Sentiment Analysis of Twitter Feed with Dual Training Using Machine Learning Techniques

CS18L1 Project

12140819 CSU13216 Christin Wilson
12140903 CSU13242 Reynah Maria John
12140857 CSU13256 Tony Simon Akkara
12140917 CSU13259 Varuna C Dev
B. Tech Computer Science & Engineering

Department of Computer Engineering
Model Engineering College
Thrikkakara, Kochi 682021
Phone: +91.484.2575370
http://www.mec.ac.in
hodcs@mec.ac.in

April 2017

Model Engineering College Thrikkakara
Department of Computer Engineering

C E R T I F I C A T E

This is to certify that, this report titled ***Project title*** is a bonafide record of the work done by

12140819 CSU13216 Christin Wilson
12140903 CSU13242 Reynah Maria John
12140857 CSU13256 Tony Simon Akkara
12140917 CSU13259 Varuna C Dev
Eighth Semester B. Tech. Computer Science & Engineering

students, for the course work in **CS18L1 Project**, which is the second part of the two semester project work, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, B. Tech. Computer Science & Engineering of **Cochin University of Science & Technology**.

Guide

Sheena Y
Assistant Professor
Computer Engineering

Coordinator

Dr. Priya S
Professor
Computer Engineering

Head of the Department

Manilal D L
Associate Professor
Computer Engineering

April 20, 2017

Acknowledgements

This project would not have been possible without the kind support and help of many individuals. We would like to extend my sincere thanks to all of them.

First of all, We would like to thank our esteemed Principal, Prof. (Dr.) V.P Devassia, for his guidance and support in maintaining a calm and refreshing environment to work in and also for providing the facilities that this work demanded.

We are highly indebted to our Project Coordinator, Dr. Priya S, Professor and Head of the Department, Dr. Manilal D L, Associate Professor for their guidance, support and constant supervision throughout the duration of the work as well as for providing all the necessary information and facilities that this work demanded.

We would like to thank our Project Guide, ... for his/her support and valuable insights and also for helping me out in correcting any mistakes that were made during the course of the work.

We offer our sincere gratitude to all our friends and peers for their support and encouragement that helped me get through the tough phases during the course of this work.

Last but not the least, we thank the Almighty God for guiding me through and enabling us to complete the work within the specified time.

Christin Wilson

Reynah Maria John

Tony Simon Akkara

Varuna C Dev

Abstract

With the growing volume of online reviews available on the Internet, sentiment analysis and opinion mining, as a special text mining task for determining the subjective attitude (i.e., sentiment) expressed by the text, is becoming a hotspot in the field of data mining and natural language processing. Twitter has provided an effective way to expose the public sentiment which is used for decision making in various domains. Also, the general practice remains limited due to some fundamental deficiencies in handling the polarity shift problem. Thus, a sentiment-reversed review can also be used used for training the classifier. On this basis, a dual training algorithm to make use of original and reversed training reviews in pairs for learning a sentiment classifier is used and applied to predict the sentiment of reviews.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the rapid development of internet, demand of online data analysis becomes key role in all areas. Sentiment analysis and Opinion mining involves the study of opinions and its related concepts such as sentiments, evaluations, attitudes and emotions. It is widely used in Data mining, Web mining, Text mining and Natural Language Processing. Data mining is the analysis step of the "knowledge discovery in databases" process (KDD). During the decision making process, "what people think" has always been an important piece of information. In the past, when an individual needed to make a decision, he typically asked for opinions from friends and family. Now it depends on online reviews. Sentiment Analysis (SA) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. SA is the text mining task for subjective attitude. Sentiment classification is a basic task in sentiment analysis, with its aim to classify the sentiment (e.g., positive or negative) of a given text.

Social network is a platform for connecting people to share information and it is an opportunity to study the propagation of ideas. With the growing volume of online reviews available on the Internet, sentiment analysis and opinion mining is becoming a hotspot in the field of data mining and natural language processing.Twitter is a worldwide popular website that offers a social network and micro blogging services, which enable the users to update their status in tweets and follow the people who are interested in retweeting posts and even communicate with them directly. Sentiment analysis on Twitter has provided an economical and effective way to expose the public sentiment, which is used for decision making in various domains.

Also, general practice in sentiment classification follows the techniques in traditional topic-based text classification, where the bag-of-words (BOW) model is typically used for text representation. In the BOW model, a review text is represented by a vector of independent words. The Polarity classification is the most classical sentiment analysis task which aims at classifying reviews into either positive or negative. Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. Negation is the most important type of polarity shift. For example, by adding a negation word "don't" to a positive

text "I like this book" in front of the word "like", the sentiment of the text will be reversed from positive to negative. The scope of this project is to create a sentiment classifier over twitter feed and reduce the polarity shift problem using dual training of the classifier.

1.1 Proposed Project

1.1.1 Problem Statement

With the growing volume of online reviews available on the Internet, sentiment analysis and opinion mining, as a special text mining task for determining the subjective attitude (i.e., sentiment) expressed by the text, is becoming a hotspot in the eld of data mining and natural language processing. Twitter has provided an eective way to expose the public sentiment which is used for decision making in various domains. Also, the general practice remains limited due to some fundamental deciencies in handling the polarity shift problem. Thus, a sentiment-reversed review is used for each training. On this basis, a dual training algorithm to make use of original and reversed training reviews in pairs for learning a sentiment classier is used and applied to predict the sentiment of reviews.

1.1.2 Proposed Solution

Social network is a platform for connecting people to share information and it is an opportunity to study the propagation of ideas. With the growing volume of online reviews available on the Internet, sentiment analysis and opinion mining is becoming a hotspot in the eld of data mining and natural language processing. Twitter is a worldwide popular website that oers a social network and micro blogging services, which enable the users to update their status in tweets and follow the people who are interested in retweeting posts and even communicate with them directly. Sentiment analysis on Twitter has provided an economical and effective way to expose the public sentiment, which is used for decision making in various domains. We propose an approach to sentiment analysis using Natural Language Processing.To address the problems regarding conventional methods for finding reviews, we suggest a system that will analyse the user generated reviews from social media platforms and come to a conclusion, about what most people think about the query user entered. It is an approach which can accurately do sentimental analysis about user submitted topics or tweets . The project involves categorising the tweets into positive, neutral or negative reviews and summarising the response. Dual training is introduced to reduce the errors that may occur such as the polarity shift problem. The input to the system will be given by the users.The user can enter their tweets , queries or the topics that they want to analyse.Based on the topic given by the user, tweets are collected , analysed and is classied to positive, negative or neutral.The tweets along with the sentiment is shown as output.

Chapter 2

System Study Report

2.1 Literature Survey

1. [1] introduces a simple yet efficient model, called Dual Sentiment Analysis (DSA). DSA is proposed to address the polarity shift problem in sentiment classification. By using the properly that sentiment classification has two opposite class labels (i.e., positive and negative), initially a data expansion technique is proposed by creating sentiment-reversed reviews. The original and reversed reviews are constructed in a one-to-one correspondence. Thereafter, a Dual Training (DT) algorithm and a Dual prediction (DP) algorithm respectively, to make use of the original and reversed samples in pairs for training a statistical classifier and make predictions.

2. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed in Twitter is 140. Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text. [2] tries to analyze the twitter posts about electronic products like mobiles, laptops etc using Machine Learning approach. By doing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification.Machine learning approach makes use of a training set to develop a sentiment classifier that classifies sentiments. Since a predefined database of entire emotions is not required for machine learning approach, it is rather simpler than Knowledge base approach.There are different feature extraction methods for collecting relevant features from text which can be applied to tweets also. But the feature extraction is to be done in two phases to extract relevant features. In the first phase, twitter specific features are extracted. Then these features are removed from the tweets to create normal text. After that, again feature extraction is done to get more features. This is the idea used in this paper to generate an efficient feature vector for analyzing twitter sentiment.

3. .[3] contributes to the sentiment analysis for customers review classification which is helpful to analyze the information in the form of the number of tweets where opinions

are highly unstructured and are either positive or negative, or somewhere in between of these two.

