DWIT COLLEGE

DEERWALK INSTITUTE OF TECHNOLOGY

SENTIMENT ANALYSIS of TWITTER DATA USING LOGISTIC REGRESSION

A Mini PROJECT REPORT

Submitted to

Department of Computer Science and Information Technology

DWIT College

In partial fulfillment of the requirements for the Bachelor's Degree in Computer Science and Information Technology

Submitted by

Raju Shrestha

Batch 2019

November, 2018

DWIT College

DEERWALK INSTITUTE OF TECHNOLOGY

SUPERVISOR'S RECOMENDATION

I hereby recommend that this project prepared under my supervision by RAJU SHRESTHA entitled "**SENTIMENT ANALYSIS OF TWITTER DATA USING LOGISTIC REGRESSION**" in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Information Technology be processed for the evaluation.

…….

Ritu Raj Lamsal

Head of Electronics Department

Deerwalk Institute of Technology

DWIT College

DWIT College

DEERWALK INSTITUTE OF TECHNOLOGY

LETTER OF APPROVAL

This is to certify that this project prepared by RAJU SHRESTHA entitled "SENTIMENT ANALYSIS OF TWITTER DATA USING LOGISTIC REGRESSION" in partial fulfillment of the requirements for the sixth semester of degree of B.Sc. in Computer Science and Information Technology has been well studied. In our opinion it is satisfactory in the scope and quality as a project for the required degree.

…..……..	……
Ritu Raj Lamsal [Supervisor/Examiner]	Dr. Sunil Chaudhary[Examiner]
Head of Electronics Department	Head of Computer Science Department
DWIT College	DWIT College

ACKNOWLEDGEMENT

I would like to thank Mr. Rituraj Lamsal, Head of Electronics Department, DWIT College, for supervising this project and giving assistance when faced with problems. Without his guidance and persistent help, this project would not have been possible.

At last but not the least my sincere thanks goes to my parents and member of my family, who have always supported me and to all of my friends who directly or indirectly helped me to complete this project report.

Raju Shrestha
Roll No. 0522
Batch 2019

STUDENT'S DECLARATION

I hereby declare that I am the only author of this work and that no sources other than that listed here have been used in this work.

……

Raju Shrestha

Batch 2019

Date: November, 2018

ABSTRACT

Sentiment Analysis is a method for judging somebody's sentiment or feeling with respect to a specific thing written in a piece of text. It is used to recognize and arrange the sentiments communicated in writings. The web-based social networking sites like twitter draws in a huge number of clients that are online for imparting their insights in the form of tweets or comments. The tweets can be then classified into positive, negative, or neutral. In the proposed work, logistic regression classification is used as a classifier and unigram as a feature vector.

Keywords: Sentiment analysis, Opinion mining, Text classification, Unigrams, Polarity, Machine learning, Logistic regression, Natural Language Processing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

POS:	Part of Speech
API:	Application Programming Interface
CSV:	Comma Separated Values
CSS:	Cascading Style sheet
HTML:	Hyper Text Markup Language
URL:	Uniform Resource Locator

Note: The system has been referred to as “Sentiment Analysis of Twitter Data Using Logistic Regression” in the rest of the document

CHAPTER 1: INTRODUCTION

1.1. Background

In the present days, Micro blogging has become a very popular communication tool among Internet users. Many users share tweets or messages everyday on prevalent sites, for example, Twitter and Facebook. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of micro blogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to micro blogging services.
As more and more users post about products and services they use, or express their political and religious views, micro blogging web-sites become valuable sources of people's opinions and sentiments. Such data can be efficiently used in research, business or social science.

Sentiment is positive or negative reviews about product or on any topics. We people can identify tweets by reading whether it is positive or negative. But if there is huge data to be read then it would be tedious and time consuming. So, if all this process could be done with the help of automated program then it would be easier and above manual process could be eliminated.

Sentiment Analysis of Twitter Data Using Logistic Regression is a web-based application which takes tweets as an input and gives sentiment value as an output.

1.1.1. Defining Sentiment

A sentiment is defined as a view or opinion that is expressed. It is a feeling of someone that he/she expresses either in textual or verbal form. A sentiment can be defined as a personal positive or negative feeling.
For example: This is the best budget smartphone. This is positive sentiment. This phone have bad resolution is consider as negative sentiment.

