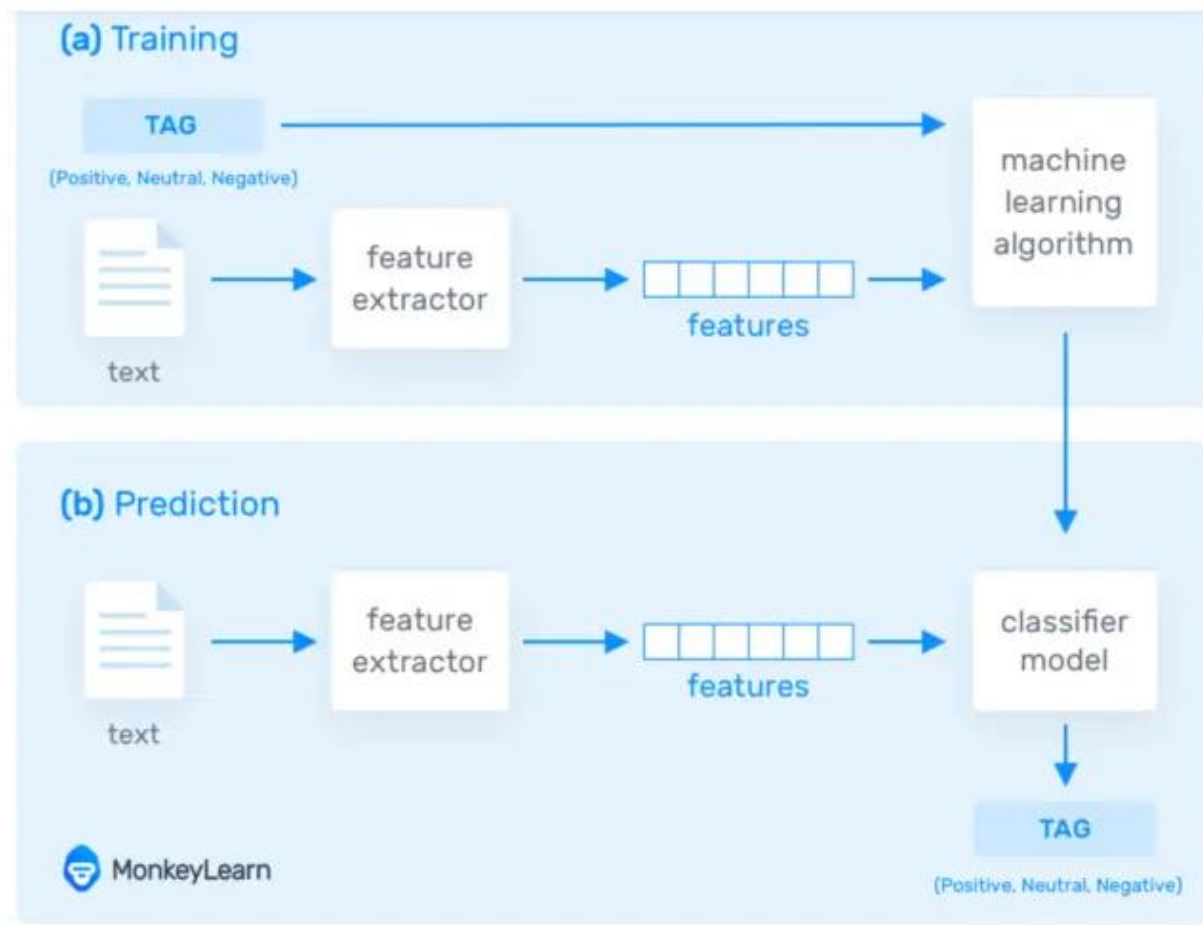


Figure 1: prediction and training of a sentiment analysis model

2.2.4 BENEFITS OF SENTIMENT ANALYSIS

Sentiment analysis has so much benefits be it in business, politics and social sectors. Some of these outline benefits include:

UPSELLING OPPORTUNITIES

Happy customers are more likely to be receptive to upselling. With sentiment analysis, you can easily identify your happiest customers. This helps you recognise chatters who might be receptive to spending more, as well as avoiding upsetting disgruntled customers with any unwelcomed sales pitches.

AGENT MONITORING

One of the most helpful benefits of sentiment analysis is its utility as a performance measurement tool. Sentiment analysis gives you a clear overview of customers' satisfaction, agent by agent. This means you can keep an eye on the quality of service each team member is offering to customers.

TRAINING CHAT BOTS

The benefits of sentiment analysis goes beyond helping your human agents. If you have a chat bot on your site, it can benefit from sentiment analysis too. That's because sentiment analysis can train your chat bot to recognise and respond to customer mood. For example sentiment analysis could detect when a chat bot needs escalating to a human agent, or route an engaged prospect through to a sales team.

HANDLING MULTIPLE CUSTOMERS

In a chat session, agents can find themselves handling more than one customer at a time. Keeping track of how a particular customer is feeling can be a challenge at any given time particularly busy hours. Sentiment analysis helps you keep track of the mood for any number of customers for your team. At a glance, you can see which chats are going smoothly and which needs further attention. So sentiment analysis reduces the risk of reading the chat room wrongly.

ADAPTIVE CUSTOMER SERVICE

Your human agents are great at providing flexible services but it can be difficult to identify the best approach for each customer earlier on. Empathetic service makes a great experience.

IDENTIFYING KEY EMOTIONAL TRIGGERS

Emotional triggers drive our decisions. Using sentiment analysis you can identify what messages and conversations act as emotive triggers and change customer mood. Understanding what messages trigger certain emotions in your customers can help you give better service, and it's also useful for creating effective marketing materials. Perhaps the phrase "please wait" for example often triggers customer annoyance or using emoji have a positive effect on the conversation's overall tone

LIVE INSIGHT

Customer's mood can change at any point in time during a customer service interaction, and this is not always clear. With sentiment analysis, not only can your agent see the mood of each customer in a session, visual indicators displays how her mood changes in real time. Your agents get a live insight into how well a chat is going.

TRACKING OVERALL CUSTOMER SATISFACTION






Sentiment analysis scoring put a quantifiable number on customer satisfaction. It enables you to see the impression and moods of customers when they approach you before they get support, and how effective is your service is at increasing satisfaction. In other words you get the bigger picture rather than just a normal case-by-case view

2.2.6 TYPES OF SENTIMENT ANALYSIS



Sentiment analysis models focus on polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad, etc.), and even on intentions (e.g. interested v. not interested). Here are some of the most popular types of sentiment analysis:

FINE-GRAINED SENTIMENT ANALYSIS

If polarity precision is important to your business, you might consider expanding your polarity categories to include:

-  Very positive
-  Positive
-  Neutral
-  Negative
-  Very negative

This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

-  Very Positive = 5 stars
-  Very Negative = 1 star

EMOTION DETECTION

This type of sentiment analysis aims at detecting emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.

One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like bad or kill (e.g. your product is so bad or your customer support is killing me) might also express happiness (e.g. this is bad ass or you are killing it).

ASPECT-BASED SENTIMENT ANALYSIS

Usually, when analysing sentiments of texts, let's say product reviews, you'll want to know which particular aspects or features people are mentioning in a positive, neutral, or negative way. That's where aspect-based sentiment analysis can help, for example in this text: "The battery life of this camera is too short", an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the feature battery life.

MULTILINGUAL SENTIMENT ANALYSIS

Multilingual sentiment analysis can be difficult. It involves a lot of pre-processing and resources. Most of these resources are available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms), but you'll need to know how to code to use them.

2.2.5 SENTIMENT ANALYSIS CHALLENGES

Computer scientists have been trying to develop more accurate sentiment classifiers, and overcome limitations in recent years. Let's take a closer look at some of the challenges they face:

1. SUBJECTIVITY AND TONE

The detection of subjective and objective texts is just as important as analysing their tone. In fact, so called objective texts do not contain explicit sentiments. Say, for example, you intend to analyse the sentiment of the following two texts:

The package is nice.

The package is red.

Most people would say that sentiment is positive for the first one and neutral for the second one, right? All predicates (adjectives, verbs, and some nouns) should not be treated the same with respect to how they create sentiment. In the examples above, nice is more subjective than red.

CONTEXT AND POLARITY

All utterances are uttered at some point in time, in some place, by and to some people, you get the point. All utterances are uttered in context. Analysing sentiment without context gets pretty difficult. However, machines cannot learn about contexts if they are not mentioned explicitly. One of the problems that arise from context is changes in polarity.

IRONY AND SARCASM

When it comes to irony and sarcasm, people express their negative sentiments using positive words, which can be difficult for machines to detect without having a thorough understanding of the context of the situation in which a feeling was expressed.

COMPARISONS

How to treat comparisons in sentiment analysis is another challenge worth tackling. Look at the texts below:

This product is second to none.

This is better than older tools.

This is better than nothing.

The first comparison doesn't need any contextual clues to be classified correctly. It's clear that it's positive.

The second and third texts are a little more difficult to classify, though. Would you classify them as positive, or negative? Once again, context can make a difference. For example, if the 'older tools' in the second text were considered useless, then the second text is pretty similar to the third text.

2.2.6 APPLICATION OF SENTIMENT ANALYSIS WITH RESPECT TO SOCIAL MEDIA

It is used to analyse tweets and posts over a period of time in order to detect the sentiment of a particular audience.

To monitor social media mentions of a particular brand.

Automatically route social media mentions to team members best fit to respond.

Gain deep insight as to what is happening at your social media channels

It tracks trends over time.

2.2.7 CLASSIFICATION OF SENTIMENT ANALYSIS ALGORITHMS

The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks:

NAÏVE BAYES

Naïve Bayes is a family of probabilistic algorithms that uses Bayes' Theorem to predict the category of a text.

LINEAR REGRESSION

Linear regression is a very well-known algorithm in statistics used to predict some value (Y) given a set of features (X).

SUPPORT VECTOR MACHINES

It is a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing texts and the regions they're mapped to.

DEEP LEARNING

It is a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data.

Hybrid Approaches

Hybrid systems combine the desirable elements of rule-based and automatic techniques into one system. One huge benefit of these systems is that results are often more accurate.

3. BIG DATA

3.1 INTRODUCTION

The term big data has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems.

These are datasets whose sizes are beyond the ability of commonly used software tools and storage system to capture, store, manage as well as process the data within a tolerable elapse time.

Big data sizes are constantly increasing, currently ranging from a few dozen terabytes to many petabytes of data in a single data set. Consequently, some of the difficulties related to big data include: capture, search, storage, sharing, analytics and visualising.

Today enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before.

3.2 DEFINITIONS RELATED TO BIG DATA

DATA

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media is data.

BIG DATA

Big data is the term used to refer to initiatives and technologies that comprises of data that is too diverse, fast evolving, and vast for ordinary technologies, infrastructure and skills to address exhaustively. In other words, it is the study of huge amount of stored data in order to extract behaviour patterns.

BIG DATA ANALYTICS

Big data analytics is where advanced analytics techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the dataset, the more difficult it is to be managed.

3.3 CHARACTERISTICS OF BIG DATA

VOLUME

The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, '**Volume**' is one characteristic which needs to be considered while dealing with Big Data.

VARIETY

The next aspect of Big Data is its variety.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

VELOCITY

The term velocity refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

VARIABILITY

This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

3.4 Types of Big Data

Big Data could be found in three forms:

- Structured
- Unstructured
- Semi-structured

STRUCTURED

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes. Looking at these figures one can easily understand why the name Big Data is given and imagine the challenges involved in its storage and processing.

UNSTRUCTURED

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples of Un-structured Data

The output returned by 'Google Search'

SEMI-STRUCTURED

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples of Semi-structured Data

Personal data stored in an XML file

3.5 BENEFITS AND ADVANTAGES OF BIG DATA ANALYSIS PROCESSING

Ability to process Big Data brings in multiple benefits, such as-

- Businesses can utilize outside intelligence while taking decisions.
Access to social data from search engines and sites like Facebook and twitter are enabling organizations to fine tune their business strategies.
- Improved customer service. Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.
- Early identification of risk to the product/services, if any

- Better operational efficiency
- Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.
- Big data processing is also applicable in the following sectors Banking, education, health care and government
- Manufacturing
Armed with the insight that big data can provide, manufacturers can boost quality and output while minimizing waste. Processes that are key in today's highly competitive market.
- Retail
Customer relationship building is critical to the retail industry and the best way to manage that is to manage big data.
- Big data analysis derives innovative solutions.
- Big data analysis helps in understanding and targeting customers.
- It helps in optimizing business processes.
- It helps in improving science and research.
- It improves healthcare and public health with availability of record of patients.
- It helps in financial trading's, sports, polling, security/law enforcement etc.
- Anyone can access vast information via surveys and deliver answer of any query.
- Every second additions are made.
- One platform carry unlimited information.

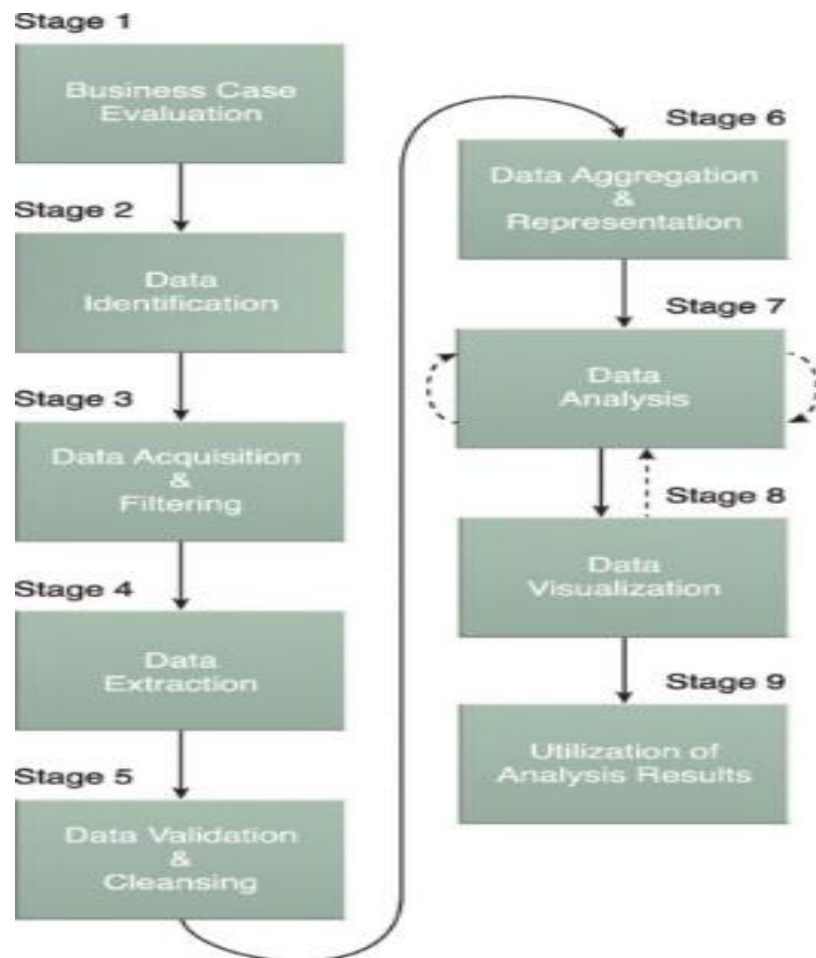
3.6 DISADVANTAGES OF BIG DATA PROCESSING

- Following are the drawbacks or disadvantages of big data :
- Traditional storage can cost lot of money to store big data.
- Lots of big data is unstructured.
- Big data analysis violates principles of privacy.
- It can be used for manipulation of customer records.
- It may increase social stratification.
- Big data analysis is not useful in short run. It needs to be analysed for longer duration to leverage its benefits.
- Big data analysis results are misleading sometimes.
- Speedy updates in big data can mismatch real figures.

