ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/273447919

Visualizing in Twitter Sentiment To Measure Consumer Insight

Thesis · September 2013

DOI: 10.13140/RG.2.1.1809.1044

CITATION
1

READS
2,530

1 author:

Mustafa Elkhunni
Bangor University

1 PUBLICATION **1** CITATION

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Visualising in Enviromental Data of Libyan North West View project

All content following this page was uploaded by Mustafa Elkhunni on 12 March 2015.

The user has requested enhancement of the downloaded file.

Visualizing in Twitter Sentiment to Measure Consumer Insight

Mustafa Bashir B El-Khunni

Department of Computer Science

Bangor University

A thesis submitted for the degree of

Master of Science

2012-2013

Abstract

Sentiment analysis study on Twitter data has recently spread widely. Due to the useful contribution of the sentiment analysis approach in review measurement, I introduce this project for sentiment classification mechanism in real time, based on Twitter data. This system provides search queries over Twitter API and database. based on user requests. Tweets are fundamentally classified into positive, negative or neutral bases, but sentiment visualization graphs show only positive and negative results. The impact of these sentiments was measured using a polarity computing algorithm. This thesis proposes techniques that improve the sentiment analysis of tweets via unsupervised classification techniques. It also includes theoretical information about supervised approaches which has been used in evaluating stage. Machinelearning techniques and tools are briefly illustrated. Lexicon and treasure in addition to tokenization were described as they have been used to build the proposed classifier to measure consumer insight.

Acknowledgement

Foremost, I would like to dedicate my thanks to my supervisors sincerely for thesupport they deliver to me during my research. In particular,my thanks go to my first supervisor Jonathan and my second supervisor Dr. *William J.Teahan* for the productive sessions we had. Those sessions resulted in new insights, which supported me in applying my research. Within the writing of my thesis, I appreciated the scientific feedback of *Jonathan C. Roberts*, especially when it was too difficult to start writing on paper, as I'm ansecond-language English speaker. I am very grateful that Jonathon spent a lot of time in assisting my project preparation.I also want to thank my second supervisor Dr. *William J Teahan*who was my course supervisor at the same time. The weekly meetings I had with him helped me to get on track, especially at the start. The feedback of both Bill and Jonathan was very insightful and invaluable. The time I have spent studyingat Bangor University has let me achieve a lot, and get involved with the practical applications in several approaches. Some of the interesting and useful modules were AI and Agents, Information Visualization and Internet Systems. Furthermore, thanks to my parents, my wife, my friends and all my brothers;all of these people were very cooperative and supportive throughout the entire period of my study.They really made me feel confident during my project, and I really love them all.

