A Project Synopsis

On

SENTIMENT ANALYSIS BASED ON COMMENTS FROM ONLINE SOCIAL NETWORK

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Degree of

Bachelor of Engineering

In

Information Technology

Submitted By

VEDANT PATIL

JAYESH THAKUR

KAPILDEV YADAV

Supervisor

Prof. DEEPTI LAWAND

Department of Information Technology

PILLAI COLLEGE OF ENGINEERING

New Panvel – 410 206

UNIVERSITY OF MUMBAI

Academic Year 2018 – 19

DEPARTMENT OF INFORMATION TECHNOLOGY
Pillai College of Engineering
New Panvel – 410 206

CERTIFICATE

This is to certify that the requirements for the BE Synopsis entitled '**Sentiment Analysis Based on Comments from Online Social Network**' have been successfully completed by the following students:

Name	**Roll No.**
VEDANT PATIL	B42
JAYESH THAKUR	B64
KAPILDEV YADAV	B36

in partial fulfillment of Bachelor of Engineering of Mumbai University in the Department of Information Technology, Pillai College of Engineering, New Panvel – 410 206 during the Academic Year 2018 – 2019.

_____

Supervisor

(Prof. Deepti Lawand)

_____ _____

Head of Department **Principal**

Dr. Sharvari Govilkar **Dr. Sandeep M. Joshi**

_____ _____

Internal Examiner **External Examiner**

DEPARTMENT OF INFORMATION TECHNOLOGY
Pillai College of Engineering
New Panvel – 410 206

SYNOPSIS APPROVAL FOR B.E

This project synopsis entitled "**Sentiment Analysis Based on Comments from Online Social Network**" by **Vedant Patil, Jayesh Thakur and Kapildev Yadav** are approved for the degree of B.E. in **Information Technology**.

Examiners:

1. _____

2. _____

Supervisors:

1. _____

2. _____

Chairman:

1. _____

Date:
Place

Declaration

We declare that this written submission for B.E. Declaration entitled "**Sentiment analysis based on comments from online social network**" represent our ideas in our own words and where others' ideas or words have been included. We have adequately cited and referenced the original sources. We also declared that we have adhere to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause for disciplinary action by institute and also evoke penal action from the sources which have thus not been properly cited or from whom paper permission have not been taken when needed.

Project Group Members:

Vedant D. Patil	&	Sign: _____
Jayesh S. Thakur	&	Sign: _____
Kapildev B. Yadav	&	Sign: _____

Date:

Place:

Table of Contents

Abstract

Internet is the platform where most of us share our happiness or other feelings. Recent years are devoted in studying and mining the data which is on social platform. This task includes understanding explicit and implicit information conveyed by sentiments. It can be extracted from the comments on social media using dictionary-based sentiment analysis or Review-Seer. Comments of the person are important to analyze the sentiments of the person at the time of writing the comment. The task is to classify the comments into positive, negative and neutral sentiments further into different emotions, for which it uses the concept of Plutchik's wheel of emotions and further makes a dictionary. The system will take input from user to classify and predict the emotions and strength of that emotion (Negative Emotions). There are basic eight emotions and system will primarily focus on negative emotions. Plutchik's wheel of emotion gives joy and sadness, anger and fear, trust and disgust, surprise and anticipation. The use of Plutchik's wheel of emotions will provide the real emotional view of comments. The polarity of emotion will be given that will indicate the strength of feeling. It uses Naïve Bayes algorithm for prediction and generates output.

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Fundamentals

The task of mining sentiments and opinions from natural language is difficult one. It involves an intense understanding of most of the implicit and explicit information which is conveyed by structure of language. The availability of a dynamic corpus contains the user generated data, such as reviews for products or polling data. Big data is the large amount of easily available data on web, Social media, remote sensing data, etc. in form of structured data, semi-structured or unstructured data. We can use this large data for sentiment analysis. Sentiment analysis is the opinion mining used on the web for identifying the text. It is nothing but to get the real voice of people for specific product, services, movies, news, issues from online social networking site like Twitter.

1.2 Objectives

The objectives of the project are as follows:

1. The objective of the project is to build a software that can initiate the analysis of the text extracted from the social media.
2. This extracted text is further analyzed using the concept of plutchik's wheel to give the depth of emotions of sentiments from the text extracted from social media.
3. The objective involves working with the most efficient and reliable algorithms and processes to give best output as possible.