1.1.2. Characteristic of Tweets

Twitter message have many unique attributes (Go, Bhayani, & Huang) which are as follows

*Tweets and Length***:** Tweets are the status posted by user which is of 140 length.

*Username***:** Username in twitter is started with @followed by text and number. Eg @barackobama.

1.2 Problem Statement

Every day millions of data is being collected on twitter which contains people opinion about many things like manufacturing products, political parties programs etc. And, the data is unstructured and not organized in a pre-defined manner. These text are usually difficult, time-consuming and expensive to analyze, understand, and sort through.

Many company wants to know how positive or negative peoples are about their product. People want to know other people how much positive or negatives tweets he/she tweet. Tweets Sentiment Analysis Using Logistic Regression Algorithm will provide the positive or negative sentiment on tweets that people have tweeted.

1.3 Objectives

The main objectives are:

To develop a web-based application for tweets sentiment analysis.

To implement Logistic Regression Algorithm for sentiment analysis.

1.4 Scope

Sentiment Analysis of Twitter Data Using Logistic Regression Algorithm can be used by people who want to know the sentiment of their tweets. It can be used by company to find out the opinion about their products. It can also be used by many organization to know opinion about the event they have organized or going to be organized.

1.5 Limitation

a) This application can only analyze the sentiment of English words.

b) This application will not give 100% accurate result.