3.7 STAGES OF BIG DATA ANALYSIS

Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processed. To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analysing and repurposing data. Below is a specific data analytics life cycle that organizes and manages the tasks and activities associated with the analysis of Big Data

Figure
stages
big
data



2:
of

analysis

1. BUSINESS CASE EVALUATION

Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis. The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.

2. DATA IDENTIFICATION

The Data Identification stage is dedicated to identifying the datasets required for the analysis project and their sources.

3. DATA ACQUISITION & FILTERING

During the Data Acquisition and Filtering stage, the data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.

4. DATA EXTRACTION

Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The Data Extraction lifecycle stage is dedicated to extracting disparate data

and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

5. DATA VALIDATION & CLEANSING

Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints. The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.

6. DATA AGGREGATION & REPRESENTATION

Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined. The Data Aggregation and Representation stage is dedicated to integrating multiple datasets together to arrive at a unified view.

7. DATA ANALYSIS

The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

8. DATA VISUALIZATION

The ability to analyse massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.

The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

9. UTILIZATION OF ANALYSIS RESULTS

Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportunities to utilize the analysis results. The Utilization of Analysis Results stage is dedicated to determining how and where processed analysis data can be further leveraged.

4 TWITTER

4.1 DESCRIPTION AND DEFINITION

4.1.1 DESCRIPTION

With more than 321 million active users, sending a daily average of 500 million Tweets, Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, it's harder for brands to quickly detect negative content, and if it goes viral you might end up with an unexpected PR crisis on your hands

Twitter is a micro-blogging and social networking service on which people chat and interact with messages known as tweets. Tweets were originally restricted to not more than 140 characters, registered users can post, like, and retweet tweets, but unregistered users can only read them. But on November 7 2017, this limit was doubled making it 280 to all countries except china, Korea and japan. Twitter users follow other users. If you follow someone you can see their tweets in your twitter 'timeline'

Monitoring Twitter allows companies to understand their audience, keep on top of what's being said about their brand and their competitors, and discover new trends in the industry.

4.2 IMPORTANCE OF TWITTER SENTIMENT ANALYSIS REAL-TIME ANALYSIS

Twitter sentiment analysis is essential for monitoring sudden shifts in customer moods, detecting if complaints are on the rise, and for taking action before problems escalate. With sentiment analysis, you can monitor brand mentions on Twitter in real-time and gain valuable insights that tell you if you need to make updates.

SCALABILITY

Let's say you need to analyze hundreds of tweets mentioning your brand. While you could do that manually, it would take hours of manual processing, and as your data grows it would be impossible to scale. By performing Twitter sentiment analysis you can automate manual tasks and gain valuable insights in a very short time.

CONSISTENT CRITERIA

Analyzing sentiment in a text is subjective. When done manually. The same tweet may be viewed differently by two members of the same team. By training a machine learning model to perform sentiment analysis on Twitter data, you can use one set of criteria to analyze all your data, so results are consistent.

THE TWITTER API

The Twitter API (the term stands for Application Programming Interface) enables software developers to access and interact with public Twitter data. Developers can interact with this API by writing their own scripts or by using one of the open source libraries available in different programming languages.

The Twitter API has 2 APIs that are useful for extracting tweets:

TWITTER STREAMING API:

This API allows you to connect to the Twitter data stream and gather tweets in real-time. You can listen to all the Tweets matching a certain keyword, mention or hashtag, as well as collect the tweets of specific users, at the time they are being posted in the Twitter platform.

Twitter offers a free version of its streaming API, as well as a paid version:

STATUSES AND FILTER (FREE) API: This API allows you to track tweets with up to 400 keywords, hashtags or mentions, monitor up to 5,000 user IDs, and up to 25 locations.

USING TWITTER API FOR SENTIMENT ANALYSIS

Before you can start analysing tweets and getting their sentiments from Twitter, you'll first need a Twitter account yourself. You'll need to create a Twitter application to get your keys

STEPS USED IN CARRYING OUT TWITTER API SENTIMENT ANALYSIS

PREPARE YOUR DATA

Once you've captured the tweets you need for your sentiment analysis, you'll need to prepare your data. As we mentioned earlier, social media data is unstructured. That means it's raw, noisy, and needs to be cleaned before we can start working on our sentiment analysis model. This is an important step because the quality of the data will lead to more reliable results. Pre-processing a Twitter dataset involves a series of tasks like removing all types of irrelevant information like emoji's, special characters, and extra blank spaces. It can also involve making format improvements, delete duplicate tweets, or tweets that are shorter than three characters. Check out this guide on how to prepare your data.

CREATE A TWITTER SENTIMENT ANALYSIS MODEL

If you want to get predictions with a high level of accuracy, adapted to your criteria and domain, then the best way is to create your own customized sentiment analysis model by training it with your industry criteria.

IMPORT YOUR TWITTER DATA

This data will be used as examples to train your machine learning model. In this case, we'll choose Twitter.

SEARCH FOR TWEETS

Write a search query to obtain tweets that will be used as training data for your sentiment analysis model. It can be a keyword, hashtag, or mention. Then, choose the column you'd like to use (it's often the text of the tweet). Once you are done, the tweets will be imported to your classifier:

TAG DATA TO TRAIN YOUR CLASSIFIER

Now, it's time to train your sentiment analysis model, by manually tagging each of the tweets as Positive or Negative based on the polarity of the opinion. After tagging the first tweets, the model will start making its own predictions. You can correct them if the answer is not correct

TEST YOUR CLASSIFIER

Once you have trained your model with a few examples, you can paste your own texts to see how the sentiment analysis model classifies it

PUT THE MODEL TO WORK

Now you've got a sentiment analysis model that's ready to analyze tons of tweets! The next step is to integrate the Twitter data you want to analyze with the sentiment analysis model you just created

VISUALIZE YOUR RESULTS

Data visualization tools help explain sentiment analysis results in a simple and effective way.

5 DEFINITION OF KEY TERMS

NATURAL LANGUAGE PROCESSING (NLP)

It is a branch of **artificial intelligence** that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

Tasks performed by nltk

CONTENT CATEGORIZATION.

A linguistic-based document summary, including search and indexing, content alerts and duplication detection.

TOPIC DISCOVERY AND MODELLING.

Accurately capture the meaning and themes in text collections, and apply advanced analytics to text, like optimization and forecasting.

Contextual extraction

Automatically pull structured information from text-based sources.

Sentiment analysis

Identifying the mood or subjective opinions within large amounts of text, including average sentiment and opinion mining.

- **Speech-to-text and text-to-speech conversion**

Transforming voice commands into written text, and vice versa.

- **Document summarization**

Automatically generating synopses of large bodies of text.

- **Machine translation**

Automatic translation of text or speech from one language to another.

TOKENIZING

Tokenization is the process of protecting sensitive data by replacing it with an algorithmically generated number called a token. Often time's **tokenization** is used to prevent credit card fraud. ... The actual bank account number is held safe in a secure token vault.

PICKLING

Pickling is the serializing and de-serializing of python objects to a byte stream. Unpickling is the opposite. You may hear this methodology called serialization, marshallng or flattening in other languages, but it is pretty much exclusively referred to as pickling in Python.

TEXT CLASSIFIERS

Text classification also known as **text** tagging or **text** categorization is the process of categorizing **text** into organized groups. By using Natural Language Processing (NLP), **text classifiers** can automatically analyze **text** and then assign a set of pre-defined tags or categories based on its content

TWEEPY

It is the python client for the official Twitter API. Install it using following pip command:

TEXTBLOB

Textblob is the python library for processing textual data. Install it using following pip command:

6 PROBLEM STATEMENT

Streaming Life Twitter Sentiment analysis using the twitter application programming interface (API)

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract and quantify and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials. Such as reviews and survey responses, online and social media and health care materials for applications that range from marketing to customer service to clinical medicine.

This thesis focusing the analysis of people about certain subjects, activity such

As commercial product, political situations

Given a message, classify whether the message is of positive or negative sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen

CHAPTER TWO: LITERATURE REVIEW

1. INTRODUCTION

The growing phenomena of social media, such as' twitter, LinkedIn, and Instagram, with each one has its own characteristics and its usages, are constantly affecting our societies. Facebook, for example, is considered as a social network where everyone in the network has a reciprocated relationship with one in the same network. The relationship in this case is indirect. Conversely, in twitter everyone in the network does not necessarily have a reciprocated relationship with others. In this case, the relationship is either directed or undirected.

In this paper, we focus on twitter for data analysis, where twitter is an online networking service that enables users to send and read short 280- character messages called "tweets". In addition to its publicity, twitter is accessible for unregistered users to read and monitor most tweets, unlike Facebook where users can control the privacy of their profiles. Twitter is also a large social networking microblogging site. The massive information provided by twitter such as tweet messages, user profile information, and the number of followers/following in the network play a significant role in data analysis, which in return make most studies investigate and examine various analysis techniques to grasp the recent seed technologies.

Sentiment analysis in the domain of micro-blogging is a relatively new research topic, so there is still a lot of room for further in this area. Decent amount of related prior work has been done on sentiment analysis. These differ from twitter mainly because of the limit of 280 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and support vector Machines. But the manual labelling required for the supervised approaches, and there is a lot of room of improvement. Various researched and semi - supervised approaches, and there is a lot of room of improvement. Various researchers testing performance. There is a need of proper and formal comparisons between these results arrived through different features and classification techniques often just compare their results arrived through different features and classification techniques in order to select the best features and most efficient classification techniques for particular applications.

2. GENERAL CONCEPTS ON TWITTER SENTIMENT ANALYSIS

To track and monitor different datasets, most studies began with collecting the desired datasets from twitter, and applied filtering techniques to remove redundant data or span tweets. Then parsed the data into a structured form and finally, analyzed the data. Below we review several types of analyses that most researchers have used.

Datasets

Analyzing structured data has been widely used. In such case, the traditional relational database management system (RDBMS) can deal with the data. With the increasing amounts of structured data on various sources (e.g., web, social media, and blog data) that are considered as big data, a single computer processor cannot process such huge amount of data. Hence, the RDBMS cannot deal with the unstructured data; a nontraditional database is needed to process the data, which is called noSQL database.

Most studies focused on tools, such as R (the programming language and the software environment for data analysis) and python. R has limitations when processing twitter data, and is not efficient in dealing with large volume of data. To solve this problem a hybrid big data framework is usually employed, such as Apache Hadoop (an open source java framework for processing and querying vast amounts of data on large clusters of commodity hardware). Hadoop also deals with structured and semi-structured data, XML/JSON files, for example. The strength of using R comes in analyzing the already-processed data. Over and beyond that, python is also an efficient and effective programming language that is used to carry out data analysis. Python is widely used in Apache Spark, which is an evolving technology used today in data analysis especially in the field of data Science.

There are different types of twitter data such as user profile data and twee messages. The former is considered static, while the latter is dynamic. Tweets could be textual, images, videos, URL, or spam tweets. Most studies do not, usually, take some tweets and automatic tweets engines into account as they can, often, affect the accuracy and add noise and bias to analysis results. The mechanism of Firefox add-on and clean tweet filter was employed to remove users that have been on twitter for less than a day and they removed tweets that contain more than three hashtags.

DATA RETRIEVAL

Before retrieving the data, some questions should be addressed like what are the characteristics of the data? And Is the data static, Such as the profile user information "Name" user ID and bio or dynamic such as user's tweets and user's network? It is important to note that it is easier to track a certain keyword attached to a hashtag rather than a keyword not attached to it.

Twitter-API is a widely used application to retrieve, read, and write data. Other studies, as used GNU/GPL application like Your TwapperKeper tool, which is a web-BASED APPLICATION THAT STORES SOCIAL MEDIA DATA IN MySQL TABLES. However, you're TwapperKeeper in storing and handling large size of data exhibits some limitations in using as MySQL and spreadsheets databases can only store a limited size data. Using a hybrid big data technology might address such limitations as suggested above.

2.2.3 Ranking and Classifying Twitters Users

There are different type of user's networks: a network of users within a specific event (hashtag), a network of users in a specific user's account and a network of user's within a group in the network. That is, twitter list. List are used to group sets of users into topical or other categories to better organize and filter incoming tweets. To rank tweet users, it's important to study the characteristics of twitter by studying the network-TOPOLOGY (NUMBER OF FOLLOWERS/FOLLOWED) for each user in the dataset. Many techniques have been employed in the ranking analysis. In twitter users are ranked by identifying the numbers of followers by studying the PageRank and by the retweet rate. In the study, 41.7 million use profiles, 1.47 billion social relations and 106 million tweets were used. A new methodology is introduced to rank twitter users by using the Twitter Lists to classify users into the Elite users (Celebrities, Media news, Politicians, Bloggers and Organizations) and the ordinary users.

Homophily

Homophily is defined as the tendency that contacts among similar users occur at a higher rate than among dissimilar users, that is similar users tend to follow each other. It requires studying the static characteristics of twitter network. Additional work had been investigated and homophily was studied using Twitter List to identify the similarity between the elite and ordinary users

Reciprocity

The characteristic nature of twitter as being both directed and undirected social network has made most studies analyze reciprocity. Reciprocity is the property of the following a user and followed back (mutual relationship), for instance, celebrities tend to follow each other, so are politicians, bloggers and ordinary users, it has been concluded that homophily and reciprocity have the same logical behavior, Reciprocal relationship is measured by analyzing the number of followers, page rank and retweet rate. Additional methodology is investigated, where the user's follower graph is studied to infer user's reciprocities.

INFORMATION DIFFUSION

Since there are different kinds of information spread over twitter, there is no agreement on what kind of information is more widely spread than others. There is also no agreement on how messages are spread over twitter network. In this area many studies have attempted to address those questions by studying the first-network topology and by measuring the retweet rate.

INFLUENCE ON TWITTER

Social influence occurs when an individual's thoughts or actions are affected by other people. Examining the influential users is related by the message propagation by answering on the following question; who are the originators of the tweets, how many audiences have and what is the retweet rate of the original tweet. Most studies agreed on analyzing the network topology and the retweet rate to identify the influential users. Additional methodology had been to examine the influence by studying the retweet mechanism through the "Centrality measures" techniques. "Centrality measures" technique used the "Degree Centrality" by counting the number of link attached to the node (user) in case of directed graph. Also employed the "Eigenvector Centrality" by answering the question of "how many users retweeted the node?"




LEVELS OF SENTIMENT ANALYSIS

There are several levels of sentiment analysis. These levels will not be completely described in this theses, however the will be enumerated. These levels include: Document Level, Sentiment Analysis, Sentence Level Sentiment Analysis, Entity or Aspect Level Sentiment Analysis. Phrase

Level sentiment analysis and the Feature Level Sentiment analysis

Sentiment Analysis Process

The figure bellow depicts the process analyzing tweets sentiment

-  Pre-processing of data sets
-  Feature extraction
-  Training

Classification: it has the following steps:

1. Pre-processing of datasets

Before analyzing the sentiments of tweets, pre-processing is carried out. This is because the tweets are the raw data and contains inconsistencies and redundancy which has to be removed to make it suitable for analysis. Pre-processing of the tweet of the following:

- ✚ Removal of hash tags and URLs.
- ✚ Removal of stop words
- ✚ Removal of non-English words.
- ✚ Removal of punctuations, symbols and bombers.
- ✚ Replace all the emoticons with sentiments

1. Feature Extraction

Pre-processed tweets possess various features which can be extracted in order to determine the positive and negative polarity of sentence. After extracting the features, the entire process becomes easy and can be done using the models like bigram or n-gram model. Machine learning techniques require these features to be extracted so that they can act as a feature vectors that can be used as training set in the learning. The features that can be extracted are:

Parts of speech tags: The part of speech like adverbs, nouns can be extracted to determine the subjectivity in the tweets. Syntactic dependency can be generated based on these using parse trees.

Words and their frequencies: The parts of speech like adverbs, nouns can be extracted to determine the subjectivity in the tweets, syntactic dependency can be generated based on these using parse trees.