1.3 Scope

The Sentiment Analysis basically contains analysis of the positive, negative and neutral comments further detecting the polarity of the emotions & uses the concept of Plutchik's Wheel of Emotions. The plutchik's wheel primarily contains the Eight basic emotions like Joy, Fear, Anticipation, Disgust, Sadness, Anger, Surprise, Acceptance. System will primarily focus on positive and negative emotions for sub-classification.

1.4 Outline

Including this the synopsis in all contains 5 chapters. The next chapter include the literature survey wherein all the previous work on the related topics have been studied and presented. The papers are deeply studied and each paper studied is mentioned along with their methodology, their advantage and the conclusion.

The third chapter, background knowledge is explained, along with the techniques and algorithms used in the development. It presents a high-level view of the system produced. It also covers the design patterns applied, the requirements gathering process, both functional and non-functional, and what methodologies have been used. Then, the Implementation Chapter demonstrates a low-level view of the system. The methods of execution or the flow of the system has been depicted with the help of Activity and Use Case diagram. The basic hardware and software requirements have also been stated.

The fourth chapter guides through the social and technical application of the studied model. Lastly, the Summary of the entire methods, techniques and algorithms has been explained in the reflective chapter. In addition, the references which are used for developing the model has been stated along with the words of Acknowledgement.

Chapter 2

Literature Survey

2.1 Introduction

Sentiment Analysis is basically used to express sentiment of the individual person. Current state of the art in sentiment analysis classified sentiments into two categories positive, negative. Some works classified sentiments as positive, negative and also in a third category objective (or neutral).

2.2 Literature Review

[1] Xianchao Wu, Hang Tony, Momo Klyen, "Fined grained sentiment analysis with 32 dimensions", University of Tokyo, 2017 IEEE. [1]

This model created understands capricious and complicated human emotions, they have proposed a sentiment analysis system wherein the input text emotion is classified into 32 categories. The model cover more detailed emotion and on the other hand, they have further measured each emotion with strength such as describing angry by range, anger and annoyance. They have used deep neural network classifiers, build the datasets using various methods and evaluated those models.

[2] Erik Tromp, Mykola Pechenizkiy, "Rule-Based emotion detection on social media: Putting tweets on Plutchik's wheel", Netherlands, 15 Dec 2014. [2]

This paper studies and analyze sentiments beyond polarity and uses Plutchik's wheel of emotions. It uses extension of Rule based emission model. This model thinks beyond the normal metrics of sentiment analysis using polarity and uses Rule-Based Emission Model (RBEM) algorithm (Tromp and Pechenizkiy, 2013) that can be used for polarity detection assigning new messages a label that is one of positive, neutral, negative. Important in algorithm is that positivity and negativity are opposite and allows negation to simply invert the emission. RBEM uses pattern matching and uses wildcards for it. The model used is compact as well as complete which works well with RBEM-Emo which is stated as extension of Rule Based detection algorithm.

[3] Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete, "Combining strength, emotions and polarities for boosting Twitter sentiment Analysis", Chile, 11 Aug 2013. [3]

This paper proposes an approach for boosting twitter sentiment classification using different sentiment dimensions as meta-level features. This research shows the combination of sentiments improves the twitter sentiment classification tasks. The scopes of tweets are categorized upon some categories as polarity, emotion, strength. It does different testing with different types of algorithm. It uses classification approach like OpinionFinder Lexicon, AFFIN Lexicon, SentiWordNet Lexicon, SentiStrength Lexicon, Senti140 method, NRC Lexicon. So, when it classifies tweets into polarity classes, we are essentially projecting these multiple dimensions to one single categorical dimension. But also, sentiment classification of tweets can lead to loss of valuable sentiment information.

[4] Dhanshri Chafale, Amit Pimpalkar, "Sentiment Analysis on Product reviews using Plutchik's Wheel of Emotion with Fuzzy Logic", Nagpur University, Dec 2014. [4]

The model consists of stemming and stop word techniques. This filtering removes almost all unwanted noise from comment. The filtered comment is then split to get the separate words for comparing. Then each single word is compared with the sentiment words dictionary. If the word is matched with the positive or negative dictionary then it is placed in the corresponding box, that is positive word in positive words text and in the same way negative words are placed. The comparison is done between number of positive word and number of negative words in a given comment. The condition is checked whether the positive words are more or negative and accordingly the comment is decided to be positive or negative. If both the positive and negative words are same or if there are no positive or negative, the comment is treated as neutral comment.