1.7 Outline of Document

The report is organized as follows

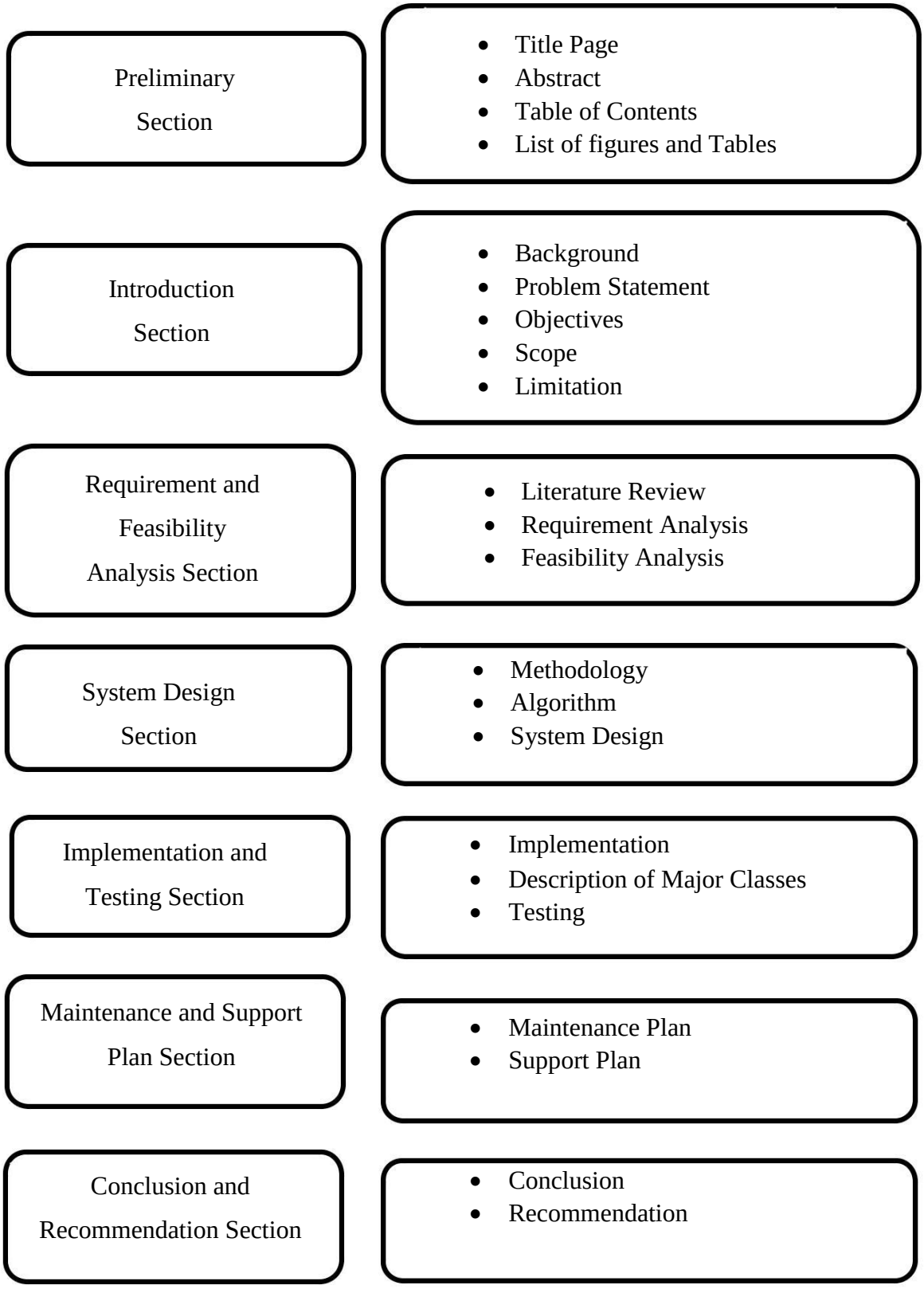

Figure 1- Outline of document

CHAPTER 2: REQUIREMENT AND FEASIBILITY ANALYSIS

2.1 Literature Review

2.1.1 Supervised machine learning for sentiment analysis

Sentiment analysis have become the growing area in the natural language processing. Supervised machine learning algorithm like Logistic Regression algorithm paly vital role in the sentiment analysis. There are many researched carried out for sentiment analysis.

(Pang, Lee, & Vaithyanathan, 2002) Studied various technique for sentiment analysis for the movies review. They compare the different classification algorithm like Naïve Bayes classification, Maximum Entropy classification and Support Vector Machine. They also consider the different factor affecting the sentiment like unigrams, bigrams, Part of speech (POS) etc. They achieved accuracy of above 80% for all three algorithm using unigrams + bigrams.

(Waykar, Wadhwani, Pooja, & Kollu, 2016) Have focused mainly on the Naïve Bayes classifier. They take the baseline for their research as (Pang, Lee, & Vaithyanathan, 2002). They display the result on pie chart for positive, negative and neutral for the specific keyword.

(S.T Indra, Liza Wikarsa & Rinaldo Turang) have focused on topic based classification based on the Logistic Regression. They also have used the confusion matrix as a classifier model. They achieved the accuracy of 92% for the tweets classification into selected topics.

(Abhilasha Tyagi1 & Naresh Sharma, 2018) Have proposed research based on Logistic Regression. They have used Logistic Regression as classifier and unigram as a features vectors. For increasing the accuracy K-fold cross validation and tweet subjectivity is used. To further speed up the classification process they also use the idea of effective word score heuristics that find out the polarity score of the words which are frequently used.

Supervised machine learning classifier required the trained data set to work. For this I have used publicly available labeled dataset.

2.2 Requirement Analysis

Table 1- Functional and non-functional requirements

Functional requirement	Non-functional requirement
Downloads tweets	Downloads tweets by specific keywords when query is submitted.
Show sentiment value of tweets	Display List of tweets with positive and negative value on it

2.3 Feasibility Analysis

2.3.1 Technical feasibility

Sentiment Analysis of Twitter Data Using Logistic Regression is a web based application that uses Flask (A Python Framework). It uses HTML, JavaScript, CSS for front-end, python, NLTK, Scikit-learn as the back end. It requires client, and internet connection to function properly.

It supports both on Windows and Linux platform for its operation. All the technology required by the Sentiment Analysis of Twitter Data Using the Logistic Regression are available freely, hence it was determined technically feasible.

2.3.2 Operational feasibility

Sentiment Analysis of Twitter Data Using Logistic Regression has a simple design and is easy to use. It uses two-tier architecture (i.e. Client and Server). It can be easily accessed from anywhere having the internet if we host it on cloud. Hence Sentiment Analysis of Twitter Data Using the Logistic Regression was determined operationally feasible.

CHAPTER 3: SYSTEM DESIGN

3.1 Methodology

This project use supervised machine learning classifier which is Logistic Regression. Logistic Regression requires labeled data for training the classifier.