For this first dataset is pre-processed , after that the adjective from the dataset that have some meaning is extracted which is called feature vector, then the feature vector list is selected and thereafter machine learning based classification algorithms are applied. The preprocessor is applied to the raw sentences which make it more appropriate to understand. The different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.

4. [4] considers the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, they find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. They conclude by examining factors that make the sentiment classification problem more challenging.

5. [5] introduces a learning based method of sentiment classification of sentences using word-level polarity.The polarities of words in a sentence are not always the same as that of the sentence, because there can be polarity-shifters such as negation expressions.The model can be trained in two different ways: word-wise and sentence-wise learning. In sentence-wise learning, the model can be trained so that the prediction of sentence polarities should be accurate. The model can also be combined with features used in previous work such as bag-of-words and n-grams. It is empirically shown that the method almost always improves the performance of sentiment classification of sentences especially when there is only small amount of training data.

6. [6] proposes a feature selection method to automatically generate a large scale polarity shifting training data for polarity shifting detection of sentences. Then, a classifier combination method is presented for incorporating polarity shifting information. Compared with previous ones, this approach highlights the following advantages. First of all, a binary classifier is applied to detect polarity shifting rather than merely relying on trigger words or phrases. This enables the approach to handle different kinds of polarity shifting phenomena. More importantly, a feature selection method is presented to automatically generate the labeled training data for polarity shifting detection of sentences. Polarity shifting marked by various linguistic structures has been a challenge to automatic sentiment classification. In this paper, a machine learning approach to incorporated polarity shifting information into a document-level sentiment classification system is proposed.

7. [7] presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a

review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., subtle nuances) and a negative semantic orientation when it has bad associations (e.g., very cavalier).the semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word excellent minus the mutual information between the given phrase and the word poor. A review is classified as recommended if the average semantic orientation of its phrases is positive.

8. With the explosive growth of user generated messages, Twitter and Facebook has become a social site where millions of users can exchange their opinion. Sentiment analysis affords as an economical and effective way to expose public opinion timely, which is critical for decision making in various domains. Similarly, due to the large volume of opinions, rich web resources such as discussion forum, review sites, blogs and news etc are available in digital form; much of the current research is focusing on the area of sentiment analysis. The information is very useful for businesses for marketing, governments and individuals. While this content meant to be helpful, analyzing the bulk of user generated content is difficult and time consuming. In the existing system there is no way of analysis and ranking the user opinions and sometimes they consider the individual opinions without conducting any reviews. [8] proposes an intelligent system which automatically extract such huge content and classify them into positive, negative and neutral type using Artificial Neural Networks based on the specified criteria.

9. Ideally, an opinion mining tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good). Begin by identifying the unique properties of this problem and develop a method for automatically distinguishing between positive and negative reviews. This classifier [9] draws on information retrieval techniques for feature extraction and scoring, and the results for various metrics and heuristics vary depending on the testing situation. The best methods work as well as or better than traditional machine learning. When operating on individual sentences collected from web searches, performance is limited due to noise and ambiguity. But in the context of a complete web-based tool and aided by a simple method for grouping sentences into attributes, the results are qualitatively quite useful.

10. [10] proposes a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus and one domain-independent corpus. This disparity is captured via a measure called domain relevance (DR), which characterizes the relevance of a term to a text collection. They first extract a list of candidate opinion features from the domain review corpus by defining a set of syntactic dependence rules. For each extracted candidate feature, they then estimate its intrinsic-domain relevance (IDR) and extrinsic-domain relevance (EDR) scores on the domain-dependent and domain-independent corpora,

respectively. Candidate features that are less generic (EDR score less than a threshold) and more domain-specific (IDR score greater than another threshold) are then confirmed as opinion features. They call this interval thresholding approach the intrinsic and extrinsic domain relevance (IEDR) criterion. Experimental results on two real-world review domains show the proposed IEDR approach to outperform several other well-established methods in identifying opinion features.

2.2 Proposed System

We propose a sentiment analysis tool using Natural Language Processing. To address the problems regarding conventional method for finding reviews, we suggest our system that will analyse the user generated data or reviews from twitter and come to a single conclusion, about what most people think about the query user entered. It is a tool which can accurately do sentimental analysis about user submitted topics or tweets .The project makes use of dual training which will increase the accuracy of the results. The project involves categorizing the tweets into positive, neutral or negative review and summarizes and visualizes the response.

Input : The input to the system will be given by the users. The web application will contain an input text area were user can enter their tweets , queries or the topics that they want to analyse by clicking on the button Search. These topics can vary from movie reviews, product reviews, socio-religious issues, political issues etc. The application will also have a limit field in which the user can specify the number of tweets that he wants to analyse. Based on the topic given by the user, tweets are collected , analysed and is classified to positive, negative or neutral. The user query field in mandatory because the analysis process cannot be commenced without a search topic

Output : The system classifies the tweets as positive, negative or neutral. The tweets along with the sentiment is shown as output.

ftp

2.2.1 Advantages Of Proposed System

- Not domain specific
- SVM classifier is used which is better than naive bayes and other classifiers used by the existing system
- User friendly
- Dual training reduces errors and improves accuracy

No.	Paper	Techniques	Advantages	Disadvantages
1.	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews by Peter D. Turney	*Pointwise Mutual Information (PMI) and Information Retrieval (IR) algorithm. *Compare similarity of review to a positive reference word (excellent) with its similarity to a negative reference word (poor).	Works by comparing the sentiment of a word.	Not very accurate. Needs four queries to calculate the semantic orientation of a phrase so, is slow.
2.	Dual Sentiment Analysis: Considering Two Sides of One Review By Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li .	Applies Dual Training and Dual Prediction to solve the polarity shift poblem	Solves the polarity shift problem.	Uses a test data set.
3.	Sentiment Analysis in Twitter using Machine Learning Techniques	A dataset is created using twitter posts of electronic products *performs a sentence level sentiment analysis.	Identifies sentiment for a given tweet.	On a very particular domain.
4.	Thumbsup? Sentiment Classification using Machine Learning Techniques	- A dataset of movie reviews from imdb fed as input to classfier. - 3 classifiers where used such as SVM, Naive Bayes and Maximum entropy.	- A dataset of movie reviews from imdb fed as input to classfier. - 3 classifiers where used such as SVM, Naive Bayes and Maximum entropy.	Poor accuracy results as worked on a limited trained dataset.

No.	Paper	Techniques	Advantages	Disadvantages
5.	Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis	a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data.	a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data.	a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data.
6.	Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis	Feature selection method to automatically generate a large scale polarity shifting training data for polarity shifting detection of sentences. classifier combination method- to incorporate polarity shifting information.	Feature selection method to automatically generate a large scale polarity shifting training data for polarity shifting detection of sentences. classifier combination method- to incorporate polarity shifting information.	Automatically generated polarity shifting training data is prone to noise.
7.	Sentimental Data Analysis on Social Media by Priyanka. B J.T. Thirukrishna	A sentimental data analysis model is proposed using Neural Networks-extracting data from social media and segregate the comments	Easy, simple and user friendly to use all types of sentiment analysis models.	-Platform dependent. No negation handling.

No.	Paper	Techniques	Advantages	Disadvantages
8.	Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews	The classifier draws on information retrieval techniques for feature extraction and scoring, and the results for various metrics and heuristics vary depending on the testing situation.	Uses metadata substitutions, variable length features and best substring algorithm which gives more accuracy	Extraction is more difficult
9.	Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance	A intercorpus statistics approach to opinion feature extraction based on the IEDR feature-filtering criterion is used	Outperforms other methods like IDR, EDR, LDA, ARM, MRC, and DP.	Neutral opinions will not be considered.
10.	Learning to Shift the Polarity of Words for Sentiment Classification	Learning to Shift the Polarity of Words for Sentiment Classification	Good, when there is only small amount of training data.	Requires a classifier for itself.

Table 2.1: Literature Survey

	Positive	Negative	Neutral	Total
Training	9667	9667	2323	21657

Table 2.2: Statistics of the Dataset used

Chapter 3

Software Requirement Specification

3.1 Introduction

3.1.1 Purpose

A method which contributes to the sentiment analysis for customers review classification which is helpful to analyze the information in the form of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two.