[5] Hamid Bagheri, Md Johirul Islam, "Sentiment Analysis of Twitter data", IOWA State University, 2015. [5]

Their implementation has followed various steps like:

1. Start with downloading and caching the sentiment dictionary.
2. Download twitter testing data sets, input it in to the program.
3.Clean the tweets by removing the stop words.
4. Tokenize each word in the dataset and feed in to the program.
5. For each word, compare it with positive sentiments and negative sentiments word in the

dictionary. Then increment positive count or negative count.

6. Finally, based on the positive count and negative count, we can get result percentage about sentiment to decide the polarity.

2.3 Literature Summary

Following are the paper referred for the project and their short description.

Paper	Advantages	Disadvantage	Inference
Xianchao Wu, Hang Tony, Momo Klyen, "Fined grained sentiment analysis with 32 dimensions", University of Tokyo, 2017 IEEE. [1]	The emotions have 32 classes which are divided from 8 basic types of emotions. This system has wider and accurate scope.	The emotions classification becomes unnecessarily detailed and dictionary becomes messy.	This paper has close resemblance with proposed system. It has vast scope as it classifies the opinions to 32 different sentiments.
Erik Tromp, Mykola Pechenizkiy, "Rule-Based emotion detection on social media: Putting tweets on Plutchik's wheel", Netherlands, 15 Dec 2014. [2]	The model proposed is relatively compact and yet complete. This classification model is highly coupled with Plutchik's wheel.	It takes amount of time for rule creation and hence increases time.	Results of this system shows RBEM-Emo is competitive to current systems supporting three different language
Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete, "Combining strength, emotions and polarities for boosting Twitter sentiment Analysis", Chile, 11 Aug 2013. [3]	This system uses existing system and further enhances the classifier accuracy with more precision	Only useful for particular type of sentence. Also dependent on platform.	The technique which is used does boost the performance of Twitter sentiment analysis system.

Paper	Advantages	Disadvantages	Inference
Dhanshri Chafale, Amit Pimpalkar, "Sentiment Analysis on Product reviews using Plutchik's Wheel of Emotion with Fuzzy Logic", Nagpur University, Dec 2014. [4]	This system uses Fuzzy logic for prediction. Additionally, which helps to predict which product is good or bad and thus has wider scope.	Accuracy is good but requires large amount of data for training.	This system basically focuses more on working on reviews, feedbacks than comments along with analysis of the sentiments and emotions. It has also has considerably good accuracy.

Table 2.3. Summary of Literature Survey

Chapter 3

Sentiment Classifier System

3.1 Overview

The system is trained to analyze the texts in comments and then classify according to classes with specific classifier. Different classifiers give different levels of accuracy.

3.1.1 Existing System Architecture:

The existing system performed aspect level sentiment analysis on twitter data based on three classes:

1) Positive

2)Negative

3)Neutral

Twitter is a Social Networking service that is restricted to 140 Characters in length using various emotions, Hashtag etc.