Words and their frequencies: The frequency counts of the words in the text can be used in the models like unigram, bigram and ngram models.

Position of the term: The relative positioning of the terms and words in a phrase can act as a can determine the opinion of the text.

Specific opinion words and phrases: In the context of the situation, specific opinion words and phrases tend to repeat more frequently than the others, these words can be identified and used as a feature

Training: After extracting the features, the training data set created according to which the classifier is trained and then it is able to work for unknown data

Approaches/Techniques for sentiment classification (sentiment analysis algorithm)

OPINION mining (often referred as sentiment analysis) refers to identification and classification of opinion expressed in the text span; using information retrieval and computational linguistics. In this section i will expound on the carious sentiment analysis approaches and levels.

MACHINE LEARNING APPROACHES

Machine learning techniques are most useful techniques for the sentiment classification for categorized text into positive, neutral categories, in machine learning technique, training and testing dataset is used to learn the documents and test dataset is used to validate the

performance. There are number of machine learning algorithms used to classify reviews. Machine learning approaches classify the data into classes. There are two machine learning techniques include

UNSUPERVISED LEARNING

It is a learning technique without the training data and relies mainly on clustering. It is regarded to learn by observation.

SUPERVISED LEARNING:

It is regarded as to learn by example. A training data set is created and fed into the system to obtain the meaningful outputs. This helps in decision making. The efficiency of the supervised learning techniques is based on the fact that how accurately the features are extracted from the pre-processed data and fed as the feature vectors into the system i order to detect the sentiment. There are two types of data sets needed in the supervised learning algorithms: Training set and test set.

NAIVE BAYES

Naive Bayes is a simple and easy but effective classification algorithm. It is mostly used for document at level classification. The basic idea is to calculate the probabilities of categories given a test document by using joint probabilities of words and categories. Naive Bayes is optimal for certain problem classes with highly dependent features. Naive Bayes classifiers are computationally fast when taking when taking decisions. It does not require large amounts of data before learning can begin.

This is another sentiment classification approach/technique. It is the branch of computer science and technology which focused on developing systems that allow computers to communicate with people using natural language. Natural Language Processing (NLP) is the interaction between computers and human (natural) Languages. To evaluate sentiment of users online, particularly on twitter, effective sentiment annotation should be used. Most studies use the three common sentiment labels: positive, neutral, and negative. New feature had been used to effectively annotate sentiments of users; "Mixed sentiment Label", it exists in tweets that have two different meanings.

Natural language processing technique plays important role to get accurate sentiment analysis NLP techniques like Bag of words, Hidden markov model (HMM). Part of speech (POS)

3. RELATED WORKS

Sentiment Analysis is the thorough research of how opinions and perspectives can be relate to one's emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached up to great achievement which can surpass the positive versus negative and deal with whole arena of behaviour and emotions for different communities and topics. In the field of sentiment analysis using different techniques good amount of research has been carried out for prediction of social opinions.

Pang and lee, July 2002 proposed the system where an opinion can be positive or negative was found out by ratio of positive words to total words. Later in 2008 the authors developed

methodology in which tweet outcome can be decided by term in the tweet. Compare to baselines that are generated by humans, the results are pretty good when machine learning techniques are used. SVM gave best result as compare to Naïve Bayes. Regardless of using different types of features the authors did not attain desired accuracies over topic based categorization.

Jiang et al, June 2020 focus on target-dependent sentiment classification. Here target-dependent means whether the sentiment is positive, negative or neutral depends on nature of the question that is asked. The authors proposed to make better target-dependent sentiment classification by joining features of target-dependent and considering related tweets. The authors also proposed that there is need of consideration current tweets to the related tweets by employing graph based optimization. As claimed by authors' experimental results, the graph based optimization increases the performance.

Tan et al, July 2020 said that the users that shared similar opinions are likely to be connected. The authors proposed the model that were generated from either by following the network that has been made by tagging different user with the help of "@" or by analysing the network of twitter follower/followee. The authors explained that by employing information of link of twitter there will be improvement in user-level sentiment analysis.

Chen et al., June 2004 employed the feed-forward BPN network and uses sentiment orientation to calculate the results at each neuron. The authors proposed a methodology based on neural network. The proposed methodology is combination of machine learning classifiers and semantic orientation indexes. In order to obtain efficiency in methodology, semantic orientation indexes used as inputs for neural network. The proposed methodology outperforms other neural networks and traditional approaches by increasing efficiency in both training as well as classification time.

Malhar and Ram, 2002 employed supervised machine learning techniques and artificial neural networks to classify twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers. The authors proposed a methodology to predict the outcome of election results by utilizing the user influence factor. To carry out reduction in dimension the authors combined the Principle Component Analysis with SVM.

Anton and Andrey, October 2001 reviewed the existing techniques and developed a model for automatic sentiment analysis of twitter messages using unigram, bigram and jointly i.e. hybrid feature. The purpose of the authors is to explore and produce approaches for analysing the accent of the messages in social media. The authors reviewed existing automatic sentiment analysis approaches and in order to maintain the context of growing methods the character feature of social media statements were studied.

Pak and Paroubek, May 2012 perform linguistic analysis and build a sentiment classifier to determine positive, negative and neutral sentiments for a document. The authors developed a sentiment classifier, which gives neutral, negative and positive statements of a document. In order to train sentiment classifier the author proposed an approach that collects corpus automatically. In order to analyse the dissimilarity in diffusion among neutral, negative and positive sets, the authors used Tree Tagger.

Kopel and Schler, April 2008 explain that it is very important to use neutral messages to get good knowledge of polarity. The authors also states that positive and negative messages alone will not give proper understanding about neutral messages. Knowing about neutral messages clear the difference between positive and negative messages. The authors found that in one of the corpus having most of the neutral documents gives no sentiment which can be used as counter for testing both positivity and negativity of a document.

Go *et al.*, March 2009 introduced a methodology for automatic sentiment classification of twitter messages. Respective of query term messages were classified as negative or positive. Here authors use distant supervision to display the results of sentiments of twitter posts with the help of the machine learning algorithms. The algorithms such as Maximum Entropy, SVM and Naïve Bayes are applied to training data which contains emoticons, gave accuracy above 80%. The authors also discuss about pre-processing steps that was helped to obtain higher accuracy. The authors came up with an idea for distant supervised learning using tweets that contain emoticons.

Christianini and Taylor, 2010 published and shared the knowledge about SVM which is machine learning algorithm. The authors manage to give deep understanding about algorithm and how to approach the SVM algorithm in order to implement it to solve the practical problems. The approach will be theoretical as when the book was published, the research was on going on every field.

Burger *et al.*, February 2020. Since, in this era the computer have become enough powerful that can handle large scale application which gives pattern recognition and statistical estimation of real world problems. The authors introduced an approach for statistical modelling based on maximum entropy. By using examples of problems in natural language processing, the authors shows maximum-likelihood methodology for automatic construction of maximum entropy models. Here the authors described the principle of maximum entropy. This principle selects the model with greatest entropy among all the consistent models. By maximizing the likelihood of training data we can obtain optimal values of given parameters.

Romero *et al.*, February 2020 Discovered that hashtags becomes the common feature of twitter used in every message and new terms are created and changing on daily basis which effects the general meaning of the original term. The authors also found structural difference among issues and learn the structure of widely used different types of hashtags. The authors also developed generative and simulation based models to study the interaction between design of latest adopters on which hashtag expands and adoption dynamics.

Li and wu, February 2020 stated emotional polarity computation as sentiment analysis which have become prospering boundary in the community of text mining. With the help of text mining and sentiment analysis, here the authors studied about hotspot detection and forecast. The authors created an algorithm which describes emotional polarity of a message and obtain a value of each word in it. To create unsupervised text mining method, this algorithm is combined with support vector machine (SVM) and K-means clustering. After the experimental study both K-means and SVM obtain the same results for top 4 hotspots of the year.

Tan and Zhang, march 2020 Until this date very less number of researches carried for the Chinese documents on sentiment analysis. The authors studied sentiment categorization on Chinese documents. The selection methods were featured as Document Frequency (DF) , CHI,

Information Gain (IG) and Mutual Information (MI). The machine learning methods that are used are support vector machine (SVM), Naïve Bayes (NB), Winnow classifier, K-nearest neighbor classifier and Centroid classifier. Size of 1021 Chinese document were investigated. For selection of sentiment terms Information gain (IG) performs best and for sentiment classification SVM outperforms all the classifiers.

Martineau and Finin, February 2020 proposed a technique called Delta TFIDF which measure word scores efficiently before classification. Delta TFIDF was easy to understand, implement and compute. For sentiment classification the authors used support vector machines to achieve better accuracy with Delta TFIDF and using data sets of movie reviews. The authors said that Delta TFIDF is better than TDFIF feature and count term raw for all sizes of documents that weights for congressional detecting support for bill, sentiment polarity classification and subjectivity detection. The authors stated Delta TFIDF is first measuring approach to boost and identify the relevance of selective words using the calculated unsupervised distribution of features before classification between the two classes.

Nielson, April 2016 developed a labelled word list in which scores of the effective words had been obtained comes into the messages analysing for sentiment analysis. Before the arrival of sentiment analysis and micro blogging there contains an effective term list for e.g. Affective Norm for English Words (ANEW) developed by the author. The author made the word list exclusively for micro blogs i.e. ANEW in comparison with other list which can be used for detecting sentiment strength for micro blogs. The author used Twitter posted messages for scoring words for sentiment analysis.

Mohammad *et al.* January 2017 developed two SVM classifiers, one is term level task which determines the sentiment of a word in the message and one is message level task which determines the sentiment of messages such as SMS and tweets. The authors took part in a competition where 44 teams came in their submissions stood first in work on tweets, getting 88.93 F-core in term-level task and 69.02 F-score in message-level task. The authors executed sentiment, semantic and surface-form features. The authors also produced two big term-sentiment associations, first with emoticons from tweets and second with sentiment –term hashtags from tweets.

Kouloumpis *et al.* march 2018 explored the advantage of semantic features for analysing the sentiment of messages of Twitter. The authors investigate the features that collects knowledge about intuitive and informal language that used in microblogging as well as advantage of existing lexical resources. The authors used the supervised learning method to the problem and to collect it hashtags are used. The authors concluded that in the experimental study part-of-speech feature not better for sentiment analysis when it comes to the domain of microblogging on twitter and it confirmed that for collecting data hashtags are very useful so that messages with negative and positive emoticon.

Denecke, June 2020 proposed an approach for deciding polarity of word in framework of multilingual. The approach influences on lexical resources available in English for sentiment analysis. In this approach first the language itself is translated into English using the standard translation software. Further in translation the document is then classified into positive and negative class for sentiment analysis. The classification can be done on the basis of the adjective present in the document. The authors concluded that it is feasible approach to sentiment analysis in the multilingual framework.

Gokulkrishnan *et al.* January 2020 proposed a methodology for the pre-processing of publically generated tweets from twitter online microblogging site and on the basis of their opinion content of irrelevant, negative or positive sentiment classified can be done; and investigating the performance different classifying methods based on precision and recall. The authors explained limitations and applications of the research. The authors also handled the skewness of the datasets by exclusively new approach called SMOTE oversampling method which helped by increases the accuracy of the classifier. Random Forest, SVM and SMO generates better performance compare to Naïve Bayesian classifier.

Neri *et al.*, March 2019 performed sentiment analysis on newscast over more than 1000 Facebook posts and then compared the sentiment for dynamic company La7 and Rai – the Italian social broadcasting company which is emerging company. The authors observations were mapped with the study conducted by the Italian research institute highly specialized in study of media at empirical and theoretical level, occupied in the study of communication of politics in the mass media known as Osservatorio di Pavia. The authors experiment done by Knowledge Mining System which is used by security related agencies and institution of government in Italy to control information contained Web Mining and OSINT.

Wilson *et al.*, June 2002 said that the methodologies for automatic sentiment analysis start with a big set of terms noted with their respective polarity. The main purpose of this study is to easily differentiate between contextual and prior polarity, with prior knowledge of understanding which the necessary features for this task are. The experiment covers the feature performance for multiple algorithms of machine learning. Except one algorithm, features when combined together gives the best results. The evaluations shows that when natural instances are present the performance of features degraded on great pace. The authors suggested that indicating features that described more complex interdependencies between polarity clues can be considered as future research work.

God bole *et al.*, january2002 proposed a system which contains phase of identification sentiment in which for a particular topic which displays some opinions and scoring phase and a sentiment aggregation that will scores relative entities in the same class. At last the authors investigates the importance of scoring methods over big dataset of blogs and news. The authors interested in the fact that sentiments can vary according to the geographic location, news source or demographic group. As future work the authors are studying in evaluating the extent to which we predict the changes of future in behaviour of market or popularity.

Benamara *et al.*, march 2000 stated that most of the work done in past is finding the strength of subjective statements within a document or expressions uses the special part of speech such as nouns, verbs and adjectives. The authors said that until their contribution there was not a single related to adverbs in sentiment analysis nor use of adverb-adjective combinations (AACs) in sentiment analysis. The authors proposed a sentiment analysis method which is based on AACs which uses a linguistic evaluation of degrees of adverbs. The authors explained the experimental results on dataset of 200 news articles and compares the proposed technique with existing techniques of sentiment analysis. Based on Pearson correlation with human objects their experimental results gives higher accuracy.

Boyd and Ellison, March 2005 stated that social networking sites (SNSs) are regularly seeking the attention of industry and academic researchers fascinated by their reach and affordance. The authors described in the introductory article the functions of SNSs and

introduce a complete definition. The authors presented an aspect on the history of such sites, explaining development and key changes. The authors finally concluded that the condition is changing drastically and people should aware of which sites is using and why and for what purposes, especially other countries than U.S.

Agarwal *et al.*, *february2006* performed sentiment analysis on twitter data. The authors proposed functions polarity prior POS- specific and studied the usage of a tree kernel to prevent the necessity for hectic feature engineering. The tree kernel and the new functions performed approximately at the same level both surpassing the state of the art baseline. The authors concluded that for twitter data sentiment analysis is not that different as sentiment analysis for different genres.

Nasukawa and Yi, April 2007 proposed an approach for sentiment analysis to find sentiments connected with negative or positive polarities from a document for specific subject, rather than classifying the whole document into negative or positive. The major problems in sentiment analysis are whether the statements points negative or positive behaviour towards the subject and to be found how sentiment are described in texts. The authors stated that it is essential to clearly find out the semantic relationships between the subject and the sentiment expressions to increase the accuracy for the analysis of sentiment. In order to identify the sentiments in news articles and web pages, their proposed system obtained high precision of 75-95%.

Wang *et al.*, march 2008 proposed a system in U.S. elections 2012 for presidential candidates using real-time evaluation of sentiment on online microblogging site twitter. In order to collect the poll data the traditional analysis of election takes much time, but with the help of this system it takes data from more people with help of twitter, a microblogging service. It helps the social people like scholars, media and politician to broadcast their future perspective of the public opinion and electoral process. The authors finally concluded that the system and approach are generic, and should be adopted easily and spread across various other domains.

Wilson *et al.* march 2009 presented a method which first describes the whether a statement is polar or neutral to phrase-level sentiment evaluation and then ascertain the polarity of polar statements. By applying this methodology, the system is capable to identify automatically the contextual polarity of sentiment statements for huge subsets, obtaining results which are greater than baseline.

Kanayama and Nasukawa, November 2000 proposed an unsupervised lexicon building approach which detected the clauses of polar that grant negative or positive effect in a particular domain. The entries that are lexical in nature to be received are called polar atoms, the lesser human-recognizable semantic models that justify clause polarity. By the usage of precision and overall density of consistency in the dataset, the statistical approximation selects necessary polar atoms through candidates, without change in the threshold values. The obtained result shows that the applied method is robust enough for datasets with different domains and also for weight of initial lexicon and the precision of polarity report from the automatically received lexicon was on average of 94%.