3.1.2 Document Conventions

- Entire document should be justified.
- Convention for Main title
  - Font face: Times New Roman
  - Font style: Bold
  - Font size: 14
- Convention for Sub title
  - Font face: Times New Roman
  - Font style: Bold
  - Font size: 12
- Convention for body
  - Font face: Times New Roman
  - Font size: 12

3.1.3 Intended Audience and Reading Suggestions

The audience targeted is wide scale, they could be data miners, members of marketing team, customers interested in a particluar product. Sentiment is a crucial factor to discover how people feel about a particular topic.

3.1.4 Project Scope

The project aims to create a system that can perform sentiment analysis on a user submitted query regarding a relevant topic or product, which can help data miners, members of marketing team, customers interested in a particluar product.

3.1.5 Overview of Developer's Responsibilities

The Developers are responsible for the building of this project, its problems and solutions and also for assuring its fitness for usage by customers.They must ensure that the software is in a usable state and free of bugs. The main feature that will be implemented is sentiment analysis .Also the product aims at being interactive in nature and at providing personalised outputs.

3.2 Overall Description

3.2.1 Product Perspective

There is a huge demand of online data analysis with the growing amounts of reviews available. This system implements sentiment analysis to study of opinions and its related concepts such as sentiments, evaluations, attitudes and emotions.Sentiment Analysis (SA) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.
Twitter has provided an effective way to expose the public sentiment which is used for decision making in various domains. Thus Twitter is used as a source to obtain reviews. The reviews here will help with understanding the opinion of a majority of people so as to help in decision making.
In the current senario, the polarity shift problem is not considered.Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. It causes errors in the anaysis of reviews. This system implements a dual training which reduces the polarity shift problem during the sentiment analysis.

3.2.2 Product Functions

- The user inputs a query to the system.
- An API call is made to Twitter to obtain tweets based on the query made by the user. The tweets are used as reviews to identify the sentiment of each user over the query.
- The reviews are passed through a sentiment classifier which identifies the sentiment of the query. The sentiment maybe positive or negative.
- An ouput to the user is created to show the reviews along with their sentiments.

3.2.3 User Classes and Characteristics

The user can be anybody who needs to know the sentiment of the general public that uses the web. It can be from any age group who may require to make a decision based on the opinion of the public.

3.2.4 Operating Environment

- Operating system: Linux
- Languages used: Python
- Processor: Intel Core i5

- RAM: 6GB
- Hard disk: 400GB

3.2.5 Design and Implementation Constraints

- There might be junk tweets which are irrelevant to the analysis as they may be created by bots.
- Tweets may contain sarcasm which will be hard to identify. The sarcasm can create an erroneous sentiment analysis of the tweet.
- The Twitter API allows only a limited number of calls in an hour.
- Creating an accurate algorithm to reverse the sentiment of a Tweet.

3.2.6 User Documentation

A help manual giving instructions on the use of the product will be given in the webpage.

3.2.7 General Constraints

- A strong internet connection is required at all times to retrieve the Tweets.

3.2.8 Assumptions and Dependencies

The proper working of the system depends on:

- Internet connection.
- Friendly interface.
- Acquiring relevent tweets.
- Proper queries by the user.

3.3 External Interface Requirements

3.3.1 User Interfaces

The user interface is a webpage of the product. The user is asked to input his/her queries. Output is also shown on the Web Browser.

3.3.2 Hardware Interfaces

The hardware should have the following specifications:

- Computer with proper internet connection
- Continuous power supply.

3.3.3 Software Interfaces

- Operating System: Windows, Linux, MacOS.
- Programming Languages: Python
- Input Data: User query

3.3.4 Communication Interfaces

The Web Page takes input from the user. It then uses Twitter APIs to get the required tweets in order to do Sentiment Analysis on the tweets based on the given query.

3.4 Hardware and Software Requirements

3.4.1 Hardware Requirements

- High performance system
- Strong internet connection

3.4.2 Software Requirements

3.4.3 Twitter API

The Twitter Search API is part of Twitters REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days. Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results.

3.4.4 Python

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale.

3.4.5 Django-Web Application

Django is a free and open source web application framework, written in Python, which follows the model view controller (MVC) architectural pattern.It is maintained by the Django Software Foundation (DSF), an independent organization established as a 501(c) non-profit. Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. Its free and open source. Its features are:
1. Ridiculously fast - Django was designed to help developers take applications from concept to completion as quickly as possible.
2. Reassuringly secure - Django takes security seriously and helps developers avoid many common security mistakes.
3. Exceedingly scalable - Some of the busiest sites on the Web leverage Djangos ability to quickly and flexibly scale.

3.4.6 Web Browser

A web browser (commonly referred to as a browser) is a software application for retrieving, presenting, and traversing information resources on the World Wide Web. An information resource is identified by a Uniform Resource Identifier (URI/URL) and may be a web page, image, video or other piece of content

3.4.7 HTML5

HTML5 is a markup language used for structuring and presenting content on the World Wide Web. It is the fifth and current version of the HTML standard. Web browsers receive HTML documents from a webserver or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

3.4.8 Operating System

An operating system (OS) is system software that manages computer hardware and software resources and provides common services for computer programs. All computer programs, excluding firmware, require an operating system to function.

3.5 Functional Requirements

Major functional requirements include:

3.6 User Input

The user is to input queries. Based on the queries that the user inputs, the tweets are received and analysed. The user may enter queries on a range of topics that the general public would tweet about in Twitter. The sentiment of the users of Twitter will be analysed based on the input.

3.7 Tweet Collection

Tweets are collected from Twitter using a Twitter API call. Only a limited number of tweets can be taken every hour. The tweets collected are what is going to be used for the Sentiment Analysis.

3.8 Data Cleaning

The tweets are limited to a size of 150 characters in which there will be terms which are not required so we need to extract relevant information.

3.9 Sentiment Analysis

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine the writer's attitude towards a particular topic. The tweets are anallysed to obtain the sentiment of the tweet. This is done with multiple tweets so as to know the sentiment of the general public on Twitter on a topic.

3.10 Non-functional Requirements

3.11 Performance Requirements

The software will provide up-to-date information, limited only by the rate of Twitter input.Any lags should be notified to the user.Resource consumption of this application should not reach an amount that renders the desktop PC or mobile device unusable.The application should be capable of operating in the background should the user wish to utilize other applications.

3.12 Safety Requirements

It is illegal to maintain tweets in any database for privacy issues, hence we are not saving the tweets in our database. There is no seperate authentication required from user end.

3.13 Security Requirements

The privacy of the twitter users should be maintained. The system does not collect any personal information about twitter users.The system is trained with available tweets and developer inputs.

3.14 Software Quality Attributes

3.14.1 Reliability

The software will meet all of the functional requirements without any unexpected behaviour. At no time should the graph output display incorrect or outdated information without alerting the user to potential errors.

3.14.2 Availability

The software will be available at all times on the users device, as long as the device is in proper working order. The functionality of the software will depend on any external services such as internet access that are required. If those services are unavailable, the user should be alerted.

3.14.3 Security

The software should never disclose any personal information of Twitter users, and should collect no personal information from its own users.

3.14.4 Maintainability

The software should be written clearly and concisely. The code will be well documented. Particular care will be taken to design the software modularly to ensure that maintenance is easy.

3.14.5 Portability

This software will be designed to run on any web browsers or operating system.

3.15 Other Requirements

The user must have a web browser and internet connectivity to use the product. The major requirement here is the Tweepy API which gives the Tweets of the general public to be reviewed.