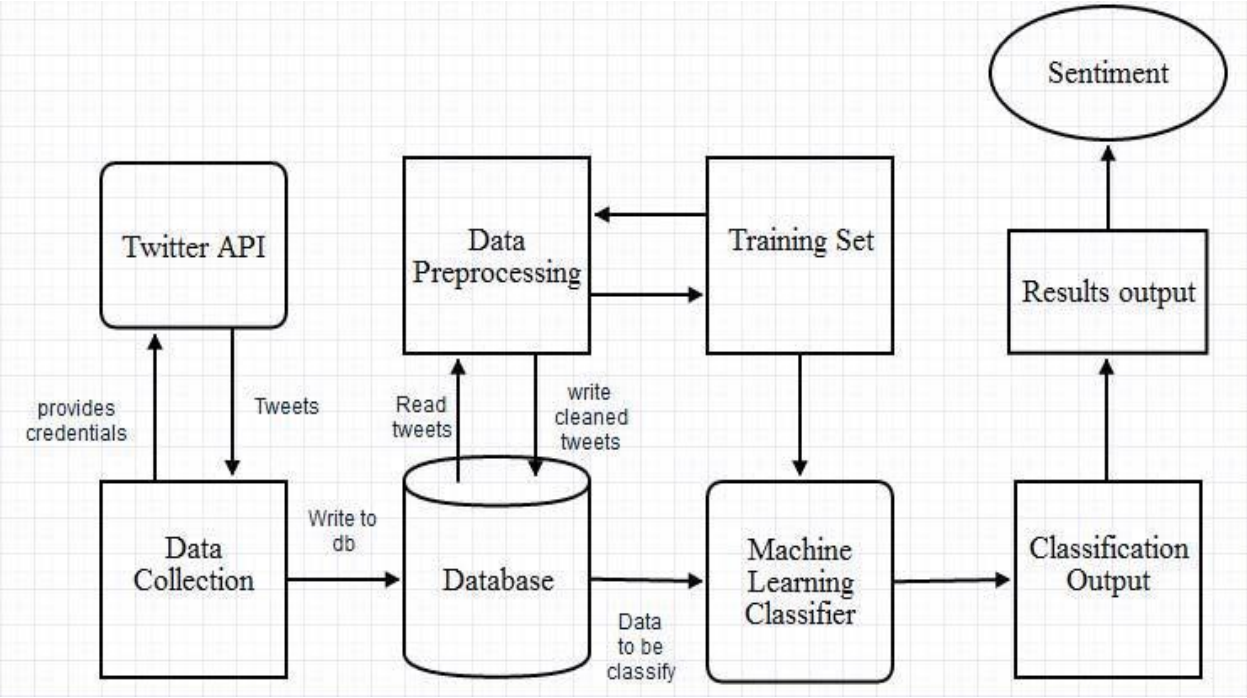

Fig. 3.1.1. Existing system architecture used for Content based Systems [8]

Existing System Has Following Steps [8]:

1)Data collection using twitter API: Publicly large sets of Twitter data are not available. Hence, they first extracted twitter data through twitter API

2)Data Preprocessing: It involved cleaning of data by spell correction punctuation etc. Reducing noise from the data.

3)Applying Classification Algorithm: The Classification Algorithm is applied on tweets to categories them with highest accuracy.

4) Classified tweets and result: The tweets are further classified three defined categories. Result of which is displayed in form of pie chart.

Data Preprocessing

Following are steps generally followed in data preprocessing:

Fig 3.1.2. Steps in Data Preprocessing

1)Converting to lower/upper case: The whole text is converted into single case which make it easier.

2)Removing URLS: hyperlinks do not play an important role in sentiment classification so they are removed.

3)Removing Unicode character: Unicode characters are used to represents emoticons and many other complex symbols are removed in order to reduce the complexity of data preprocessing step.

3.1.2 Proposed System Architecture:

Today people are uploading and expressing themselves on social media platform. These comments will affect the strategy designed by companies. It will also have impact on ups and downs of market. The proposed system will primarily classify emotion into positive, negative, neutral and further into 8 basic emotions. Firstly, the data required for analysis will be divided

into testing and training datasets, this dataset are downloaded from official twitter APIs. The Naive Bayes classifier will be trained according to this dataset.