Choi and Cardio October 2003 studied that the essential cooperation in event of compositional semantics and presents a learning based technique that connects structural assumption by compositional semantics for learning method. The authors conducted

experiments that shows compositional semantics based on natural heuristics that can outperform the learning based techniques which does not integrate compositional semantics, whereas a technique which consolidate semantics compositional onto learning which is greater than other all alternatives. The authors also studied that for describing expression-level polarity, content word negator plays an important role. Finally the authors concluded that accuracy of classification of expression level linearly decreases as context that is gradually determined.

Melville *et al.*, May 2016 presented a uniformed framework with respect to world-class associations using background lexical information and improve the information by using one of the available training examples to a particular domain. Experimental results shows that the authors methodology better performs than using training data or background knowledge within separation and text classification with lexical knowledge using to optional methodology. The authors concluded that they made two contributions. Firstly, they described a uniformed framework for combining knowledge of lexical for categorization of text in supervised learning and secondly, successfully applied the described methodology to analysis of classification of sentiment.

Paltoglou and Thelwall, may 2003 stated that a large number of sentiment analysis methodologies have used support vector machines as their baselines with the weights of binary unigram. The authors in this paper explored if there is any reliable feature weighted schemes which can improve accuracy of classification with the help of retrieving the information. The authors shows that alternatives of the *tf.idf* scheme gives notable increase in accuracy for sentiment analysis, with the use of sub linear function for smoothing of document frequency and term frequency weights. The methodology was tested on large data set and obtained highest accuracy.

Fernandez *et al.*, January 2004 developed a system introduced for the Subtask B Sentiment analysis of twitter i.e. SemEval 2014 task9. The authors system comprises of supervised methodology using techniques of machine learning, which using the text in dataset as features. This work is totally independent of any external resources and knowledge. The originality of author methodology depends on the use of skipgrams, n-grams and words as features. In the experimental study, it is clearly proves that skipgram shows better results than the ngram or word for the given datasets.

Mullen and Malouf January 2005 developed initial tests of statistics on a fresh datasets postings of group of political discussion that indicates the post that made response direct to post of others that having a greater likelihood that presents the perspective of opposing politics that of original post. The authors concluded that's the approaches of traditional text classifications is insufficient for this task in this dataset of sentiment analysis and the improvement can be made by utilizing information about how posters cooperate with one another.

Harb *et al.*, February 2006 stated that the previous approaches until this paper were written suffered from drawback i.e. for a particular topic either the adjective is not available or from another topic it meaning is different. The authors proposed a new methodology which consists of two steps. Firstly, for a particular topic the authors extract a learning dataset from the internet. Secondly the authors extracting from the dataset, they made two classes that are negative and positive adjectives with respect to the topic. The experimental study on the real

dataset shows the importance of authors' methodology. The experiments are performed on dataset that are cinema reviews and blogs shows that with the author methodology, it is easy to extract the desired adjectives for a particular topic.

Kim and Hovy may 2007 stated that the identification of a sentiment was challenging problem. The authors developed a system for a particular topic it automatically search the users who posts their views on that topic and the sentiment of each views. The systems consists a module for describing sentiment of a word and other for merging the sentiments into a statement. The authors did experiment with different models of classification and merging the sentiment at sentence and word level, given better results. For the improvement of recognition of Holder, the authors are using parser to attach areas that are more reliable with Holders. The learning techniques that are used is this system are support vector machines and decision list.

Martalo *et al.* march 2008 investigated how factors that are affective impact on the dialogue patterns and whether this impact may be explained and identified by Hidden Markov Models (HMMs). The goal of the authors is to study the chance of applying this model to classify behaviour of users for the purposes of adaptation. The authors obtained the initial results of their research and present a debate of problems that are open. With the help of the results, the author claims that the complicated interaction between the pragmatic level and the acoustic level comprises an important facet of emotions contained in voice expressions.

Daoud, October 2009 proposed a classifier and the introduced classifier contains four components which are AdaBoost which is a piece of an algorithm, Bayesian neural network, support vector machine and a technique for feature selection that is Signal-to-Noise. To confirm the efficiency of introduced classifier, the authors applied seven traditional classifiers to four datasets. The experimental study shows that applying the introduced classifier increases the rates of classification for all datasets. The author stated that SVMs key features are the control over capacity attained by margin optimization, sparseness of solution, the lack of local minima and the usage of kernels.

Yessenov and Misailovic march 2000 presents study of effectiveness of techniques of machine learning in text message classification by semantic meaning. The authors use comments of movie reviews from Digg that is social network which is popular as authors dataset and text classification can be done by negative or positive and objectivity or subjectivity attitude. The authors suggested different methodologies in text feature extraction such as using knowledge of WordNet synonyms, bounding word frequencies by threshold, handling negations, restricting to adjectives and adverbs, using large movie review corpus and a bag-of-words model. The authors analyse their performance on accuracy using four methodologies of machine learning that are K-Means clustering, Maximum Entropy, Decision Trees and Naïve Bayes. Finally, the authors concluded that bag-of-words model perform better than relative models.

Kang *et al.* June 2001 stated that the senti-lexicon existed does not properly adopt the word sentiment used in the restaurant review. The author introduced a senti-lexicon of restaurant reviews for the sentiment analysis. Using supervised learning technique a review document is classified as negative sentiment and positive sentiment, hence there is chance for the accuracy of positive classification to greater than 10% than the accuracy of negative classification. The author also introduced an improved version of Naïve Bayes to deal with these types of problems. The authors improved Naïve Bayes had managed to low the gap between positive

and negative accuracy by 3.8% when applied with unigram + bigram and 28.5% when compared with SVM.

Afroze Ibrahim Baqapuri, 2012. (Twitter Sentiment Analysis).

Citing from other similar works that has been carried out in sentiment analysis (We can deduce some key concept about sentiment analysis.

The bag-of-words is one of the most widely used feature model for almost all text classification task due to its simplicity coupled with good performance. The model represents the text to be classified as a bag or collection of individual word with no link or dependence of one word with the other, i.e. it completely disregards grammar and order of words within the text. This model is also very popular in sentiment analysis and has been used by various researchers. The simplest way to incorporate this model in our classifier is by using unigrams as features.

Generally speaking, n-gram is a contiguous sequence of "n" words in our text, which is completely independent of any other words or grams in the text. So inigrams are just a collection of individual words in the text to be classified and we assume that the probability of occurrence of one word will not be affected by the presence or absence of any other word in the text, This is a very simplifying assumption but it has been shown to provide rather good performance.

one simple way to use unigrams as features is to assign them with a certain prior polarity and take the average of the overall polarity of the text, where the overall polarity of the text could simply be calculated by summing the prior polarities of individual unigrams. Prior polarity of the word would be positive if the word is generally used as an indication of positivity, for example the word "sweet"; while it would be negative if the word is generally associated with negative connotations, for example "evil". There can also be degrees of polarity in the model, which means how much indicative is that word for that particular class. A word like "awesome" would probably have strong subjective polarity along with positivity, while the word "decent" would although have positive prior polarity but probably with weak subjectivity.

There are three ways of using prior polarity of words as features. The simpler un-suspensive approach is to use publicly available online lexicons/dictionaries which map a word to its prior polarity. The Multi-Perspective-Question-Answering (MPQA) is an online resource with such a subjectivity lexicon which maps a total of 4,850 words according to whether they are "positive" or "negative" and whether they have "strong" or "weak" subjectivity. The SentiWordNet 3.0 is another such resource which gives probability of each word belonging to positive, negative and neutral classes. The second approach is to construct a custom prior polarity dictionary from our training data according to the occurrence of each word in each particular class. For example, if a certain word is occurring more often in the positive labelled phrase in our training dataset (as compared to other classes) then we can calculate the probability of that word belonging to positive class to be higher than the probability of occurring in any other class.

This approach has been shown to give better performance, since the prior polarity of words is more suited and fitted to a particular type of text and is not very general like in the former approach. However, the latter is a supervised approach because the training data has to be labelled in the appropriate classes before it is possible to calculate the relative occurrence of a word in each of the class.

Kouloumpis et al. noted a decrease in performance by using the lexicon word features along with custom n-gram word features constructed from the training data, as opposed to when the n-gram were used alone.

The third approach is middle ground between the above two approaches. In this approach we construct our own polarity lexicon but not necessarily from our training data, so we don't need to have labelled training data. One way of doing this as proposed by turnkey et al, is to calculate the prior semantic orientation (polarity) of a word or phrase by calculating its mutual information with the word "excellent" and subtracting the result with the mutual information of that word or phrase with the word "poor". They used the number of result hit counts from online search engines of a relevant query to compute the mutual information. The final formula they used is as follows

$$\text{Hit (phrase NEAT excellent).hits} * \text{"poor"}$$

$$\text{Polarity (phrase)} = \log_2 \text{hits}$$

$$(\text{Phrase NEAR poor}).\text{hits ("excellent")}$$

Where hits (phrase NEAR "excellent") means the number documents returned by the search online in which the phrase (whose polarity is to be calculated) and word "excellent" are co-occurring while hits ("excellent") means the number of documents returned which contain the word "excellent". prabowo et al. have gone ahead with this idea and used a seed of 120 positive words and 120 negative to perform the internet search [12]. So the overall semantic orientation of the word under consideration can be found by calculating the closeness of that word with each one of the seed words and taking an average of it.

Another graphical way of calculating polarity of objective has been discussed by Hatzivassiloglou et al. The process involves first identifying all conjunctions of adjectives from the corpus and using a supervised algorithm to mark every pair of adjectives as belonging to the same semantic orientation or different. A graph is constructed in which the nodes are the adjectives and links indicate same or different. A graph is constructed in which the nodes are the adjectives and links indicate same or different semantic orientation. Finally, a clustering algorithm is applied which decides the graph into two subsets such that nodes within a subset mainly contain links of same orientation and links between the two subsets mainly contain links of same orientation. One of the subsets would contain links of different orientation. One of the subsets would contain positive adjectives and the other would contain negative.

Many of the researchers in this field have used already constructed publicly available lexicons of sentiment bearing words, while many others have also explored building their own prior polarity lexicons.

The basic problem with the approach of prior polarity approach has been identified by Wilson et al. who distinguish between prior polarity and contextual polarity. They say that the prior polarity of a word may in fact be different from the way the word has been used in the particular context.

The paper presented the following phrase as an example: Philip clapp, president of the National Environment Trust. Sums up well the general thrust of the reaction of environmental movements: "There is no reason at all to believe that the polluters are suddenly going to become reasonable."

In this example all of the four underline words "trust:", "well", "reason", and "reasonable" have positive polarities when observed without context to the phrase, but here they are not being used to express a positive sentiment. This concludes that even though generally speaking a word like "be" used in positive sentences, but this doesn't rule out the chances of it appearing in non-positive sentences as well. Henceforth prior polarities of individual words (WHETHER THE WORDS GENERALLY CARRY POSITIVE OR NEGATIVE CONNOTATIONS) may alone not enough for the problem. The paper explores some other features which include grammar and syntactical relationships between words to make their classifier better at judging the contextual polarity of the phrase.

The task of twitter sentiment analysis can be most closely related to phrase level sentiment analysis.

A seminal paper on phrase level sentiment analysis was presented in 2005 by Wilson et al., which identified a new approach to the problem by first classifying phrases according to subjectivity (polar) and objectivity (neutral) and then further classifying the subjective - classified phrases as either positive or negative. The paper noticed that many of the objective phrases used prior sentiment bearing words in them, which led to poor classification of especially objective phrases. It claims that if we use a simple classifier which assumes that the contextual polarity of the word is merely equal to its prior polarity gives a result of about 48%. The novel classification process proposed by this paper along with the list of ingenious features which include information about contextual polarity process.

The most commonly used classification techniques are the Naive Bayes Classifier and State Vectors

Machine. Some researchers like Barbosa et al. publish better result for SVMs while other like Park et al. support Naive Bayes.

It has been observed that having a larger training sample pays off to a certain degree, after which accuracy of the classifier stays almost constant even if we keep adding more labelled tweets in the training data. Barbosa et al. used tweets labelled by internet resources, instead of labelling them by hand, for training the classifier. Although there is loss of accuracy of the labelled samples in doing so (which is modelled as increase in noise) but it has been observed that if the accuracy of training labels is greater than 50% the more the labels. The higher the accuracy of the resulting classifier. So in this way if there are extremely large number of tweets,

the fact that our labels are noisy and inaccurate can be compensated for. On the other hand, Pak et al and Go et al. use presence of positive or negative emoticons to assign labels to the tweets like in the above case they used large number of tweets to reduce effect of noise in their training data.

Some of the earliest work in this field classified text as positive or negative, assuming that all the data provided is subjective, While this is a good assumption for something like movie reviews but when analyzing tweets and blogs there is a lot of objective text we have to consider, so incorporating neutral class into the classification process is now becoming a norm. Some of the work which has included neutral class into their classification process.

There has been very recent research of classifying tweets according to the mood expressed in them, which goes one step further. Bollen et al. explores this area and develops a technique to classify tweets into six distinct moods: tension, depression, anger, vigor, fatigue, and confusion. They use an extended version of profile of mood states (POMS): a widely accepted psychometric instrument. They generate a word dictionary and assign them weights corresponding to these six dimensions. However not much detail has been provided into how they built their customized lexicon and technique did they use for classification.

2.4 Partial conclusion

The aim of this chapter was to give a literature review of sentiment analysis especially on twitter data. In this chapter, the description of the method through which data is retrieved from twitter through the tweeter API. Making sentiment analysis on tweets is challenging from the natural language processing perspective, but also in terms of performance when huge amounts of tweets should be processed. Both challenges are addressed in this chapter. Also, there are two main approaches used to carry out sentiment analysis. In addition, there are various sentiment analysis level and the sentiment analysis process was also explained in this chapter. The research on sentiment analysis has been going for a long time. Sentiment analysis in present days becomes the major issue in field of research and technology. Due to day by day increase in the number of users on the social networking websites, huge amount of data produced in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find whether the sentiment is negative or positive.

CHAPTER THREE: ANALYSIS AND DESIGN

3.1 Introduction

This chapter is focused on the analysis and design of sentiment analysis using twitter. It presents the high level view of the system by beginning with the description of the software development life cycle used to be accomplished .its pros and cons ,the design using the unified modelling language and the description of the resolution. The system was design using the tools as describe in chapter 4. Several diagrams were modelled to ease the system development, these include: Use case Diagram, Class diagram, State Chart Diagram, and Entity Relational Diagram.

3.2 Proposed Methodology

The software development life cycle I used is the agile methodology in the prototype model

A sprint is a period of time allocated for a particular phase of a project. Sprints are considered to be complete when the time period expires. There may be disagreements among the members of the team as to whether or not the development is satisfactory; however, there will be no more work on that particular phase of the project. The remaining phases of the project will continue to develop within their respective time frames.

3.2.1 Definition of agile methodology

The Agile Method and methodology is a particular approach to project management that is utilized in software development. This method assists teams in responding to the unpredictability of constructing software. It uses incremental, iterative work sequences that are commonly known as sprints.it refers to a group of software development methodologies based on iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams.

3.2.2The General Principles of the Agile Method

- Satisfy the client and continually develop software.
- Changing requirements are embraced for the client's competitive advantage.
- Concentrate on delivering working software frequently. Delivery preference will be placed on the shortest possible time span.
- Developers and business people must work together throughout the entire project.
- Projects must be based on people who are motivated. Give them the proper environment and the support that they need. They should be trusted to get their jobs done.
- Face-to-face communication is the best way to transfer information to and from a team.
- Working software is the primary measurement of progress.

- Agile processes will promote development that is sustainable. Sponsors, developers, and users should be able to maintain an indefinite, constant pace.
- Constant attention to technical excellence and good design will enhance agility.
- Simplicity is considered to be the art of maximizing the work that is not done, and it is essential.
- Self-organized teams usually create the best designs.
- At regular intervals, the team will reflect on how to become more effective, and they will tune and adjust their behaviour accordingly.