Chapter 4

System Design

4.1 System Architecture

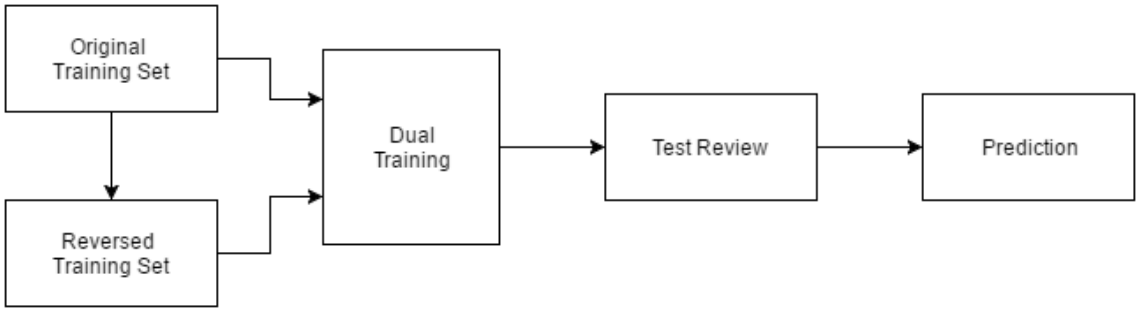

Figure 4.1: Overall System Design

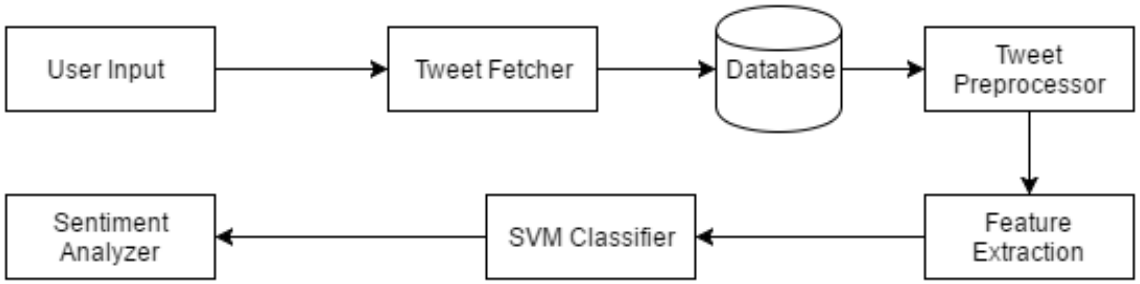

Figure 4.2: System Working Model

4.2 Input Design

The input that is taken from the user is the query to be searched and the number of tweets that should be fetched from the twitter API. The user should enter the input on a webpage.

Once the processing is done, the user will be redirected to a new webpage which will display the results.

4.3 Libraries and Packages Used

- nltk
  - The Natural Language Toolkit (NLTK) is an open source Python library for Natural Language Processing.
  - It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
- wordnet
  - WordNet is a lexical database for the English language.
  - It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.
  - WordNet can thus be seen as a combination of dictionary and thesaurus.
- re
  - A regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern.
  - The module re provides full support for Perl-like regular expressions in Python.
- csv
  - CSV files are used to store a large number of variables or data.
  - The CSV module is a built-in function that allows Python to parse these types of files.
- libsvm
  - LIBSVM is a library for Support Vector Machines (SVMs).
  - The goal is to help users to easily apply SVM to their applications.
  - LibSVM allows for sparse training data. That is, the non-zero values are the only ones that are included in the dataset.

4.4 Module Description

4.4.1 User Input

The user input module contains input query given by the user.

4.4.2 Tweet Fetching

The tweets based on the user searched topic or query is collected and downloaded using Tweet downloader. This is stored in the database so that they can be preprocessed and their sentiment could be found out.The user input module takes in input query, tweet limit and stock symbol as user inputs and stores it to the database.

4.4.3 Tweet Pre-processing

Pre-processing of data is the process of preparing and cleaning the data of the dataset for classification. Reducing the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis. Stop word removal - A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include a, of, the, I, it, you, and and these are generally regarded as functional words which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words .Hence it is practical to remove those words which appear too often that support no information for the task.

4.4.4 Dual Training

The classifier is trained by using an original and reversed training set. And original training set is used initially to train the classifier. The training set and label is reversed and used to train the classifier.

4.4.5 SVM Classifier

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these

formed groups.Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector , and we want to know whether we can separate such points with a (p-1)-dimensional hyperplane. This is called a linear classifier.
More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Support vector machines are universal learners. Remarkable property of SVM is that their ability to learn can be independent of dimensionality of feature space. SVM measures the complexity of Hypothesis based on margin that separates the plane and not number of features. SVM learning Algorithms for Text Categorization - SVM has defined input and output format. Input is a vector space and output is 0 or 1 (positive/negative).Tweets in original form are not suitable for learning. They are transformed into format which matches into input of machine learning algorithm input. For this pre-processing on tweets is carried out.Then we carryout transformation.Each word will correspond to one dimension and identical words to same dimension. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings.SVM has been proved one of the powerful learning algorithm for text categorization.

4.4.6 Sentiment Analyzer

The Tweets are obtained by the Tweet Fetching module.The obtained tweets are analyzed to know the opinion of the general public.