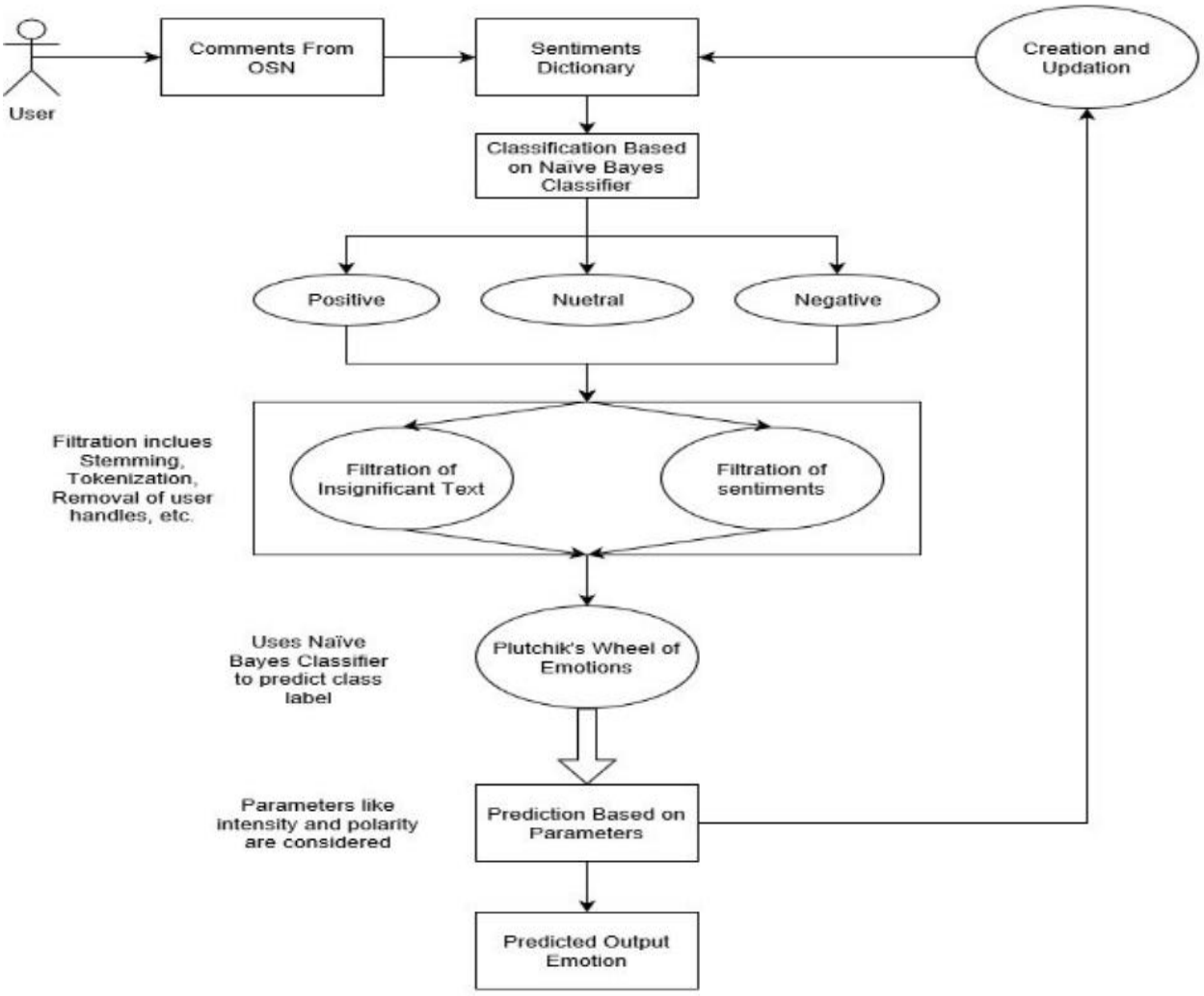

Fig 3.1.3. Architecture of system

After data collection filtration of data will be done using pre-defined tokenizers. This tokenizer does feature extraction, part of speech, stemming and other filtration of data. Filtered data is further used for classification into emotions.

Naive Bayes classifier will train using previous training and test datasets. This naive Bayes classifies the comments into 8 classes named as anticipation, joy, anger, sadness, surprise, disgust, fear and trust. This comment will be saved along with its tag of class into sentiment dictionary for further learning of system.

The next prediction of system is done based on previous results by system.

3.2 Requirements for Implementation

3.2.1 Implementation Details

Data Preprocessing:

The techniques which are used for data preprocessing is done using predefined library function of python called textblob. Textblob preprocesses data by stemming the data and further steps like lemmatization, tokenization as well as part-of-speech tagging.

Naive Bayes Classifier:

After preprocessing, the sentence which needs to be classified using same textblob library function. The predefined Naive Bayes classifier function of textblob is trained using training dataset. The dataset is chosen with .csv file. The training dataset is used with f_read function.

We can train classifier with

cl = NaiveBayesClassifier(train)

We can also find accuracy of test set using

cl.accuracy(test)

After training output which has to be given using naive Bayes classification technique. We know the formula which can be used for classification as,

$P(A|B) = P(B|A) * P(A) / P(B)$

Where, $P(A|B)$ is probability the A belongs to class B

$P(B|A)$ is evidence

$P(A)$ is probability of class A is seen and similar with B.

The $P(A|B)$ is calculated for each word and then the class tag is selected with maximum probability. Maximum probability is selected and saved into dictionary for further increasing accuracy of classifier.

3.2.2 Activity Diagram and Use Case Diagram

Following shows the activity flow of the system. It starts from used entering comments and then processing that comment through naïve Bayes algorithm which will give results to user that comment is positive, negative, neutral along with emotion.