3.1.1 Data collection

Data used by the Sentiment Analysis of Twitter Data Using Logistic Regression Algorithm was collected from the publically available data which is already being placed by other researchers. It is collected from the (kaggle.com). The dataset consists of 1,00,000 training and 15,034 testing data in csv format and is labeled 0 for negative and 1 for positive. So this project uses only the positive and negative datasets.

3.1.2 Data Preprocessing

The twitter data consist of different properties in which most of it is not useful for sentiment analysis. Data preprocessing includes various step.

1. *Usernames*: Twitter consists of username which consist of symbol @ at the beginning.eg @sparkingroshan. It is replaced by the word AT_USER in data sets which is started by @ in the datasets.
2. *Usages Link*: User includes the link in the tweets for the more detail information which is not useful for sentiment analysis. The link is replaced by the word 'URL'.
3. *Stop Words*: Stop word are those filler words which are not useful in for sentiment. These words includes most repeated word like a, an, the, for, etc. These words does not give any sentiment hence they are filtered out form the datasets.
4. *Removing Hash-tags:* Hash-tag in a tweet is used to emphasize a particular word or sentence, for example #thisisgood. Removal of these hash-tags is important because these hash-tags do not define any sentiment. Thus pre-processing is done and hash-tags before any word are removed.

5. *Repeated Letters*: Tweets contain the very causal language (Waykar, Wadhwani, Pooja, & Kollu, 2016) so the word such as hurrayyy is replaced with actual word hurray. The letter repeated more time reduced to the one.
6. *Stemming:* Change a word in the text into its base term or root term. Example, happiness to happy.

3.1.3 Feature Extraction

After preprocessing the tweets, tweets is converted into feature vector. Feature vector are the most important concept in implementing classifier (ravikiranj, n.d.).Feature vector is used for building the model and is used to train the model which is further used to classify the unseen data. Feature vector is the n-dimensional vector of numerical features that represent the some object. In tweets we can consider the presence or absence of words that appear in the tweets. The tweets in training data is split into words and each words into feature words. The feature words may consist of words unigram or bigrams. This project consider unigram as feature words. For eg. This is the ball is represented as this, is, the, ball as unigrams. The entire feature vector will be the combination of each of this feature words.

3.2 Algorithm

The algorithm used is Logistic Regression. Logistic Regression is predictive analysis model based on binary classification. It classify the tweets based on the probability given to tweets belong to that particular class. To predict the tweets into positive and negative. I have used label dataset with probability value 0 for negative and 1 for the positive tweets. Logistic regression, use a Logistic function, for instance, a sigmoid function to estimate probabilities between positive or negative label and data features.

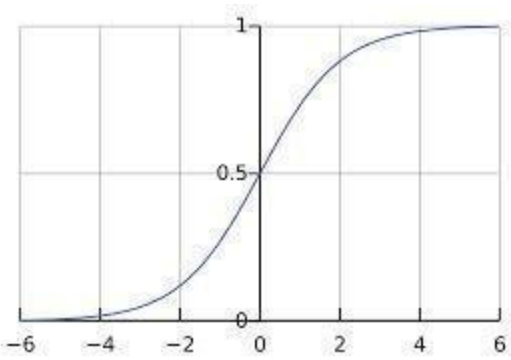

Figure 2 - Sigmoid Function

Logistic Regression is a discriminative model which means computing P(y|x) by discriminating among the different possible values of the class y based the given input x. The equation for this is as shown below:

$$P(c|x) = \sum_{i=1}^{N} w_i . f_i$$

To generate a value of P(y|x) of an output that is in between value 0 and 1, the following exp function is used:

$$P(c|x) = \frac{1}{z} \exp \sum_i w_i . f_i$$

To change the normalization factor Z and specify the number of features as N is as follows:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i . f_i)}{\sum_c \exp(\sum_{i=1}^{N} w_i . f_i)}$$

The final equation for computing the probability of y being of class c given x is:

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i . f_i(c,x))}{\sum_{c' \epsilon C} \exp(\sum_{i=1}^{N} w_i . f_{i(c',x)})}$$