4.5 Activity Diagram

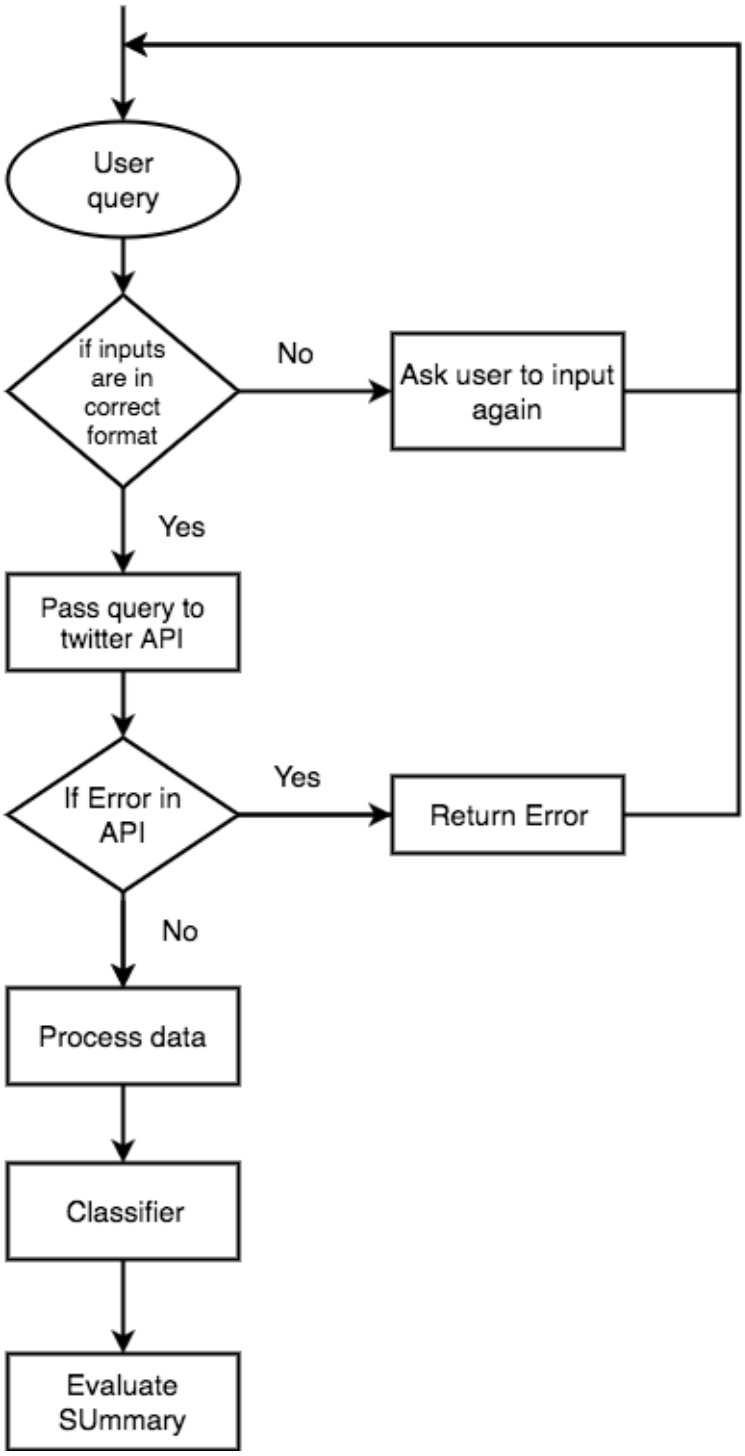

Figure 4.3: Activity Diagram

4.6 Class Diagram

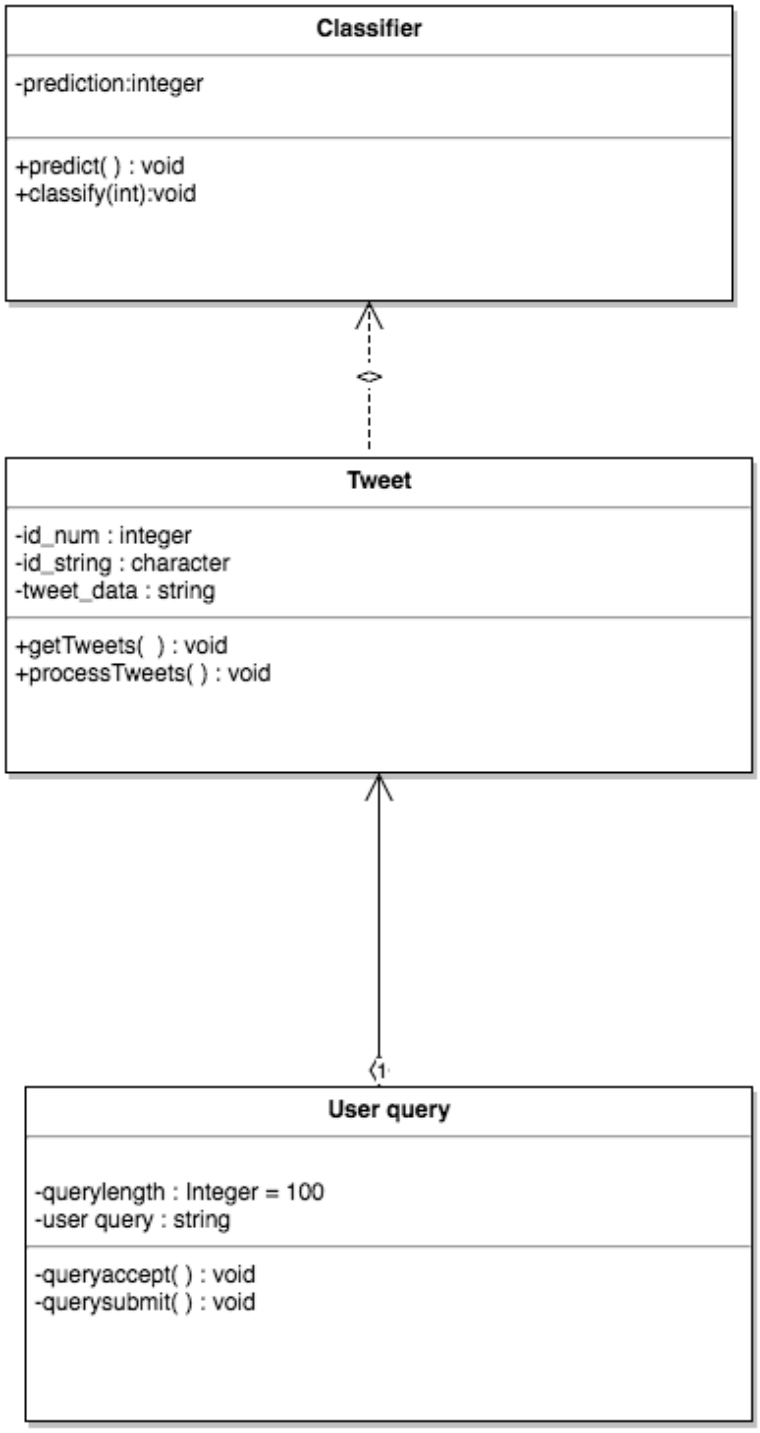

Figure 4.4: Class Diagram

4.7 Use Case Diagram

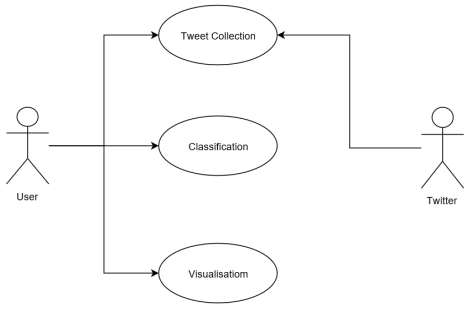

Figure 4.5: Use Case Diagram

Chapter 5

Data Flow Diagram

5.1 Level 0 DFD

Figure 5.1: DFD: Level 0

5.2 Level 1 DFD

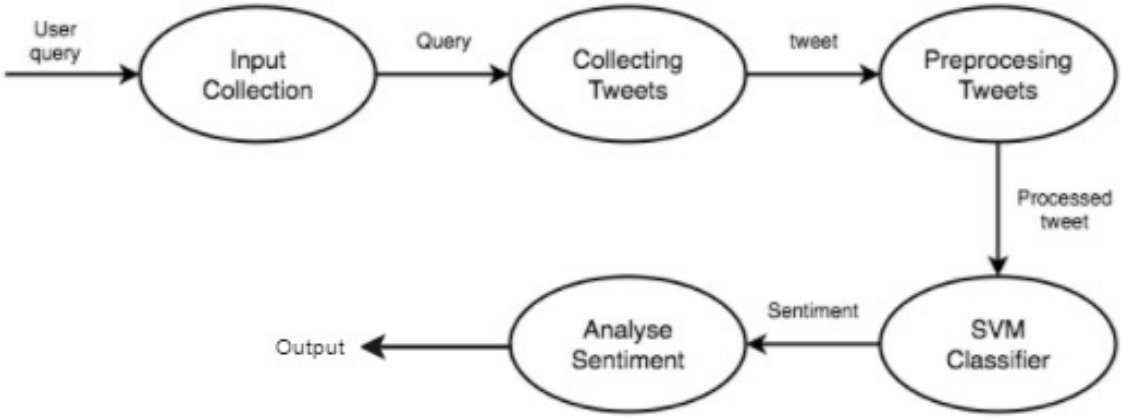

Figure 5.2: DFD: Level 1

5.3 Level 2 DFD

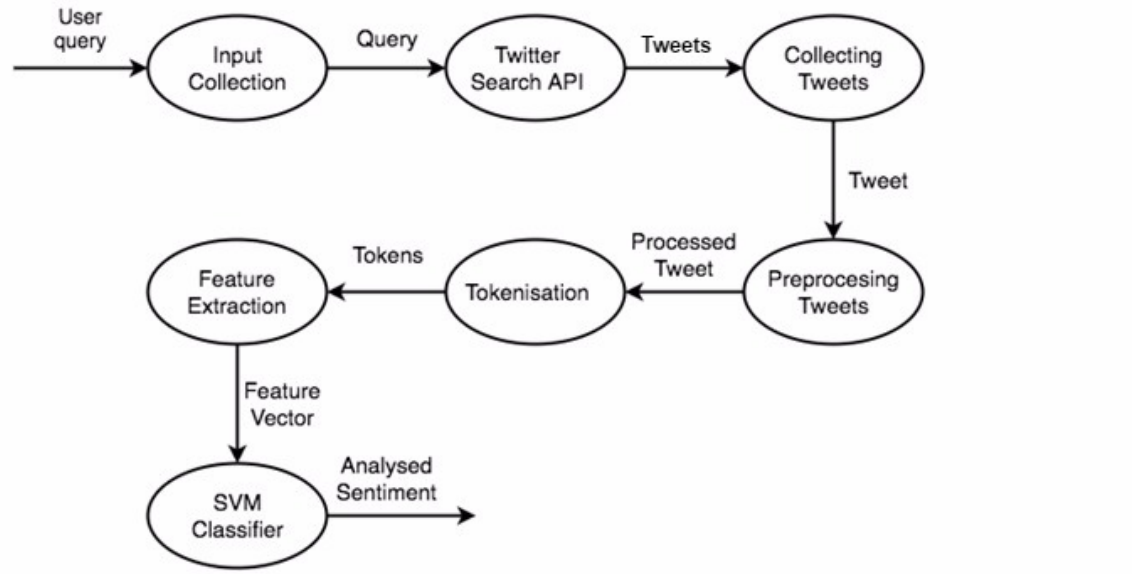

Figure 5.3: DFD: Level 2

Chapter 6

Implementation

6.1 Algorithms

6.1.1 Tweet Collection

1. Receive the query and limit from user input module.
2. Connect to twitter via tweepy API using the consumer key, consumer secret and other required credentials.
3. Download tweets and save it.