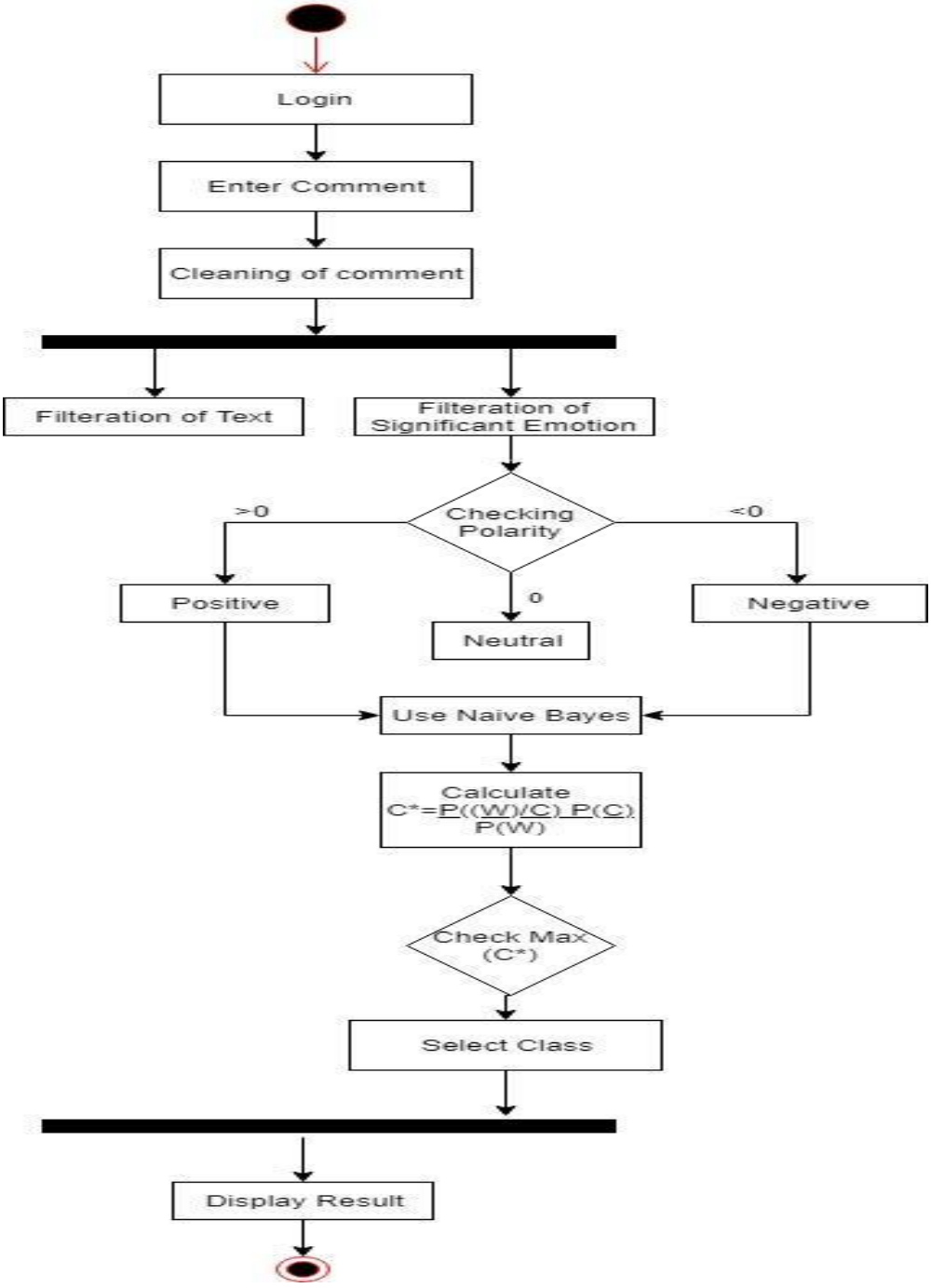

Fig 3.2.1. Activity Diagram

Following is use case diagram which show the interaction between the user and system. User entering login credentials is first interaction. Afterwards user enters comment and gets result.