3.3 System Design

3.3.1 System Architecture Diagram

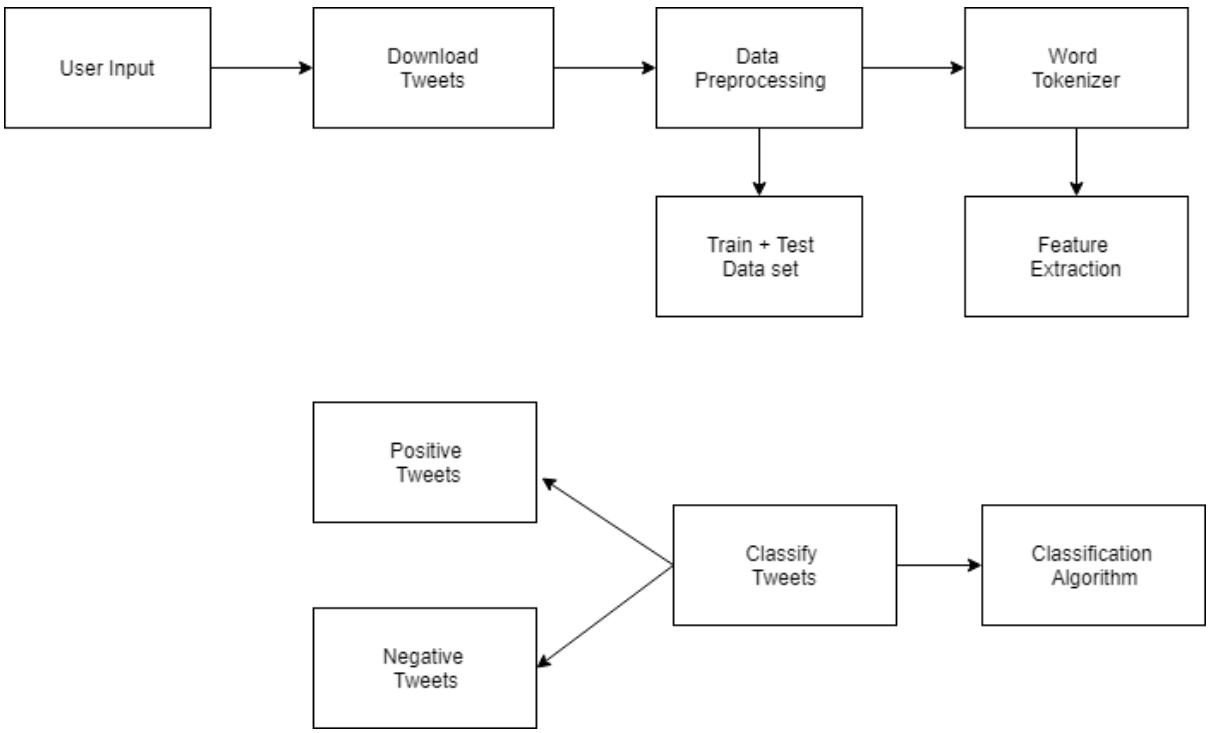

Figure 3 - System architecture diagram

3.3.2 Sequence Diagram

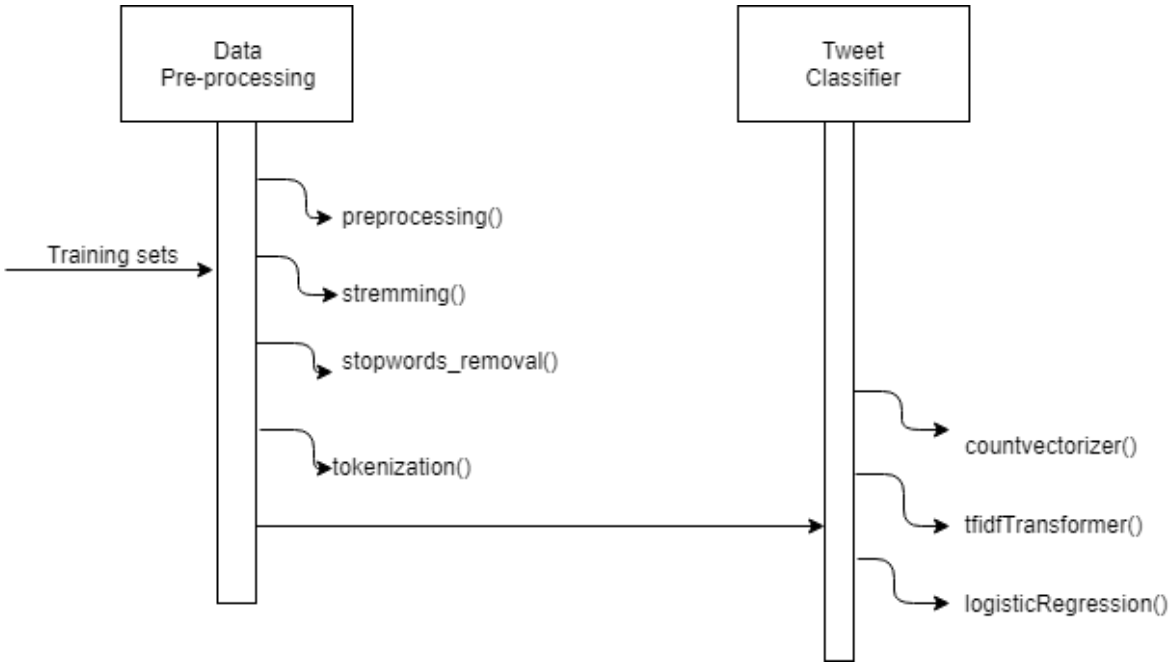

Figure 4 - Sequence diagram

CHAPTER 4: IMPLEMENATION AND TESTING

4.1 Implementation

User can access the application through a browser and see the interface. They can

4.1.1 Tools used

CASE tools:

a) Draw.io

Client side:

a) HTML used to display the content in the browser.
b) CSS is used to adjust the layout, look and design of the HTML content.
c) Bootstrap CSS framework is used for beautifying the HTML elements to improve the user experience.
d) Flask web framework is used for dynamic webpage generation and to display the predicted result in the browser as well as to handle page requests.
e) JavaScript is used to program the behavior of web pages.

Server Side:

a) Python programming language is used to implement the core program logic.
b) Numpy is to manipulate the large multidimensional arrays and matrices.
c) Pandas is used for data manipulation and analysis.
d) NLTK is used for natural language processing which contains text processing libraries.
e) Scikit-Learn is used to implement the machine learning algorithm.

This section describes the technologies used in Tweets Sentiment Analysis Using Logistic Regression Algorithm. Tweets Sentiment Analysis Using Logistic Regression Algorithm is a web application that uses flask python framework. HTML, Twitter Bootstrap CSS, JavaScript are used to develop front-end and python, NLTK, Scikit-learn are used to develop back-end. HTML is used for presentation technology. JavaScript are implemented to show the result of the application in a dynamic way.

All the algorithms for the application are written in Python. Algorithms used in Sentiment Analysis of Twitter Data Using Logistic Regression is Predictive analysis model. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables. The algorithm is codded in python programming language.