6.1.2 Preprocessing

1. Convert tweet to lower case
2. Remove web addresses
3. Convert emoticons to sentiment words
4. Remove punctuations
5. Remove stop words
6. Remove hashtags
7. Remove additional white spaces
8. Append the rest to feature vector

6.1.3 Feature Extraction

1.Extract words to be converted to feature vector 1.1. Split the tweets into words.
1.2. Remove the duplicate words.
1.3. Find the words that are relevant in finding the sentiment of the tweet.
1.4. Save those words as feature words.
2. Create a list containing all the sentiment defining words in the dataset.
3.Sort the above list and name it as sortedlist.
4.Create a map and assign to zero for each word in sortedlist.
5.Now access the tweet and for each word in tweet check if that word is present in the map,

if present assign map[word] as one.
6.Obtain the values of the feature word which are the support vectors.

6.1.4 Text Reversal

1. Read from csv row by row.
2. For the second column of each row
2.1. Reverse the string by calling the reverse() function.
2.2. If the second column is reversed then reverse the first column of the row.
3. Write the reversed data into a csv file.

reverse()
1. Preprocess the text
2. For each word in the text
2.1. If the word is not or nor then continue
2.2. For each synset of word
2.2.1 If pos is a,s or v
2.2.1.1 If lemma of the synset of the word has an antonym set the reversed word as the antonym.
3. Return the reversed text

6.1.5 SVM Training

1. Open the training data set .
2. Extract the sentiment from first column and tweet text from second column.
3. Pass tweet text to preprocessing unit.
4. Featurelist is created which consists of all the unique words present in the training dataset.
5.Obtain the feature vectors and labels to train the svm model.
6.Save the Featurelist variable to a file.
7.Define svm parameter and kernel type.
8.Train the SVM model.
9.Save the model for classification.

6.1.6 SVM Classification

1.Load the SVM trained model.
2.Open the feature list saved while training and also take the tweet words to obtain the feature vector.
3.Feed the feature vector to the loaded trained model.
4.Obtain tweet sentiment.

6.2 Development Tools

- Django Framework: Django is a free and open-source web framework, written in Python, which follows the model-view-template (MVT) architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent organization established as a non-profit. Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models. Some well-known sites that use Django include the Public Broadcasting Service, Pinterest, Instagram, Mozilla, The Washington Times,Disqus,Bitbucket,and Nextdoor.

  Despite having its own nomenclature, such as naming the callable objects generating the HTTP responses views, the core Django framework can be seen as an MVC architecture. It consists of an object-relational mapper (ORM) that mediates between data models (defined as Python classes) and a relational database (Model), a system for processing HTTP re- quests with a web templating system (View), and a regular-expression-based URL dispatcher (Controller).

- Sublime Text: Sublime Text is a proprietary cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and its functionality can be extended by users with plugins, typically community- built and maintained under free-software licenses.

- Git Version Control: Git is a version control system (VCS) for tracking changes in computer files and coordinating work on those files among multiple people. It is primarily used for software development, but it can be used to keep track of changes in any files. As a distributed revision control system it is aimed at speed, data integrity, and support for distributed, non-linear workflows.

- Machine Learning: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed." Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

- SVM: In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

  In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

  When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

  More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

  The original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it

was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function selected to suit the problem. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings.

Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback. This is also true of image segmentation systems, including those using a modified version SVM that uses the privileged approach as suggested by Vapnik.

Hand-written characters can be recognized using SVM.

The SVM algorithm has been widely applied in the biological and other sciences. They have been used to classify proteins with up to 90 percentage of the compounds classified correctly. Permutation tests based on SVM weights have been suggested as a mechanism for interpretation of SVM models. Support vector machine weights have also been used to interpret SVM models in the past. Posthoc interpretation of support vector machine models in order to identify features used by the model to make predictions is a relatively new area of research with special significance in the biological sciences.

- LIBSVM: LIBSVM and LIBLINEAR are two popular open source machine learning libraries, both developed at the National Taiwan University and both written in C++ though with a C API. LIBSVM implements the SMO algorithm for kernelized support vector machines (SVMs), supporting classification and regression. LIBLINEAR implements linear SVMs and logistic regression models trained using a coordinate descent algorithm.

  The SVM learning code from both libraries is often reused in other open source machine learning toolkits, including GATE, KNIME, and scikit-learn. Many bindings to it exist for programming languages such as Java, MATLAB and R.

- Google chart: The Google Chart API is an interactive Web service (now deprecated) that creates graphical charts from user-supplied data. Google servers create a PNG image of a chart from data and formatting parameters specified by a user's HTTP request. The service supports a wide variety of chart information and formatting. Users may conveniently embed these charts in a Web page by using a simple image tag.

Originally the API was Google's internal tool to support rapid embedding of charts within Google's own applications (like Google Finance for example). Google figured it would be a useful tool to open up to web developers. It officially launched on December 6, 2007. Currently, line, bar, pie, and radar charts, as well as Venn diagrams, scatter plots, sparklines, maps, google-o-meters, and QR codes are supported. Google deprecated the API in 2012 with guaranteed availability until April 2015. Google now reserves the right to turn it off without notice, although as of April 2016, there are no plans to do so. Google recommends the successor service Google Charts.

Chapter 7

Testing

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. It involves the execution of a software component or system component to evaluate on or more properties of interest. In general, these properties indicate the extent to which the component or system under test : meets the requirements that guided its design and development. responds correctly to all kids of inputs. performs its functions within an acceptable time. is sufficiently usable, can be installed and run in its intended environments. achieves the general result its stakeholders desire.

7.1 Testing Methodologies

- Blackbox Testing : It is the testing process in which the tester can perform testing on an application without having any internal structural knowledge of the application. Usually, test engineers are involved in the blackbox testing
- Whitebox Testing : It is the testing process which tests the internal structure or working of an application, as opposed to its functionality. In whitebox testing, an internal perspective of the system, as well as programming skills, are used to design test cases.Usually, the developers are involved in whitebox testing
- Graybox Testing : It is the process in which the combination of blackbox and whitebox testing techniques are used.

A testcase in software engineering is a set of conditions or variables under which a a tester will determine whether an application or software system is working as per the requirements specified or not. We have used Whitebox testing in the testing phase of the project.

7.2 Unit Testing

In software engineering, unit testing is a software testing method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, are tested to determine whether they are fit for use.

7.2.1 Input

- Test Case
  - Test case deals with entering the user query and the limit.
  - Test case should check for proper inputs and make sure that all the required necessary information is entered by the user
- Verdict
  - The correct form of inputs were obtained.
  - Input was correctly processed.

7.2.2 Tweet Collection

- Test Case
  - Check to see whether user input was successfully obtained and sent to twitter API and correct tweets are obtained.
- Verdict
  - The Module collected an array of tweet objects from the twitter servers successfully.

7.2.3 Preprocessing

- Test Case
  - An array of tweets should be taken and should return preprocesed tweets
- Verdict
  - The tweets are preprocessed successfully.

7.2.4 Feature Extraction

- Test Case
  - An array of preprocessed tweet should be taken and it should return an array of feature vectors objects corresponding to each tweet object.
- Verdict
  - Feature vectors were obtained successfully.

7.2.5 Dual Training

- Test Cases
  - A CSV file is given as input and output should be a reversed sentiment in the CSV file.
- Verdict
  - The reversed sentiments were obtained successfully.

7.2.6 SVM Classification

- Test Cases
  - Feature vectors are taken as inputs, and should return an array of sentiment values corresponding to each feature vector.
- Verdict
  - Classification is done successfully.

7.2.7 Result Visualization

- Test Case
  - An array of sentiment values are taken in as input and result should be visualized as pie chart.
- Verdict
  - The result was succesfully obtained and a pie chart was displayed.

7.3 Integration Testing

Integration testing is the phase in software testing in which individual software modules are combined and tested as a group.The purpose of this level of testing is to expose faults in the interaction between integrated units.Integration Testing is performed after Unit Testing and before System Testing. All the modules of the project were integrated and tested and the output was successfully obtained.