Author

Contents

List of Figures..... 8

List of Tables..... 10

Chapter 1. Introduction..... 11

1.1 Problem Description..... 12

1.1.1 Research Questions..... 13

1.2 Motivation..... 14

1.3 Hypothesis..... 16

1.4 Objects and Aims..... 16

1.5 Methodology and Structure of Thesis..... 17

Chapter 2.Background and Literature Review..... 20

2.1 Basic Concepts..... 20

2.1.1 Micro-blogging Twitter..... 20

2.1.2 Twitter Slang..... 21

2.1.3 Twitter Message Statistics..... 22

2.1.4 Twitter API..... 22

2.2 Consumer Insight..... 23

2.3 Text Mining and Natural Language Processing..... 24

2.4 Sentiment Analysis..... 24

2.4.1 Sentiment Analysis as a Scientific Problem..... 26

2.4.2 Data Representation..... 26

2.4.3 Sentence Extraction. 27

2.4.3.1 Supervised Approach. 28

a Naive Bayes Classifier. 28

b Max Entropy Classifier..... 30

c Support Vector Machine Classifier..... 30

2.4.3.2 Unsupervised Approach..... 32

a Data Preprocessing..... 32

b Feature Extraction..... 33

c Tokenization..... 34

2.5 Twitter Sentiment Applications. 36

2.6 Data Mining Tools……. 37

2.6.1 WEKA..... 38

2.6.1.1 WEKA Main Features……..... 39

2.6.1.2 WEKA GUI……. 39

2.6.1.3 WEKA API and Tools……..... 39

2.6.2 NLTK..... 40

Chapter 3.Risk, Legal, Social, Ethical and Professional Issues..... 41

Chapter 4. Design..... 43

4.1 Sentiment Classification Module..... 43

4.1.1 Twitter API Access..... 44

4.1.2 Data Processing Module..... 44

4.1.2.1 Opinion Extracting Stage (OES)..... 44

4.1.3 Sentiment Classification Module..... 44

4.1.3.1 Polarity Assessment Stage (PAS)..... 45

4.1.3.2 Polarity Classification..... 45

4.1.4 Feature Extraction Stage (FES)..... 46

4.2 Sentiment Visualization Module..... 47

4.2.1 Design..... 47

4.2.1.1 Methodology..... 47

4.2.1.2 Five Design Sheet (FdS)..... 48

4.3 Project Requirement Specification..... 53

4.3.1 Scope of the Project..... 53

4.3.2 SystemRequirements..... 54

Chapter 5. Implementation and Experiments..... 55

5.1 Developing an ExistingModel..... 55

5.2 Project Platform Preparation..... 56

5.2.1 Register for an Application with Twitter..... 56

5.2.2 Global Variables and Authentication..... 57

5.2.3 Workstation or PC Specification..... 57

5.3 Project Flowchart..... 57

5.4 Twitter Connector..... 57

5.5 PreprocessingTechniques..... 60

5.6 Sentiment Extraction and Polarity Scores..... 64

5.6.1 Lexicon..... 64

5.6.2 Negations and Modifiers..... 66

5.6.3 Tweets Ranking..... 67

5.7 Data Storage..... 68

5.8 Consumer Insight Visualization and User Interface..... 68

5.8.1 Web Application Configuration. 69

5.8.2 Developing Visualization Tools..... 71

5.9 Implementation Challenges and Technical Issues..... 73

Chapter 6. Analysis and Results...... 75

6.1 Evaluation..... 75

6.1.1 Case Study..... 75

6.2 Measuring Algorithm Effectiveness..... 77

6.2.1 Results..... 81

Chapter 7. Discussion..... 94

Chapter 8. Conclusion..... 96

Chapter 9. References..... 98

List of Figures

Figure 2.1 Screenshot of TwitterMessage Interface(Twitter.com)..... 21

Figure 2.2 Separable Dataset in Linear SVM (de Groot, 2012)..... 32

Figure 2.3 Screenshot of Twitter Sentiment User Interface (Sentiment140)..... 35

Figure 2.4 Screenshot ofTwitrratr(Twitrratr website)..... 36

Figure 2.5 WEKA Platform Graphical User Interface (Weka Ver.3 GUI)..... 37

Figure 4.1 Main Design Module Diagram..... 43

Figure 4.2 Design Schematic..... 46

Figure 4.3 Scanned FdS, Sheet-1..... 48

Figure 4.4 Scanned FdS, Sheet-2..... 49

Figure 4.5 Scanned FdS, Sheet-3..... 50

Figure 4.6 Scanned FdS,Sheet-4..... 51

Figure 4.7 Scanned FdS, Sheet-5..... 52

Figure 5.1 Registration Confirmation on Twitter API Application..... 56

Figure 5.2 Project Flowchart..... 59

Figure 5.3 Polarity Scale..... 65

Figure 5.4 MYSQL Database Tables..... 68

Figure 5.5 Graphical User Interface of the System..... 72

Figure 5.6 Sentiment Visualization from User Keyword Search Action..... 73

Figure 5.7 Used Dependencies..... 74

Figure 6.1 Comparative Column Graph for Apple and Sony..... 75

Figure 6.2 Sentiment Time Line of Both Apple and Sony..... 76

Figure 6.3 Evaluation Concept Diagram..... 77

List of Tables

Table 2.1 Instances..... 27

Table 2.2 Word Scoring..... 29

Table 2.3 Extracted Features..... 34

Table 5.1 Lexicon Database..... 65

Table 6.3 Random Tweets Collected with Sentiment..... 76

Table 6.4 Data after Preprocessing..... 81

Table 6.5 Predictive Instance Generated by Naive Bayes…..... 85

Table 6.6 Test data labeled by both system and human 90

Table 6.6 Test data labeled by both system and human..... 92

Chapter1.Introduction

A few years after Web 2.0 boomed, social networks and micro-blogs services are widespread, and in fact, they are today one of the most common communication facilities(Han et al., 2013;(Saif et al., 2012).