History of Agile Method

Many of the agile ideas surfaced in the 1970s. Studies and reviews were conducted on the Agile Method that explains its emergence as a reaction against traditional approaches to project development.

In 1970, Dr. William Royce published a paper that discussed the managing and developing of large software systems. The paper outlined his specific ideas about sequential development. His presentation stated that a project could be developed much like a product on an assembly line. Each phase of the development had to be complete before the next phase could begin. The idea required that all developers must first put together all of the requirements of a project. The next step was to complete all of its architecture and designs. This is followed by writing the code. The sequences continue in complete increments. As these steps are completed, there is little or no contact between specialized groups that complete each phase of the project.

Pioneers of the Agile Method believed that if developers studied the process, they would find it to be the most logical and useful solution to software development.

3.3.3. Companies that Use the Agile Method

Although there is no official list of companies that use the Agile Method for their projects, IBM is one of the companies that openly uses this method to develop software. Many companies will adopt the use of this method within their development structure, but they aren't always open about their choice to use it.

According to IBM, the use of the Agile Method means that significant organizational changes will take place. They believe that many agile software development teams will increase their chances of success by partnering with a trusted guide. They help clients implement their own agile software development strategies for their projects. They provide critical guidance that will help agile software development teams to avoid common adoption, expansion, and implementation pitfalls.

3.3.4. Benefits of Using the Agile Method

The Agile Method grew out of the experience with the real-life projects of leading software professionals from the past. Because of this, the challenges and limitations of traditional development have been discarded. Subsequently, the Agile Method has been accepted by the industry as a better solution to project development. Nearly every software developer has used the Agile Method in some form.

This method offers a light framework for assisting teams. It helps them function and maintain focus on rapid delivery. This focus assists capable organizations in reducing the overall risks associated with software development.

The Agile Method ensures that value is optimized throughout the development process. The use of iterative planning and feedback results in teams that can continuously align a delivered product that reflects the desired needs of a client. It easily adapts to changing requirements throughout the process by measuring and evaluating the status of a project. The measuring and evaluating allows accurate and early visibility into the progress of each project.

It could be stated that the Agile Method helps companies build the right product. Instead of trying to market software before it is written, the Agile Method empowers teams to optimize the release during its development. This allows the product to be as competitive as possible within the marketplace. It preserves the relevance of the critical market, and it ensures that a team's work doesn't wind up collecting dust on a shelf. This is why the Agile Method is an attractive developmental option for stakeholders and developers alike.

There are many critics of the Agile Method; however, this method produces results that clients can take to the bank. Although a project may not turn out exactly as the client envisions, it will be delivered within the time that it needs to be produced. Throughout the process, the client and the team are changing the requirements in order to produce the quality needed by the client. Clients are happy with the results, and the team satisfies the client's needs. The ongoing change can sometimes give both the client and the team more than they had originally envisioned for the product. The Agile Method really is a winning solution for everyone involved in software development.

3.3.5. CRITICISM OF AGILE DEVELOPMENT

- It is developer-centric rather than user-centric.
- Agile focuses on processes for getting requirements and developing code and does not focus on product design.
- Agile methodologies can also be inefficient in large organizations and certain types of projects.

3.3.6. DIFFERENCE BETWEEN AGILE AND TRADITIONAL (WATERFALL OR SPIRAL) DEVELOPMENT

Fundamental Assumptions

Traditional: Systems are fully specifiable, predictable, and can be built through meticulous and extensive planning.

Agile: High-quality, adaptive software can be developed by small teams using the principles of continuous design improvement and testing based on rapid feedback and change.

Control

Traditional: Process-centric

Agile: People-centric

Management Style

Traditional: Command-and-control

Agile: Leadership-and-collaboration

Knowledge Management

Traditional: Explicit

Agile: Tacit

Role Assignment

Traditional: Individual—favors specialization

Agile: Self-organizing teams—encourages role interchangeability

Communication

Traditional: Formal

Agile: Informal

Customer's Role

Traditional: Important

Agile: Critical

Project Cycle

Traditional: Guided by tasks or activities

Agile: Guided by product features

Development Model

Traditional: Life cycle model (Waterfall, Spiral, or some variation)

Agile: The evolutionary-delivery model

Desired Organizational Form/Structure

Traditional: Mechanistic

Agile: Organic

Technology

Traditional:	No	restriction
Agile:	Favors object-oriented technology	

3.3.7. AGILE METHODOLOGY GLOSSARY

Acceptance Test: An acceptance test confirms that a story is complete by matching a user action scenario with the desired outcome. Acceptance testing is also called beta testing, application testing, and end user testing.

Customer: A customer is a person with an understanding of both the business needs and operational constraints for a project who provides guidance during development.

Domain Model: The domain model is the application domain responsible for creating a shared language between business and IT.

Iteration: An iteration is a single development cycle, usually measured as one week or two weeks.

Planning Board: A planning board is used to track the progress of an agile development project. After iteration planning, stories are written on cards and pinned up in priority order on a planning board.

Planning Game: A planning game is a meeting attended by both IT and business teams that are focused on choosing stories for a release or iteration.

Release: A release is a deployable software package that is a culmination of several iterations of development.

Release Plan: An evolving flowchart that describes which features will be delivered in upcoming releases.

Spike: A spike is a story that cannot be estimated until a development team runs a time-boxed investigation.

Stand-up: This is a daily progress meeting; literally everyone stands up and meets to keep engaged and motivated. A stand-up is traditionally held within a development area.

Story: A particular business need to be assigned to the software development team. Stories must be broken down into small enough components that they may be delivered in a single development iteration.

Timebox: A timebox is a defined period of time during which a task must be accomplished.

Velocity: The velocity is the budget of story units available for planning the next iteration of a development project. Velocity is based on measurements taken during previous iteration cycles.

Wiki: A wiki is a server program that allows users to collaborate in forming the content of a Web site.

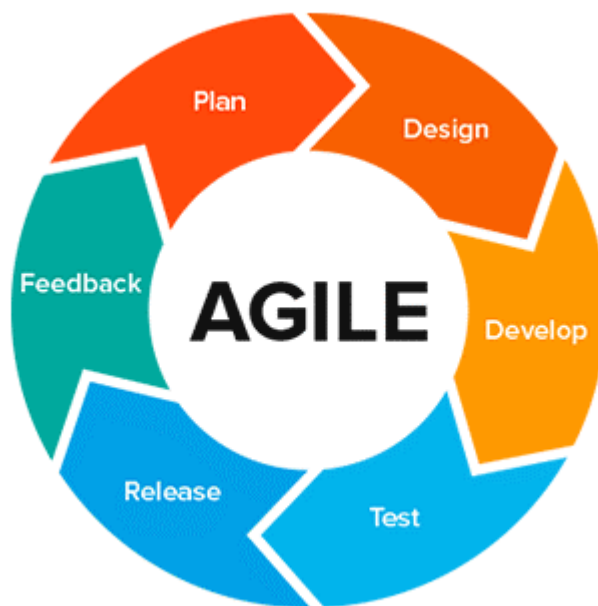


Figure 3: describing the different phases in agile development

Though I carried out this project alone I made sure I wrote my daily activities on sticky notes daily and evaluated myself at the end of each day.

I also pushed code to my github account weekly and made sure I asked people who have more experience in coding than me to check my progress weekly.

I also deployed my application in bits and so each version was a better prototype of the predecessor .and so since the project was a whole mystery for me especially at the beginning I had to implement the prototype model together with the agile methodology

3.4 BRIEF DESCRIPTION OF THE PROTOTYPE MODEL

The prototype model is one of the most popularly used Software Development Life Cycle **Models (SDLC models)**. This **model** is used when the customers do not know the exact project requirements beforehand. Once the customer figures out the problems, the **prototype** is further refined to eliminate them.

We see a lot of processes repeated for both the agile and the prototype models so the only time I used the prototype was for version and their noticed changes.

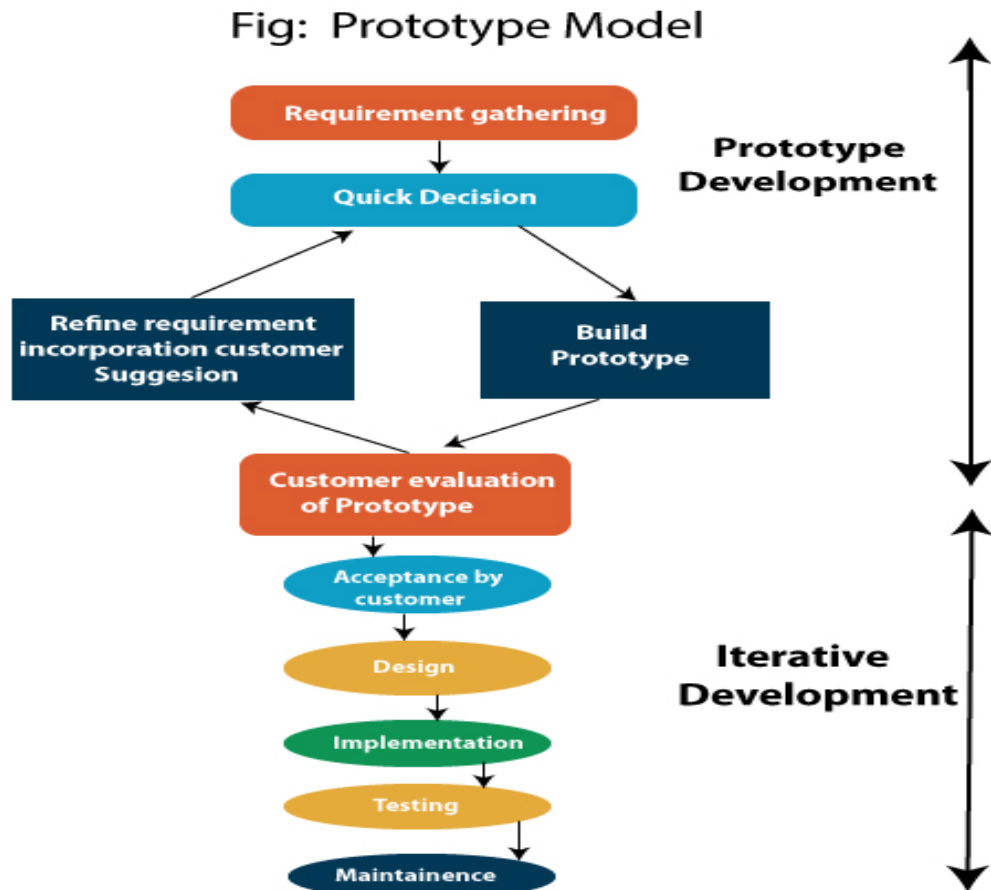


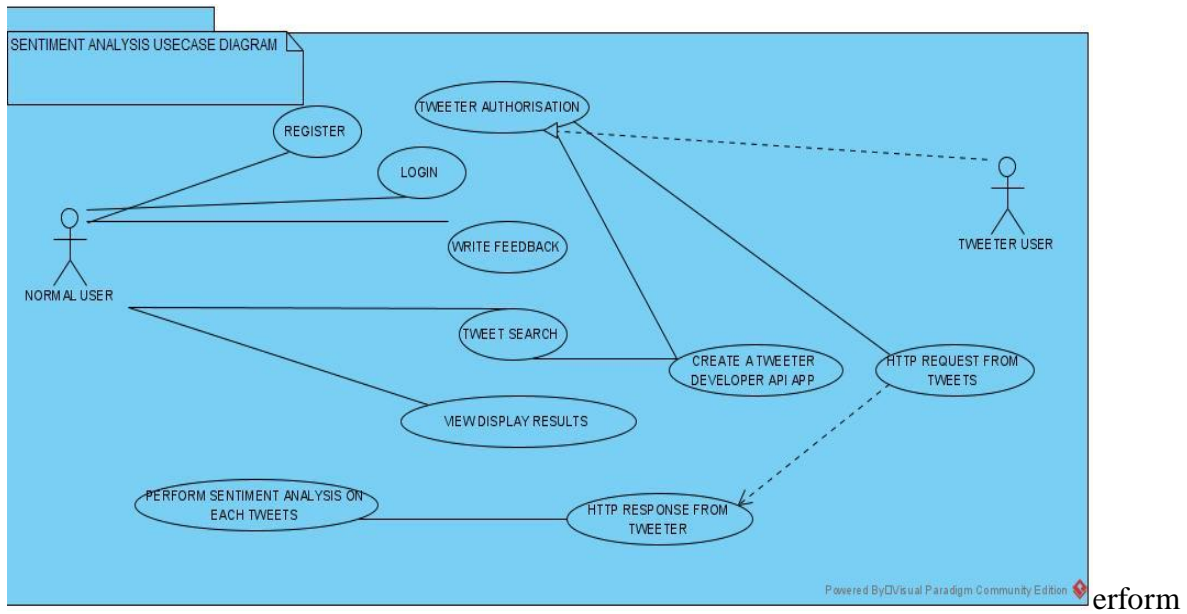
Figure 4 : The prototype model

3.5 System Design and Architecture

For the design I made use of the universal unified modelling language where I designed the following diagrams.

3.5.1. Use Case Diagram

Use-case diagrams usually referred to as behaviour diagrams is use here to describe a set of actions (use cases) that the system should or can p



in collaboration with one or more external users of the system (actors).