7.4 System Testing

System Testing is a level of the software testing where a complete and integrated software is tested.The purpose of this test is to evaluate the systems compliance with the specified requirements.All the integrated students were system tested a couple of times and output was successfully obtained.

Chapter 8

Results

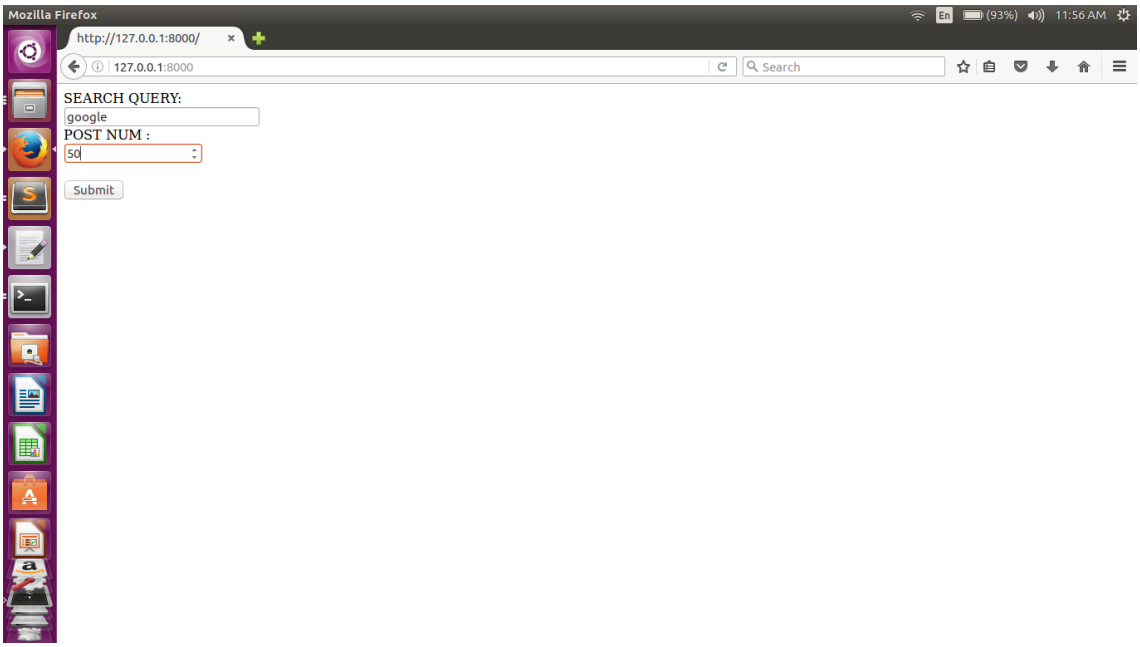

Figure 8.1: Screenshot Of Front End Webpage

This is the webpage that the user sees. The user should input the query and the number of tweets to be queried. The user should submit the query by clicking on the button.

Figure 8.2: Screenshot Of Tweet Collection Module

The tweet collection module takes in the user query and uses tweepy to collect the tweets. It then returns the tweets in a file.

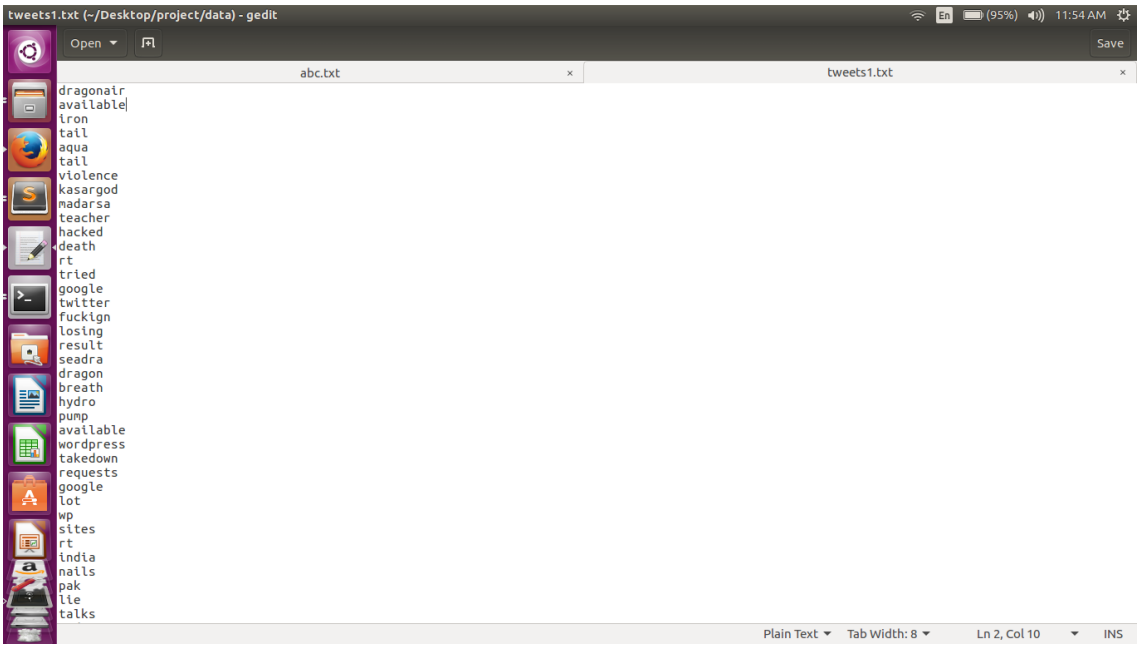

Figure 8.3: Screenshot Of Feature Vector Module

The feature vector is used to build a model which the classifier learns from the training data and further can be used to classify previously unseen data.The tweetwords are then filtered and saved in a file.

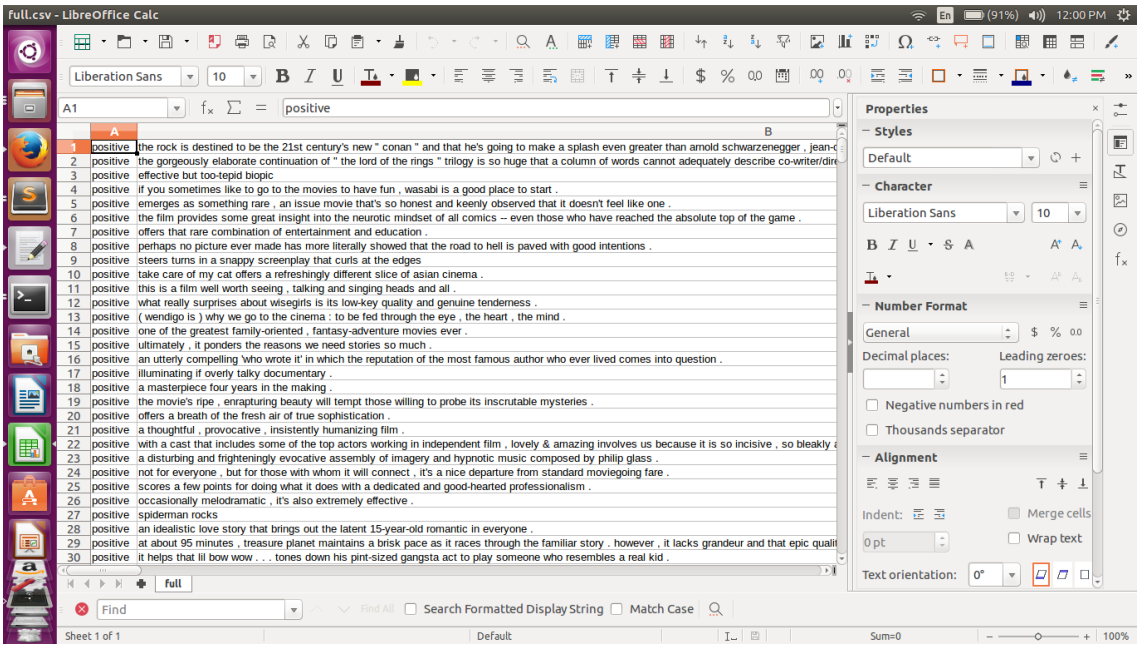

Figure 8.4: Screenshot Of Dataset Used For Training

Twenty thousand tweets are trained.