4.2 Description of Major Function

The major function in the application are:

4.2.1 Preprocessing

This is the function which is run for processing the tweets.
Input: It takes the inputs as tweets
Process: It call other function like remove URL, filter stop words, etc.
Output: It gives the list of the process tweets.

4.2.2 Feature extractor

This function is implemented after the preprocessing data.
Input: This takes pre-processed data as in input.
Process: It then uses the method, extract_feature () to process the taken input, process and extract the feature.

4.2.3 Classifier

This class implements the Logistic Regression algorithm which take feature from Feature extractor.
Input: It takes input from the feature extractor.
Process: It classify the tweets as positive or negative and return the list of tweets with its sentiment value.
Output: It gives classified tweets with positive and negative tweets value.

4.3 Testing

Among the total data 85% of the data is used for training and 15% is used for testing. For testing hit and trial method is followed and the testing module from Scikit-learn is used for testing.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Tweets Sentiment Analysis Using Logistic Regression Algorithm was successfully implemented using Flask python framework. The accuracy is quite low and can be improve by providing more datasets.

5.2 Recommendation

This project did not consider the other classifier algorithm. Unigram bag of words is only consider. Further accuracy can be obtained through the combined used of the unigrams + bigrams in support vector machine algorithm (Pang, Lee, & Vaithyanathan, 2002).

REFERENCES

Go, A., Bhayani, R., & Lee, H. (n.d.). *Twitter Sentiment Classification using Distant Supervision.*

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning.* on Empirical Methods in Natural Language Processing (EMNLP).