Figure 5 usecase diagram for twitter sentiment analysis

3.5.2 System Flow Diagram

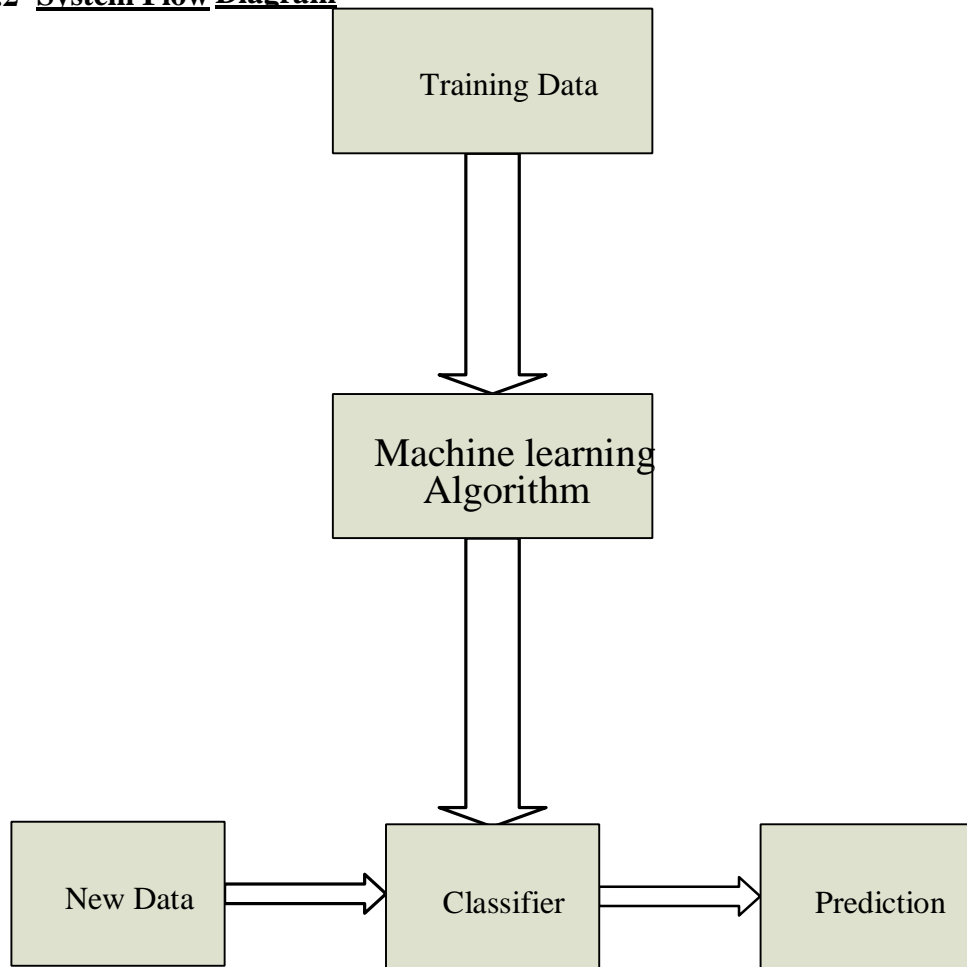


Figure 6 flow chat diagram for twitter sentiment analysis

5.4 Class Diagram

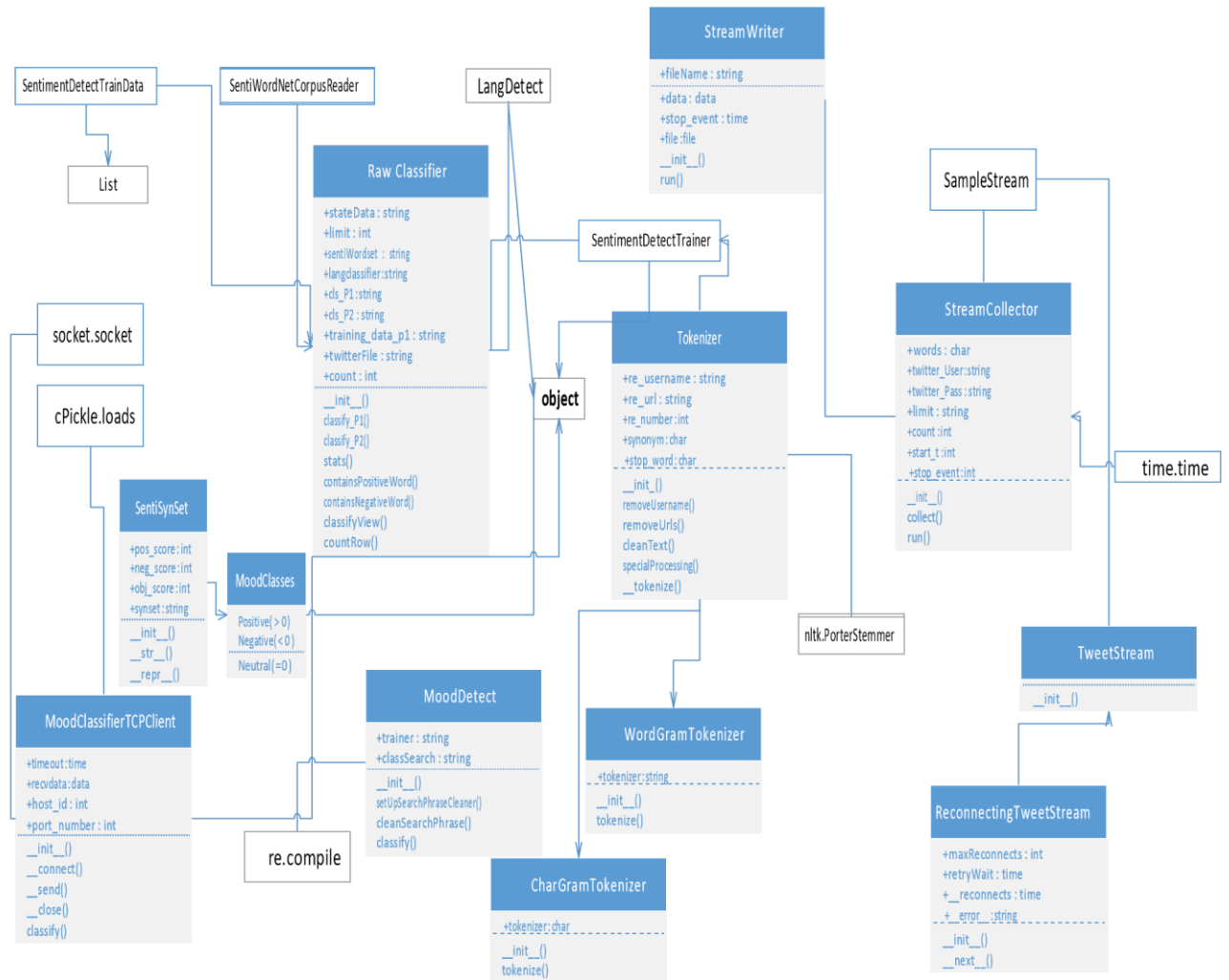


Figure 7 Class Diagram for Sentiment Analysis using Twitter API and Big Data

3.5.4 Activity Diagram

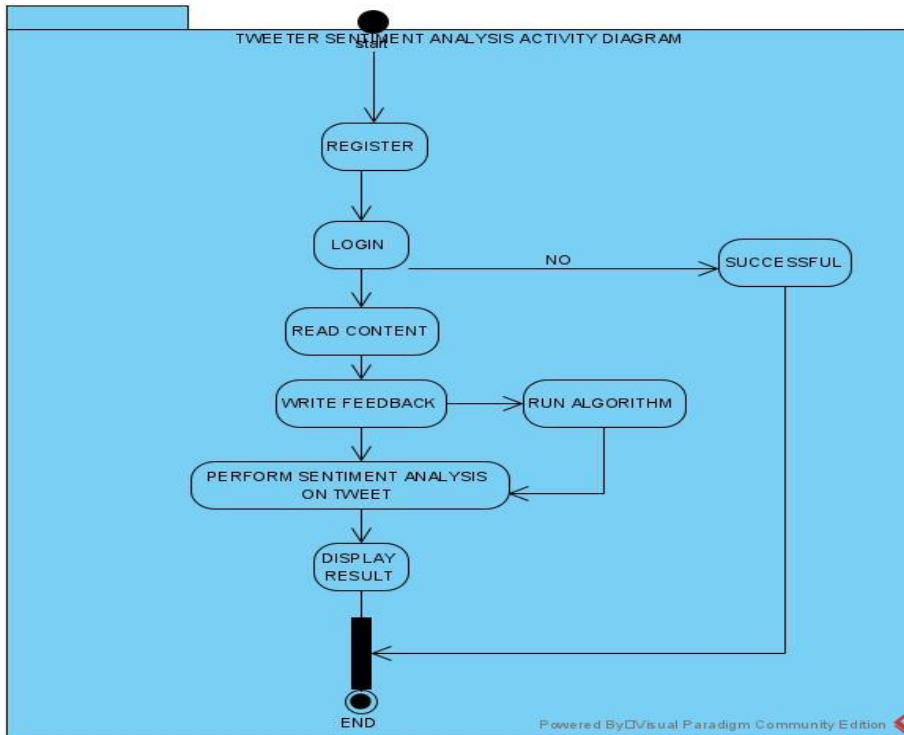


Figure 8 Activity diagram for twitter sentiment analysis

3.5.5. Data Flow Diagram

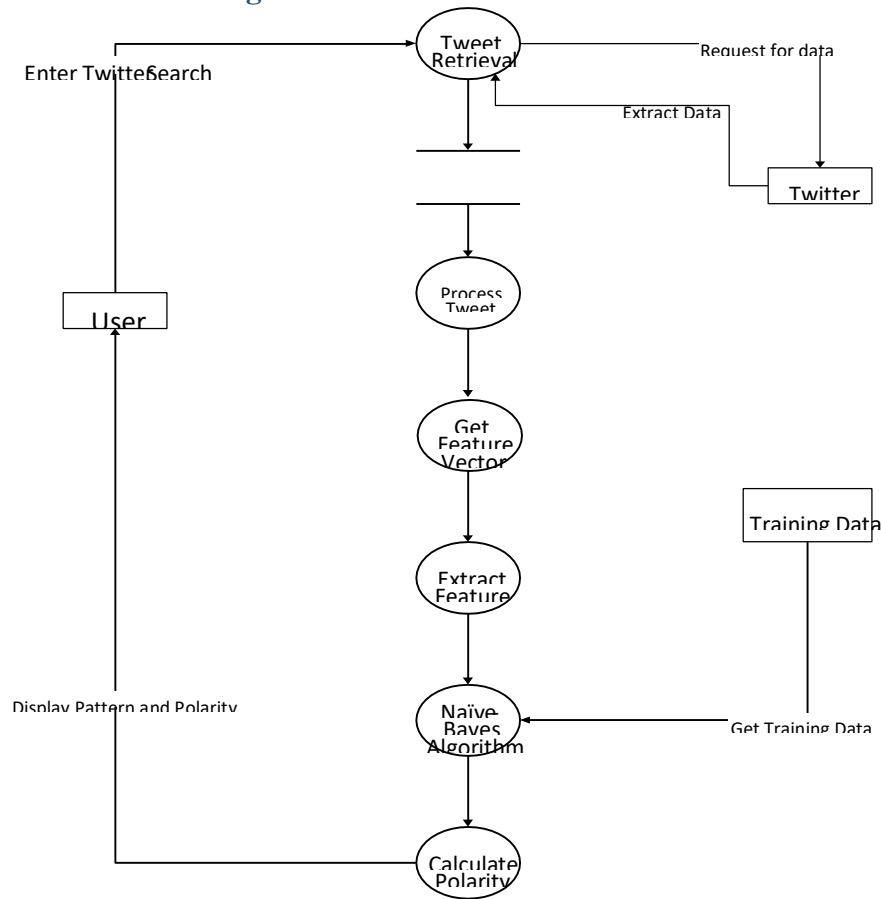


Figure 9 twitter sentiment Analysis data flow diagram

3.5.6 Flow Chart

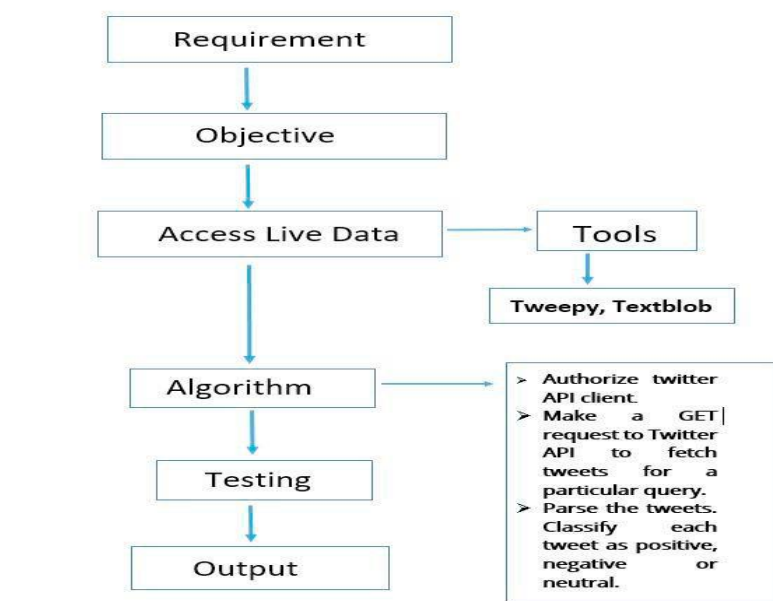


Figure 10 twitter sentiment analysis flow chat.

3.5.7. SEQUENCE DIGRAM

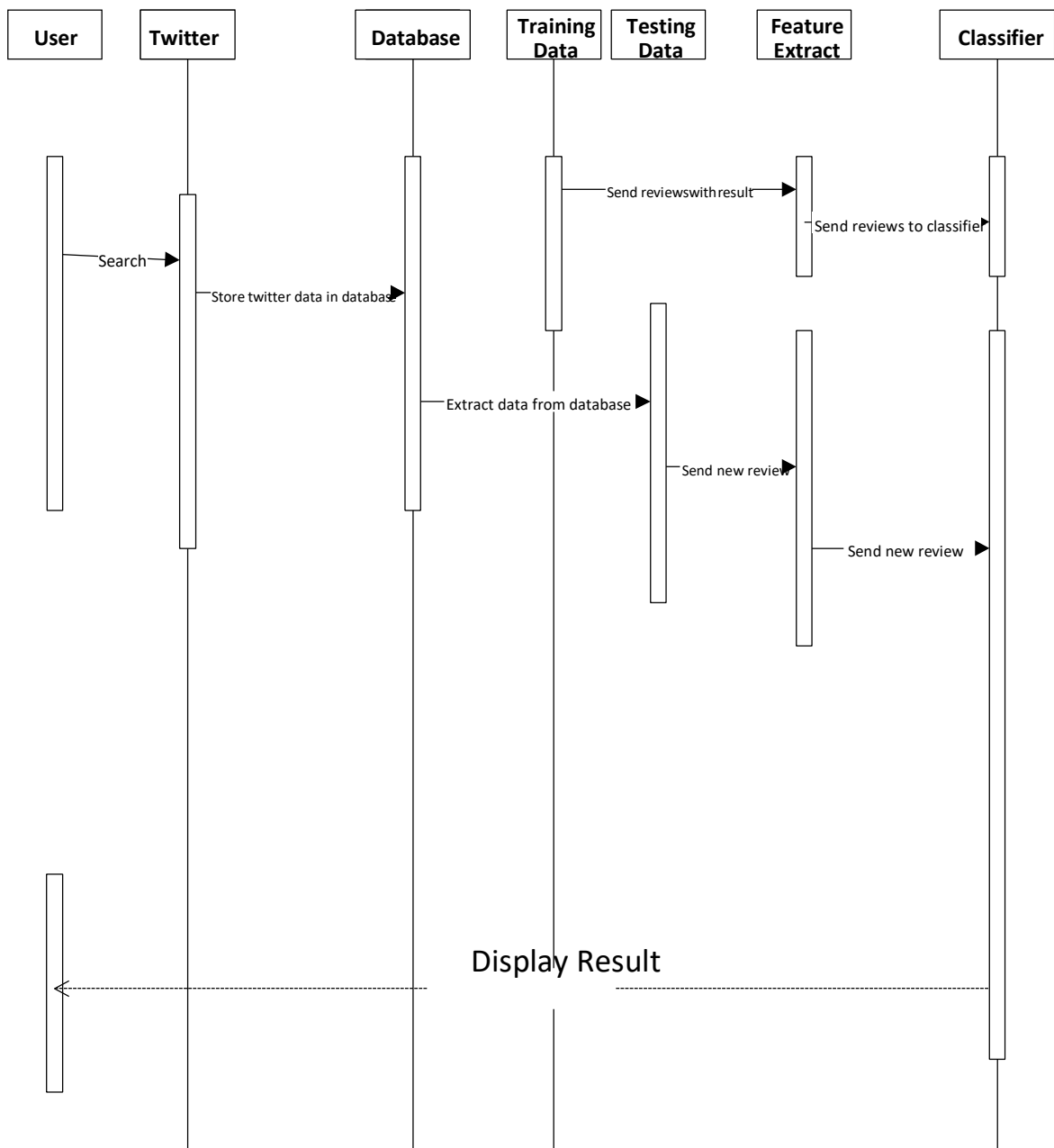


Figure 11 twitter life streaming sentiment analysis sequence diagram

3.6. GLOBAL ARCHITECTURE OF THE SYSTEM

This section is a brief discussion on how I was able to use my chosen methodology which is the agile methodology to get my solution up and running.

During the development of this application, after choosing the methodology to be used I knew I had to start work immediately:

3.6.1 PLAN

During this phase I went and saw my supervisor who further explained my project and its requirements to me in details and he asked me to write and submit a project proposal so he can find out if I actually understand what am supposed to do .I actually did it and he approved and told me to commence work.

3.6.2 DESIGN

I then moved on to doing more research and then started designing my UML diagrams using the visual paradigm for community edition.

After designing the class diagram, use case diagram, activity diagram, sequence diagram and flow chat diagrams, I moved to the next phase which is the implementation phase.

3.6.3 IMPLEMENTATION

Here I created a github repository so I could be doing self -follow-up by pushing code each time I work on the project.

And I had to look for people good in machine learning and interview them so I get a wider explanation of what am to do.