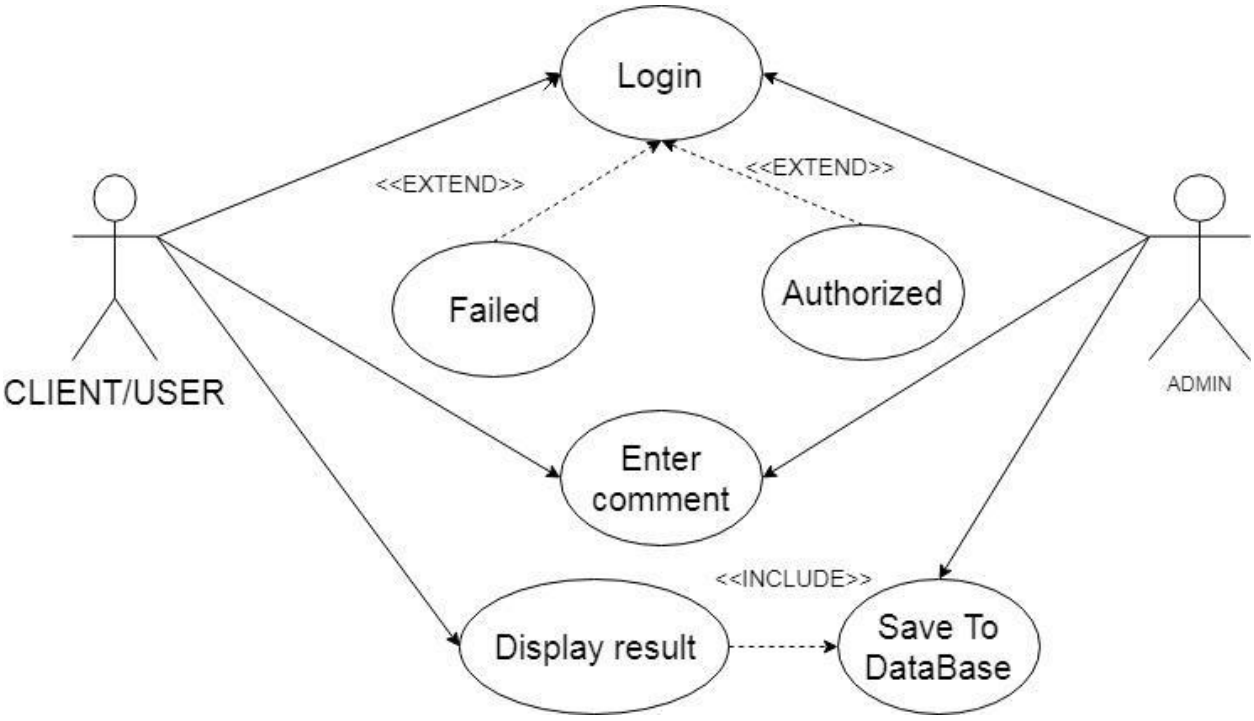

Fig 3.2.2 Use Case Diagram

3.2.3 Sample Dataset Used

The dataset contains sentences labelled with positive or negative sentiment. It contains sentences labelled with positive or negative sentiment. Score is either 1 (for positive) or 0 (for negative). The sentences come from three different websites/fields:

1. imdb.com
2. Amazon.com
3. twitter.com

For each website, there exist 500 positive and 500 negative sentences. Those were selected randomly for larger datasets of reviews.

We attempted to select sentences that have a clearly positive or negative emotion, the goal was for least neutral sentences to be selected.

Dataset	No. of Comments	Emotions
imdb.com	1000	Positive, Negative, Neutral
amazon	1000	Positive, Negative, Neutral
twitter	1000	Positive, Negative, Neutral

Sentence	Sentiment (0 for negative, 1 for positive)
Omg it’s already 7:30	1
Baddest day ever	0
Bathroom is clean...now more enjoyable tasks...	1
but I'm proud.	0
Going to See Harry Sunday Happiness	1
I missed new moon trailer	0
I think my bf is cheating on me	0
OK that’s it you win	0
Thanks to all guys	1
Miss you already	0

Table 3.1 Dataset Used for Experiment [10]

3.2.4 Hardware and Software Specifications

The experiment setup will be carried out on a computer system which has the different hardware and software specifications given in Table 3.2 and Table 3.3 respectively.

Processor	2 GHz Intel Pentium
HDD	180 GB
RAM	2 GB

Table 3.2 Hardware Details

Operating System	Windows 10
Programming Language	Python (Flask Framework), HTML
Database	MySQL

Table 3.3 Software details

Chapter 4

Applications

There are various applications of this domain system. The application is listed here.

4.1 Social

Criminal activities: The most popular platform such as social media which can be major tool in criminal activities, the suspect make fake profile, fake web pages, in order to phis the particular user by pretending as legit profile, webpage. Sentiment analysis identifies such activities through. Deciding the emotion for particular site, profile etc. which helps in solidifying law and order.

Detecting Personality Traits: Personality of a person shows according to its traits and the way they express their feelings. So, this automatic sentiment analysis can also be used to detect personality traits.

Politics: This area has tremendous interest as analysis. This will certainly have big impact on wide set of people. It will be also useful for predicting the favor for each candidate.

Detection of negative public assault: It can also be used for prediction and detection of potentially dangerous situation.

4.2 Technical

Tone detector: It can be used be used in different documentation setting. Tone detection technique used in Outlook, to identify the type of document. When user writes mail this algorithm analysis the particular tone of that mail and adapts the writing of mail according to the detected tone.