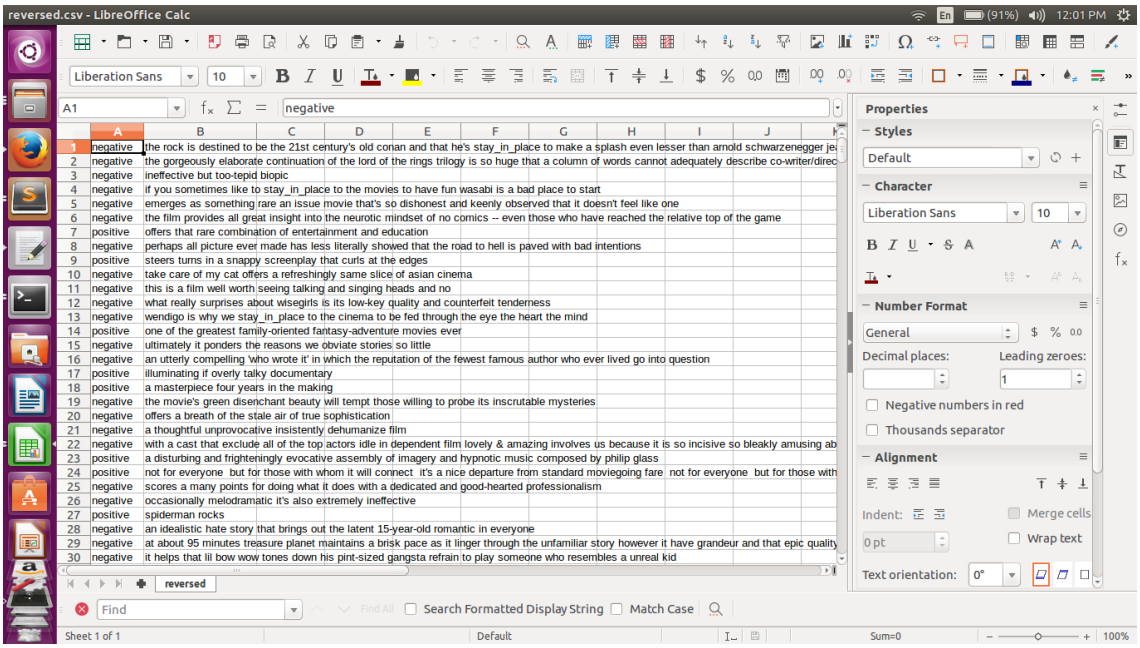

Figure 8.5: Screenshot Of Reversed Dataset Used For Training

The original dataset is reversed and twenty thousand tweets are taken for training.

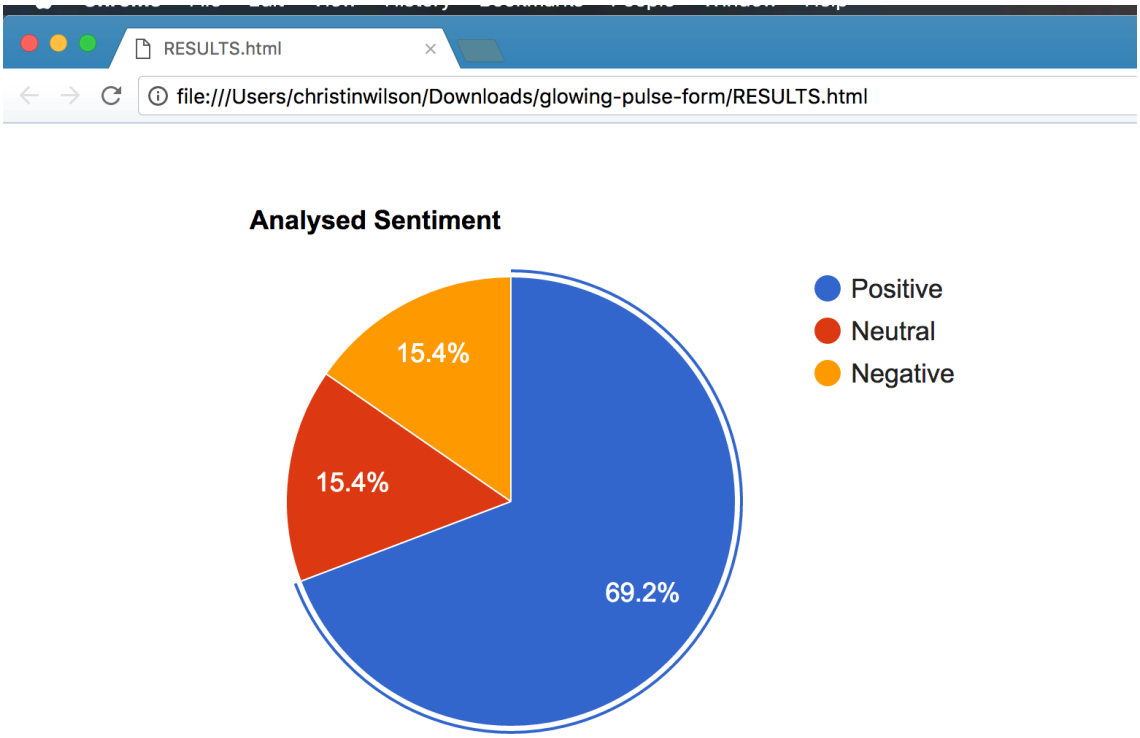

Figure 8.6: Screenshot Of Visualized Result

The classified tweets are then visualized in the form of a pie chart.

Chapter 9

Conclusion

The current scenario has a rapid demand of online data analysis. Sentiment analysis and Opinion mining involves the study of opinions and its related concepts such as sentiments, evaluations, attitudes and emotions. Twitter is an online news and social networking service where users post and interact with messages, "tweets," restricted to 140 characters. Registered users can post tweets, but those who are unregistered can only read them. Twitter has 140 million users and sees 340 million tweets per day. The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socio economic phenomena like stock exchange.

This project analysis the sentiment of a query given by the user. It searches twitter for related tweets and then analysis the tweets. They are classified as positive , negative or neutral. The current analysis methods such as bag of words causes the polarity shift problem. The polarity shift problem arises when Linguistic phenomenon in which the polarity of sentiment can be reversed (i.e., positive to negative or vice versa) by some special linguistic structures. For example, "I like this book" and "I don't like this book" would be classified as similar by the BOW classifier.

We use dual training of by using the reversed sentiment of the original data set to reduce the polarity shift problem. We obtain a sentiment analyser that gives us the opinion of the general public on a user query.

Chapter 10

Future Scope

1. Can improve efficiency of classifier by using a more efficient dataset.
2. Can use a combination of classifiers such as naive baise or maximum enthropy to improve efficiency.
3. Provision for collecting tweets of multiple languages other than english.
4. A provision to increase the limit of number of tweets that can be accessed from twitter.
5. Provision to collect tweets from a particular user (such as twitter account of a company) alone.

References

[1] Dual Sentiment Analysis: Considering Two Sides of One Review
by Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi and Tao Li, IEEE Transactions on Knowledge and Data Engineering

[2] Sentiment Analysis in Twitter using Machine Learning Techniques
by Neethu M S,Rajasree R

[3] Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis
by Geetika Gautam, Divakar yadav

[4] Thumbs up?: Sentiment classification using machine learning techniques
by B. Pang, L. Lee, and S. Vaithyanathan (2002), in Proc. Conf. Empirical Methods Natural Language Process.

[5] Learning to Shift the Polarity of Words for Sentiment Classification
by Daisuke Ikeda, Hiroya Takamuraz, Lev-Arie Ratinov and Manabu Okumura.

[6] Sentiment classification and polarity shifting

[7] Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

[8] Sentimental Data Analysis on Social Media
Priyanka. B and J.T. Thirukrishna, - International Journal for Scientific Research & Development Vol. 3, Issue 09, 2015

[9] Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews
by Kushal Dave, Steve Lawrence and David M. Pennock.

[10] Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance, by Zhen Hai, Kuiyu Chang, Jung-Jae Kim and Christopher C. Yang.