Gautam G. , Yadav P. , Sentiment Analysis of Twitter Data using Machine Learning Approaches and Semantic Analysis.

ravikiranj. (n.d.). *how to build a twitter sentiment analyzer ?*
Retrieved from ravikiranj.net:
https://www.ravikiranj.net/posts/2012/code/how-build-twitter-sentiment-analyzer/

Stanford.(n.d.). *For Academics*. Retrievd from Sentient140 :
http://help.sentiment140.com/for-students/

Waykar, P., Wadhwani, K., Pooja, M., & Kollu, A. (2016). *Sentiment Analysis of Twitter tweets using supervised.* Int. Journal of Engineering Research and Applications

Jason Brownlee , Getting Started with Applied Machine Learning?
Retrived from: https://machinelearningmastery.com/blog/

APPENDIX

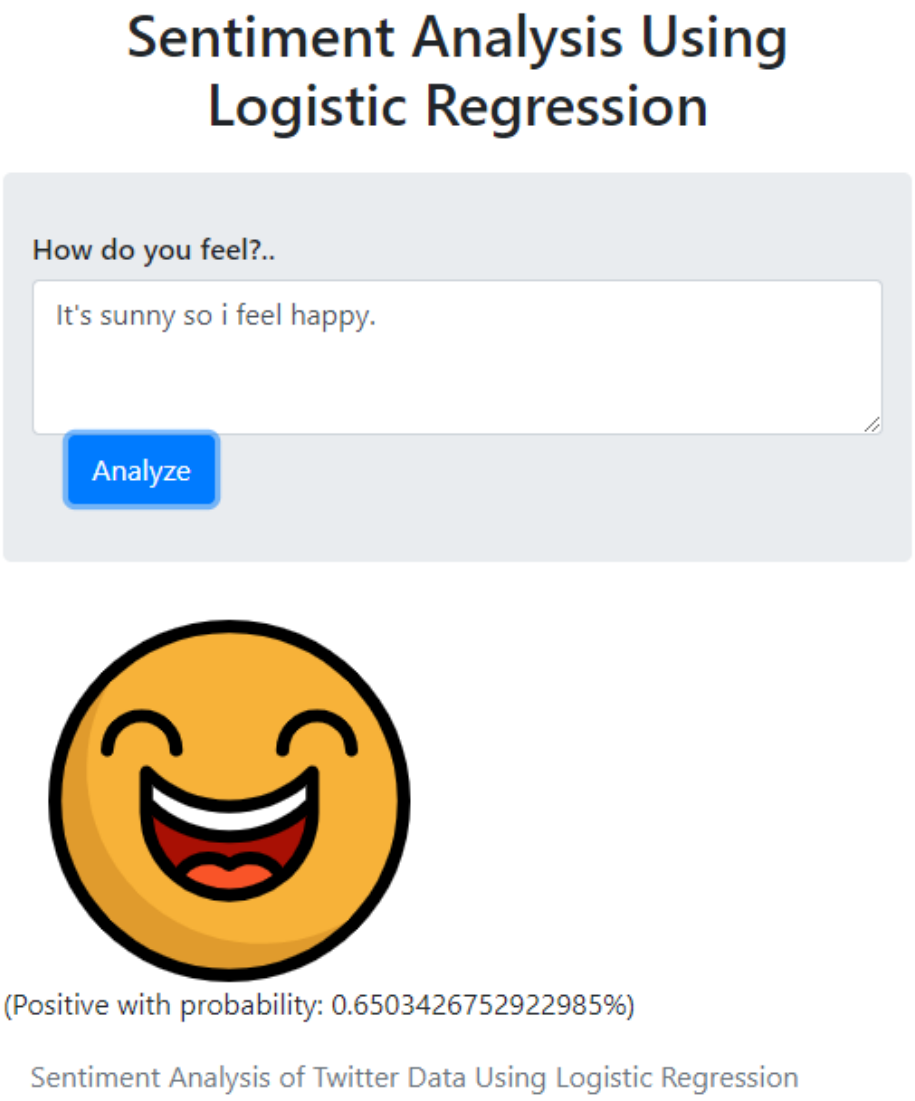

Figure 5 – Sentiment Analysis Result