Data analysis: This analytical system will surely hop up the process of Artificial intelligence. Machine can easily learn different aspects of human interaction and expression using comments from social platform.

Chapter 5

Summary

The following chapter aims to clarify the techniques used throughout the project. A broad definition will be given for the core concepts involved in the development of the artefacts: Text Mining, Natural Language Processing and Machine Learning. Text mining refers to the analysis of data contained in natural language text, (e.g. messages retrieved from twitter). The main goal of text mining is to process data into a structured format ready for analysis, via application of natural language processing and other analytical methods.

Natural Language Processing (NLP). For the purpose of explaining further concepts, Twitter will be used as a running example. The data retrieved from twitter presents a certain amount of structuring, in the sense that the maximum length of a tweet is 140 characters long. The advantage of the length limit is reflected in the complexity of the analysis for an individual piece of text. However, this project aims to analyze data in a continuous manner, where a large amount of data will be analyzed.

Tokenization is another technique used in developing the model. The first task, that must be completed before any processing can occur, is to divide the textual data into smaller components. This is a common step in a Natural Language Processing (NLP) application, known as tokenization.

Stemming and Lemmatization is very important technique. The goal of both stemming and lemmatization is to reduce inflectional forms and derivations of a word to a common base form. For example, the following words: “connection”, “connections”, “connective”, “connected”, “connecting” will have the same base, which is “connect”. Stemming, is a crude heuristic process that chops off the ends of words, so that only the base form is kept. By contrast, lemmatization uses the morphological analysis of the words, returning their dictionary form (base). The rest of the section uses Machine Learning Algorithms to produce which are used to classify the text into positive, negative and further into 8 basic emotions. The Python is programming language which will be used for developing the model.

References

[1] Xianchao Wu, Hang Tony, Momo Klyen, “Fined grained sentiment analysis with 32 dimensions”, University of Tokyo, 2017 IEEE.

[2] Erik Tromp, Mykola Pechenizkiy, “Rule-Based emotion detection on social media: Putting Tweets on Plutchik’s wheel”, Netherlands, 15 Dec 2014.

[3] Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete, “Combining strength, emotions and polarities for boosting Twitter sentiment Analysis”, Chile, 11 Aug 2013.

[4] Dhanshri Chafale, Amit Pimpalkar, “Sentiment Analysis on Product reviews using Plutchik’s Wheel of Emotion with Fuzzy Logic”, Nagpur University, Dec 2014.

[5] Hamid Bagheri, Md Johirul Islam, “Sentiment Analysis of Twitter data”, IOWA State University, 2015.

[6] Aliza Sarlan, Chayanit Nadan, Shuib Basri, “Twitter Sentiment Analysis”, Universiti Teknologi Perak, Malaysia, 18 Nov 2014.

[7] Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala, “Sentiment Analysis of English Tweets using RapidMiner”, GGITS Jabalpur, India, 2015 IEEE.

[8] M Trupthi, Suresh Pabjour, G Narsimha, “Sentiment Analysis using Streaming API”, CBIT Telangana, 2017 IEEE.

[9] Zhaoxia Wang, Chee seng Chong, Landy Lan, Yinping Yang, Seng Beng Ho, Joo Chaun Tong, “Fined grained sentiment of social media analysis using emotion sensing”, IHPC Singapore, 7 Dec 2016.

[10] UCI Machine Learning Repository: Sentiment Labelled Sentences Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Sentiment Labelled Sentences. [Accessed: 21-Oct-2018].

Acknowledgement

We would like to express our specials thanks of gratitude to our **Principal Dr. Sandeep Joshi** who gave us the golden opportunity to do this wonderful project in the topic Sentiment analysis using Plutchik's wheel, which also helped us in doing lot of research and we came to know about so many new things.

We would also like to thank **Dr. Sharvari Govilkar** for giving us opportunity to get practical exposure into actual environment and solidifying our theoretical thinking and changing it to applicative way.

We also convey our thanks to Project Coordinator **Prof. Gayatri Hegde** and Project Supervisor **Prof. Deepti Lawand** for providing us constant support, encouragement and guidance and sharing their pearls of wisdom with all of us, which was great help.

We are also thankful to everyone who supported us, for that we have completed our report successfully, efficiently and moreover on time.

Lastly, but not least, we are thankful to all of those who gave us helpful comments which helped us a lot in preparing this assignment.

Vedant Patil
Jayesh Thakur
Kapildev Yadav