I also structured the work and saw that since the application was in 3 parts I had to begin with the most difficult which is training a model using natural language processing.

So I trained the model using a couple of algorithms and after I was done I went on to creating a tweeter developer account which I was granted access to after a few days .I then created my twitter application so as to get access to access to the twitter developers Application programming interface(API)

After that I went on to my backend where I was able to get tweets from twitter where I performed sentiment analysis on them and classified them using positive and negative. And I came up with the graph.

I then moved on to my front end where I built the UI and filled it with lots of beautiful content and made sure to leave a form for users to fill and so I use those fillings to get the sentiments in the text.

After that linking everything was what I did.

3.6.4 TESTING

After doing all the work I had to test the application so I carried out the following types of tests on this application

Unit testing

Here I successfully tested each individual functionality of the application.

Integrations test

Here I had to combine these said functionalities and test its output.

3.6.5 RELEASING

After carrying out all this testing and more I had to release the application which I did by deploying it and giving my friends the link to go and check and leave a feedback.

3.6.6 FEEDBACK

I got a lot of feedback from my friends and colleagues where I used it as future works in this thesis.

3.7 DESCRIPTION OF THE RESOLUTION PROCESS

3.7.1 INTRODUCTION

When this research topic was given to me and after I had figured out how to go about it, I was normally supposed to choose a methodology and start working with respect to how that method unfolds but am new to machine learning and all its processes so I normally had to start with what I know and could start with easily.

3.7.2 DESCRIPTION

After receiving the project topic from the faculty of engineering and technology, university of buea, and attending a meeting with my supervisor, I had to gather requirements and so I started my research and came out with the system requirements and specification document where I later moved on to the implementation and then came to the design and after that I was able to test the application if it works the same way as I want it as specified in my design document

3.8. PARTIAL CONCLUSION

This chapter was the summary of all I had to do in other to make this project stand out better than all those stated in my literature review section.

Here I elaborated the methodology to be used which is the software development life cycle and I also talked about the cons and pros of the methodology. I also showed how the methodology was used to get the desired result.

CHAPTER FOUR: IMPLEMENTATION, REALIZATION AND PRESENTATION OF RESULTS

4.1 Introduction

This system explains in details the implementation phase of the methodology used as explained in chapter 3. It begins by presenting the functionality of the system, followed by the tools and materials used to implement the system. It then heads to the evaluation of the system and then a brief conclusion.

This system is basically getting tweets from the twitter API and classify them under positive or negative by studying the sentiment in them.

4.2 TOOLS AND MATERIALS USED

This project basically has three parts namely:

The backend

The front

The machine learning Model.

Hence the tools used at every part include:

FRONTEND			
FRAMEWORKS	PROGRAMMING LANGUAGES	LIBRARIES	OTHERS
ANGULAR 9	JAVASCRIPT	NODE PACKAGE MANAGER(NPM)	
BOOTSTRAP 4	HTML		
TYPESCRIPT	CSS		
BACKEND			
FLASK	PYTHON 3.9	NUMPY,MATPLOTLIB,TWEEPY,TEXT BLOB	TWITTER API
MACHINE LEARNING	PYTHON 3.9	PICKLE,SKLEARN,	NLTK

4.2.1 DEFINITION AND DESCRIPTION OF SOME MATERIALS USED.

1. **Angular** is a Typescript-based open-source web application framework led by the Angular Team at Google and by a community of individuals and corporations.

2. **Flask** is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries.

3. **Python** is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991.

4. **The Natural Language Toolkit**, or more commonly **NLTK**, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP). It is a suite of open source Python modules, data sets, and tutorials supporting research and development in Natural language processing.

5. **THE TWITTER API** lets you read and write **Twitter** data.

6. **TWEEPY** is an easy-to-use Python library for accessing the Twitter API.

7. **SCIKIT-LEARN** (formerly scikits. learn and also known as **sklearn**) is a free software machine learning library for the Python programming language.

8. **MATPLOTLIB** is a plotting library for the Python programming language and its numerical mathematics extension **NUMPY**. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.

5 DESCRIPTION OF THE IMPLEMENTATION PROCESS

Before I was able to come up with this application I did the following:

5.1. GETTING DATA FROM TWITTER STREAMING API

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements: Create a twitter account if you do not already have one.

Go to <https://apps.twitter.com/> and log in with your twitter credentials.

Click "Create New App"

Fill out the form, agree to the terms, and click "Create your Twitter application"

In the next page, click on "API keys" tab, and copy your "API key" and "API secret".

Scroll down and click "Create my access token", and copy your "Access token" and "Access token secret".

5.2. GETTING TWITTER API KEYS

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

- Create a twitter account if you do not already have one.
- Go to <https://apps.twitter.com/> and log in with your twitter credentials.
- Click "Create New App"
- Fill out the form, agree to the terms, and click "Create your Twitter application"
- In the next page, click on "API keys" tab, and copy your "API key" and "API secret".
- Scroll down and click "Create my access token", and copy your "Access token" and "Access token secret".

5.3. Connecting to Twitter Streaming API and downloading data

We will be using a Python library called Tweepy to connect to Twitter Streaming API and downloading the data. If you don't have Tweepy installed in your machine, go to this [link](#), and follow the installation instructions.

Next create, a file called `twitter_streaming.py`, and copy into it the code below. Make sure to enter your credentials into `access_token`, `access_token_secret`, `consumer key`, and `consumer secret`.

After getting these twitter API access keys we need to generate our tweets and then display it in a graph for our users to see.

5.4. PRESENTATION AND INTERPRETATION OF RESULTS

5.4.1. FRONTEND

Here is the code I wrote to get the beautiful user interface below for the enjoyment of all those who will wish to test sentiments of tweets using my application.

Figure 10 front end code

This is the user dashboard of the application where I wrote some interesting content for users to read and learn about telecommunication companies in Cameroon.

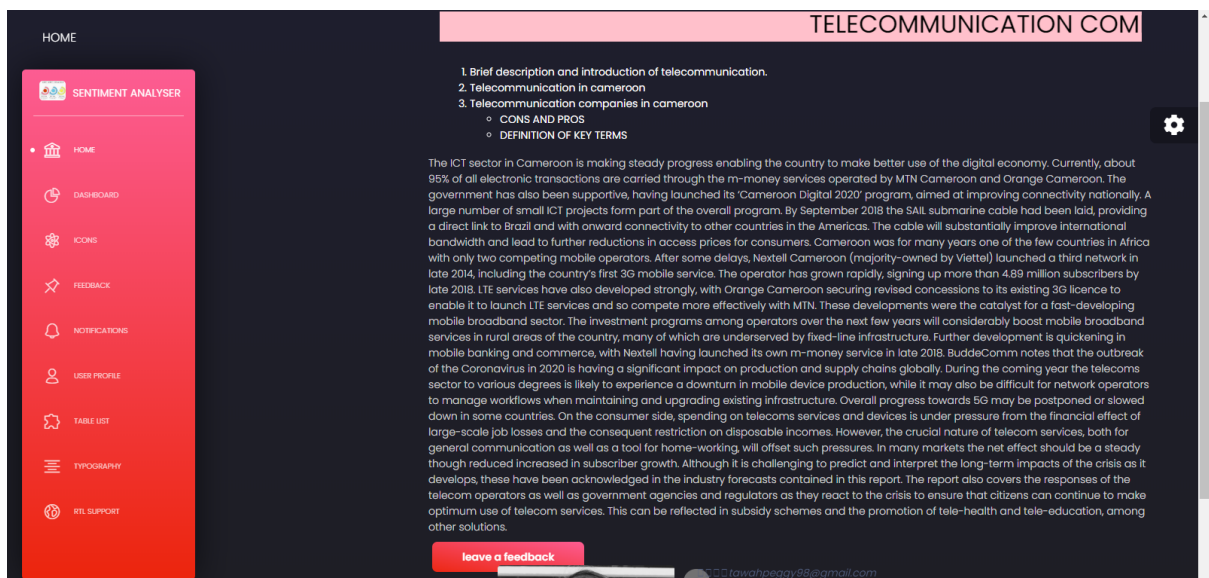


Figure 11 front end home page

The form below is where the user inputs text to get its sentiment.

FEEDBACK

SENTIMENT ANALYSER

HOME

DASHBOARD

ICONS

FEEDBACK

NOTIFICATIONS

USER PROFILE

TABLE LIST

TYPOGRAPHY

RTL SUPPORT

leave a feedback please

Email address

tawahpeggy98@email.com

Full Name

tawah peggy

Message

good job guys!

Save

About Sentiment Analyser

Lorem ipsum dolor sit amet consectetur adipisicing elit. Ut eum

Figure 12 front end feedback form

5.4.2 BACKEND

After training my model I used it to connect to twitter to get tweets with Cameroon In it and then I classified each tweet weather it is positive or negative.


```

1  from tweepy import Stream
2  from tweepy import OAuthHandler
3  from tweepy.streaming import StreamListener
4  import json
5  import dataset as s
6  # consumer key, consumer secret, access token, access secret.
7  ckey = "uyWQogbuzUaBwc5cl5pGi0zeR"
8  csecret = "zGmpLfNaBFLixWV2AJKR92zPKtJzAnPXZuUSIK46yyuQKUJTLz"
9  atoken = "1262338809011802112-LM9aM8Tyj5FBKV9VK53Z6QBcDI3lt4"
10 asecret = "PFWUh07fAvkuZm00dGC5TATmEEUWmSK7n7KSjcr47tDDh"
11
12
13 class listener(StreamListener):
14
15     def on_data(self, data):
16         try:
17             all_data = json.loads(data)
18
19             tweet = all_data["text"]
20             sentiment_value, confidence = s.sentiment(tweet)
21             print(tweet, sentiment_value, confidence)
22
23             if confidence * 100 >= 80:
24                 output = open("twitter-out.txt", "a")
25                 output.write(sentiment_value)
26                 output.write('\n')
27                 output.close()

```

Figure 13 backend code

Here are few of the tweets I was able to get.

```

Obama Quotes FDR Saying We Are All Descended from Immigrants and Revolutionists
http://t.co/bPSBKglEu2 via @politicususa neg 0.8
TV watcher! Obama has never struck me as even "book smart", let alone "wise". T
V Watcher is perfect. @janevonmises https://t.co/QqQ6yGkY9e pos 1.0
RT @briaatortillaa: Let's take tax and delivery charges off pizza @Obama @Congre
ss neg 0.8
RT @TACM_Literature: http://t.co/tnFAucxVEO
http://t.co/1vLDpTbY8k

#AntiChrist #Treason #Obama #America #USA #wakeupAmerica #Wakeupnow htt... neg 0.8
BOOK: 'The Communist: Frank Marshall Davis, The Untold Story of Barack Obama's M
entor' - Breitbart http://t.co/8R7eG3eEka via @BreitbartNews neg 0.8
@tonyzump I Said that before they elected Obama for his second term. Don't giv
e them too much credit, most went to Government schools. pos 1.0
RT @jorgeramosnews: @HillaryClinton ofrece "ciudadania" para indocumentados y ex
pandir la acci3n ejecutiva de Obama a padres de Dreamers. L... neg 0.8
Electing @HillaryClinton would not equal a 3rd #Obama term in office-

- it would be a 5th #Bush term in office pos 1.0
SILENCE FROM #OBAMA ON #ISLAMIC #TERROR ATTACK IN TEXAS http://t.co/5bFMLSbv88 #
WakeUpAmerica #PatriotsUnited #CCOT http://t.co/noyp98fotu neg 0.8

```

Figure 14 back end result.

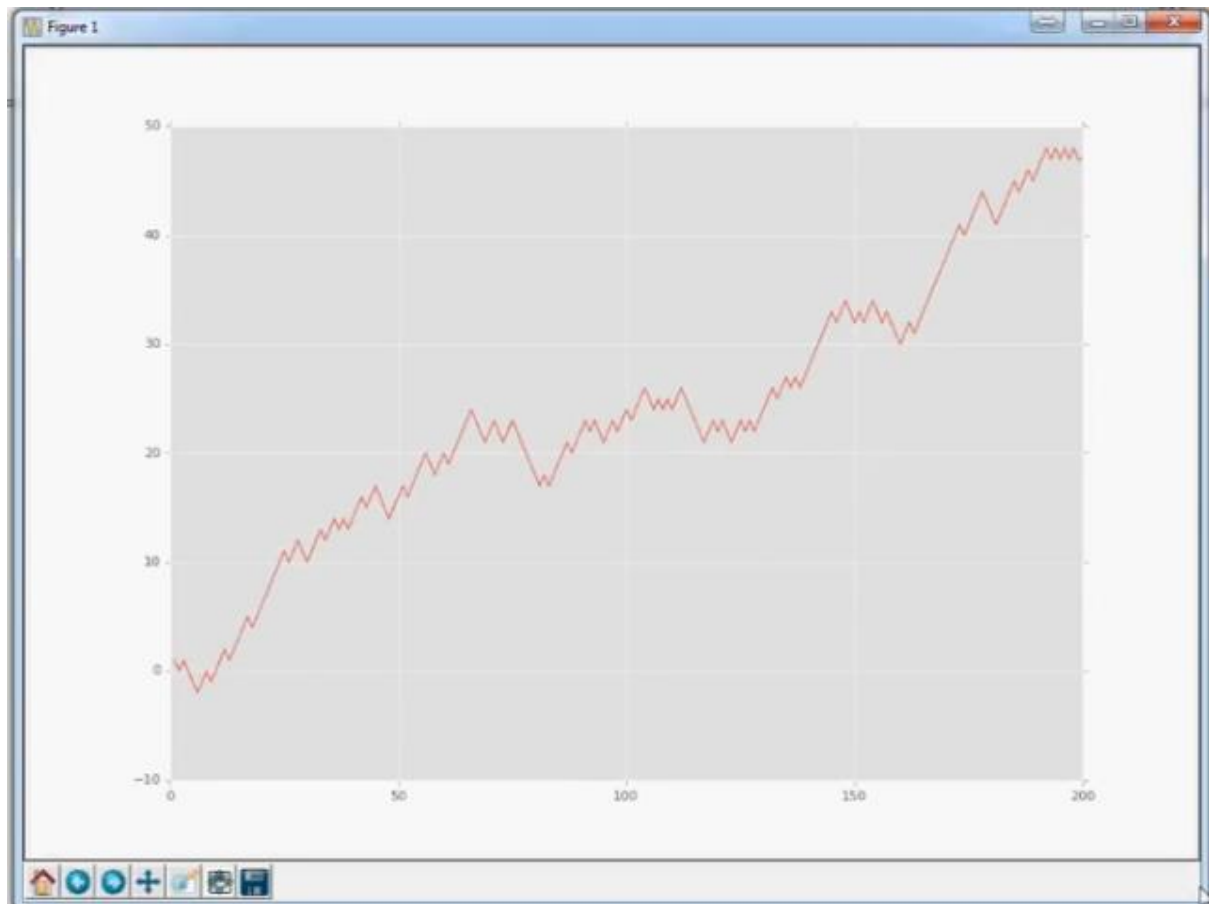
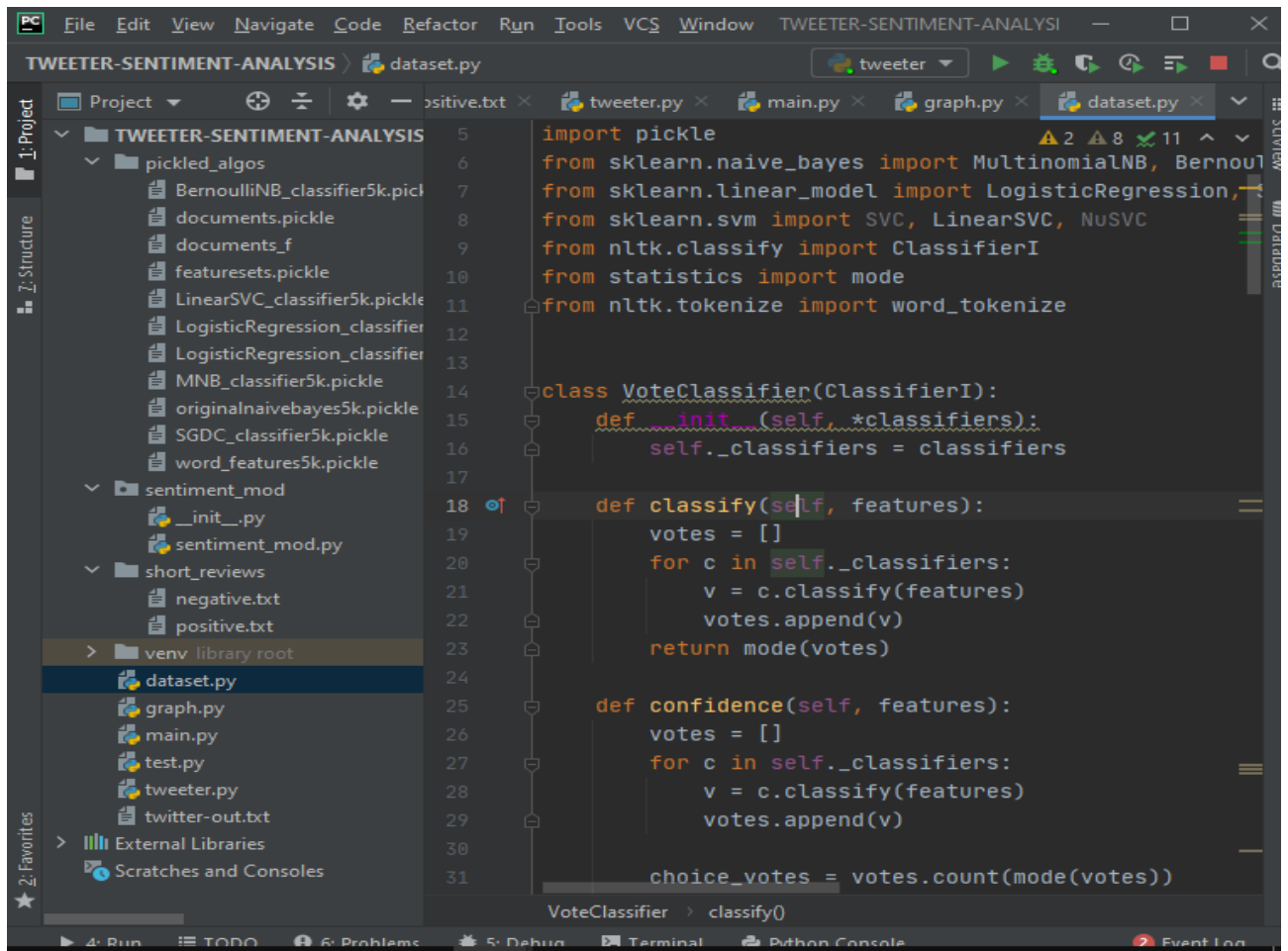


Figure 15 sentiment analysis graph

5.4.3. MODEL

The specifications for creating the model was

I7, 64bit processor, python3.9 about 2bg data bundle for downloading its dependencies and libraries and also for download the natural processing tool kit.



```
5 import pickle
6 from sklearn.naive_bayes import MultinomialNB, BernoulliNB
7 from sklearn.linear_model import LogisticRegression, LinearSVC
8 from sklearn.svm import SVC, LinearSVC, NuSVC
9 from nltk.classify import ClassifierI
10 from statistics import mode
11 from nltk.tokenize import word_tokenize
12
13
14 class VoteClassifier(ClassifierI):
15     def __init__(self, *classifiers):
16         self._classifiers = classifiers
17
18     def classify(self, features):
19         votes = []
20         for c in self._classifiers:
21             v = c.classify(features)
22             votes.append(v)
23         return mode(votes)
24
25     def confidence(self, features):
26         votes = []
27         for c in self._classifiers:
28             v = c.classify(features)
29             votes.append(v)
30
31         choice_votes = votes.count(mode(votes))
32         conf = choice_votes / len(votes)
33         return conf
```

Figure 16 code for training my model

Here is the result of the trained model with an **ACCURACY OF 73%**

```

10628
Original Naive Bayes Algo accuracy percent: 73.24840764331209
Most Informative Features
    engrossing = True          pos : neg   =   19.6 : 1.0
      mediocre = True          neg : pos   =   17.0 : 1.0
        generic = True          neg : pos   =   16.4 : 1.0
        routine = True          neg : pos   =   15.7 : 1.0
    inventive = True          pos : neg   =   15.6 : 1.0
        flat = True            neg : pos   =   14.2 : 1.0
        boring = True          neg : pos   =   13.9 : 1.0
    refreshing = True          pos : neg   =   13.0 : 1.0
        warm = True            pos : neg   =   12.6 : 1.0
    wonderful = True          pos : neg   =   11.8 : 1.0
        lame = True            neg : pos   =   11.7 : 1.0
    realistic = True          pos : neg   =   11.6 : 1.0
        stupid = True          neg : pos   =   11.0 : 1.0
    mesmerizing = True        pos : neg   =   11.0 : 1.0
        dull = True            neg : pos   =   10.7 : 1.0

MNB_classifier accuracy percent: 72.77070063694268
BernoulliNB_classifier accuracy percent: 75.15923566878982
LogisticRegression_classifier accuracy percent: 72.29299363057325
LinearSVC_classifier accuracy percent: 70.54140127388536
SGDClassifier accuracy percent: 72.29299363057325

Process finished with exit code 0

```

Figure 17: trained model analysis

```

dataset import sentiment as s
t(s("This movie was awesome! The acting was great, plot was wonderful, and there were pythons...so yea!"))
t(s("This movie was utter junk. There were absolutely @ pythons...I don't see what the point was at all. Hom

```

Figure 18 : testing the model

```

('pos', 1.0)
('neg', 1.0)

Process finished with exit code 0

```

Figure 19: result of the model output

Here is a picture of the developer twitter account I created so as to have access the twitter API

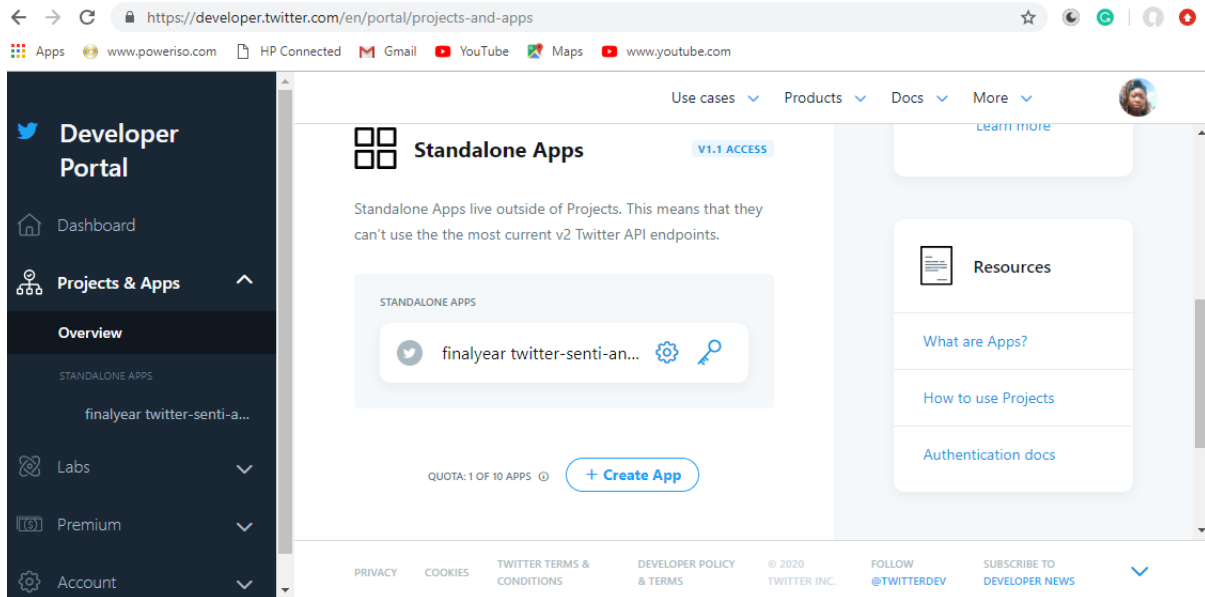


Figure 20: twitter developer account dashboard

5.5. Evaluation of the solution

During the research of this project I got to find many other existing projects with same functionalities as explained in chapter 2 and so here I will be stating the difference between my system and an already existing system.

5.6. Partial conclusion

I must say this chapter is the most interesting part of the research as it explains how the whole system is brought to a reality. It covered the comprehensive implementation of the proper work carried out is expounded.

This chapter covered the comprehensive implementation of the project from when tweets were streamed, then processed and finally the sentiment of each tweet is gotten. The most important part of this research is the output or final result which has also been provided in this chapter. This output helps the user to make good predictions and inferences which contributes greatly in decision making per the subject of interest.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Summary of findings

In the days before big data opened doors to seemingly exponential analytical insight, sales and marketing team often played a guessing game. Where campaigns working? Who were they reaching? What did people think about them? Audience targeting was one way to cut through the noise, but even that was a bit of a crapshoot. Today, though, big data (using it, not just having it) has created a functional shift from estimated actions to data driven, predictive choices. The growth of sentiment analysis as a marketing tool that is, technology that determines the emotional tone of statements made online about brands--goes one step further. The possibilities here are huge for your brand as big data and sentiment analysis team up to form a marketer's dream team.

Simply reading a post will let you identify whether the author had a positive stance or a negative stance on the topic, but that's if you're well versed in the language. However, a computer has no concept of naturally spoken language so we need to break down this problem into mathematics (the language of a computer). It cannot simply deduce whether something contains joy, frustration, anger, or otherwise. Without any context of what those words mean. Sentiment Analysis solves this problem by using Natural Language Processing. Basically, it recognizes the necessary keywords and phrases within a document, which eventually help the algorithms to classify the emotional state of the document or tweets like the case of the project.

Applying Sentiment analysis to mine the huge amount of unstructured data has become an important research problem. Now business organizations and academics are putting forward their efforts to find the best system for sentiment analysis. Although, some of the algorithms have been used in sentiment analysis gives good results, but still no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has Limitations. More future work is needed on further improving the performance of the sentiment classification. There is a huge need in the industry for such applications because every company wants to know how consumers feel about their products and services and those of their competitors. Different type of techniques should be combined in order to overcome their competitors. Different type of techniques should be combined in order to overcome their individual drawbacks and benefit from each other's merits and enhance the sentiment classification performance.

5.2 Contribution to engineering and technology

Big data and sentiment analysis is a very important aspect in the world of engineering and technology. Sentiment analysis has many applications and benefits to a business and to a business organization. With focus on this project, the sentiment analysis carried out an organization to find out how people feel about their product brand or services. This can help users of the system to make good pre-emptive decisions in order to step up in their business and also to be proactive.

The solution of this project is not just circumscribed to the business sphere. This application can also help an individual to find out how people feel about them especially in politics and to make good predictions and decisions to step their game.

Since the major factor of engineering is to solve problems, this application contributes much in engineering by solving some pertinent problems of industries and even individual problems. It helps people to enhance their decision appertaining to their subject of interest.

5.3 Recommendations

For this application to be deployed on another system and work effective, the following recommendations must be taken into account

All the necessary requirements must be installed. These requirements are the tools which are outline in chapter three, section 3.2.

The user must have a twitter account and must create a developer account so that he will be given the necessary credentials that will give him access to stream tweets from twitter API. It should be noted that, according to twitter policy, rules and agreement under taken by twitter developers, this credential (that is, consumer key, consumer secret, access token, and access token secret) is confidential. It therefore means that two people cannot share the same credential or it cannot be explored to the public.

5.4 DIFFICULTIES ENCOUNTERED

The challenges encountered during this project's creation are enormous and if I want to mention all of them this chapter will be more than 50pages so I shall only name a few.

1. Sentiment Analysis being a very broad field in computer science and computer engineering, I faced difficulty to fully comprehend the entire project and how sentiment analysis is carried out in particular. In addition to this, getting dataset from twitter was difficult since it entails a lot of procedure and applications.
2. Throughout this projects life span inconsistency of light was a great set back and I only manage to implement it right up this part.
Thank God for Go-Groups who ones in a while made provisions for a generator and it helped a lot for the progress of this write-up
3. There has also been a greet decrease in the internet connectivity quality of the whole country and so the research took me longer time than expected. But I was able to overcome this tragedy by doing more work instead because poor connection is at its peak during the day.
4. The pandemic covid19 also played a big roll to the sluggishness of this project as there were days I could not go out of the house to look for a quiet place to work since our house is usually noisy. I finally came up with an idea of going to close myself and work in the kitchen and this helped me immensely to succeed with this task.

With all this challenges sited god was still able and capable to see me through all to His glory.

5.5 Further works

I and everyone who helped me to accomplish this project did a great job not leaving out the fact that what was accomplished was the best of what I could accomplish with respect to my resource at hand.

However as stated above I encountered some challenges that slowed my work and I didn't finish all I wanted to do, they include:

1. Make a beautiful home page to attract more users.
2. I planned on authenticating the application so not every user has same privileges.
3. I also plan on deploying not only the frontend and the backend but also the model so people can just access it and use later in the future. To test the sentiments on text.
4. It will also be good if the application can have a frontend field where the user inputs the particular word he or she wants twitter to output and display the sentiment and graph.

This and many more are the things I will do as future work whenever I have time in order to improve my portfolio and also my skills

References

Big data: <https://www.guru99.com/what-is-big-data.html> 07/17/2020

Stages of big data <https://www.informit.com/articles/article.aspx?p=2473128&seqNum=11>
on 07/17/2020

Everything there is to know about twitter sentiment analysis

<https://monkeylearn.com/blog/sentiment-analysis-of-twitter/> on 07/17/2020

Disadvantages of big data

[https://www.google.cm/search?q=disadvantages+of+big+data&oq=disadvantage
s+of+big+data&aqs=chrome..69i57.3118777j0j4&sourceid=chrome&ie=UTF-8](https://www.google.cm/search?q=disadvantages+of+big+data&oq=disadvantage+of+big+data&aqs=chrome..69i57.3118777j0j4&sourceid=chrome&ie=UTF-8)
on 07/17/2020

Characteristics of big data <https://www.edureka.co/blog/big-data-characteristics/> on
07/17/2020

Everything there is to know about sentiment analysis [https://monkeylearn.com/sentiment-
analysis/](https://monkeylearn.com/sentiment-analysis/) 07/17/2020

Example of a sentiment analysis report

[https://pdfs.semanticscholar.org/0c24/bfdeac9e25cbd8c08a60a3fd28b8dca09b76.
pdf](https://pdfs.semanticscholar.org/0c24/bfdeac9e25cbd8c08a60a3fd28b8dca09b76.pdf) visited on 07/20/2020

Installing pip and a virtual environment [https://packaging.python.org/guides/installing-using-
pip-and-virtual-environments/](https://packaging.python.org/guides/installing-using-pip-and-virtual-environments/) on 7/21/2020

Sentiment analysis flow chat

[https://www.google.com/search?q=flowchart+diagram+for+sentiment+analysis&
oq=flow+chat+for+sentiment+&aqs=chrome.2.69i57j0l3.14483j0j4&sourceid=c
hrome&ie=UTF-8](https://www.google.com/search?q=flowchart+diagram+for+sentiment+analysis&oq=flow+chat+for+sentiment+&aqs=chrome.2.69i57j0l3.14483j0j4&sourceid=chrome&ie=UTF-8) visited on Sunday 2nd of August 2020

Streaming live twitter Data <http://adilmoujahid.com/posts/2014/07/twitter-analytics/> visited
on 8/ 4th /2020