Different types of online social network platforms range from micro-blogs,blogs,such as,(Twitter) andsocial networks (e.g. Facebook) to content communities such as YouTube, etc.Such networks 'provide a powerful reflection of the structure and dynamics of the society and the interaction of the Internet generation with both technology and other people'(European Commission, 2010). This revolution in the way we communicate has led to major changes in the traditional concept of communication and created a new era of communication technology. Today, people are keen to use the web not only to browse web pages or doonline shopping but also to interact with the public to see what is going on. Furthermore, emerged multimedia applications– for instance, Voice over IP (VOIP),multimedia messaging(MSM) andvideo conferencing – not only encourage people to log into the web but have also changed the way we communicate.

The popularity of Twitter as a micro-blogging social platform has increased more and more in the last couple of years. With more than 350 million tweets a day exchanged by about 550 million users,Twitter has become the most used micro-blogging social medium on the web.Messages (tweets) that represent the way people share interests interactively are an important source of information.(de Groot, 2012). From a data analysis point of view, through these platforms, in particular Twitter and Facebook, tons of information that expresses people's feelings, opinions, feedback and thinking are published and exchanged among users every day(Bifet, 2011).Consumer feedback about brands, products and services has been difficult to track in the past. Today it can be tracked by collecting what is written about them online, e.g. in micro-blogs such as Twitter. Consumer insight sentiment mining

based on social media sources for language that may impact reputation in both positive or negative ways can be an effective tool for an organization's strategy, plans and marketing enhancement schemes(Damaschin et al.,2012).

Visualization in the sentiment of social media is a technique that focuses on knowledge representation generated by a data mining process;an example is making different graphs that represent the variation between brands or companies with multiple views and comparison among different categories . In this thesis, visualization is carried out to represent sentiments of consumer insight,based on a module that will analyse Twitter sentiment to measure consumer insight and represent it(WeihongQianet al,2010).

1.1. Problem Description

This chapter describesthe problem and associatedquestions of this research.

A huge amount ofinformation can be extracted from Twitter messages content through analysis techniques, The obtained information can be helpful to improve assessment of customer satisfaction. As well as, it can assist marketers both to assess levels of competitionand to improve business strategies based on customers' feedback. This approach is known as *sentiment analysis,*and it studies opinions on the web (Liu, 2012). All emails, social networks dialogues, blogs and social media on the web are called *internal data*,while any data sources off of the web are called*external data.*

The main point of investigation for this thesisis how customer feedback visualization improves marketer evaluation based on Twitter data.

1.1.1 Research Questions

Main question: How can visualization ofTwitter sentiment contributes to improve opinion classification?

The data analysis research field involves algorithms for the Twitter Sentiment analysis mechanism that has been developed. Considered research areas are: data mining, natural language processing and sentiment analysis. In data analysis different types of algorithms have been developed. Those algorithms can be categorized into supervised and unsupervised learning algorithms. The difference between these two models is that in supervised learning a model is created on a training set, and in unsupervised learning a predefined model is supposed. The question hereis which algorithm will achieve the best accuracy.

Sub question 1: Do features found in Twitter messages enhance the accuracy of classification?

Algorithms in data analytics are developed to perform specific tasks.This is also based on the domain to which they are meant to be applied; therefore, different algorithms do different tasks. There is noperfect algorithm able to perform all analysis functions; at the same time, each specific task has its own superior algorithm. By minimizing the scope of the algorithm, results can be more accurateand vice versa. Narrowing and expanding ofscope leads to comparative and significant levels of accuracy. For instance, variation of accuracy will be gradually improved if restriction to a specific topic is applied. Additional information about the userscan add more figures to the analysis results. For example, the majority of companies classify customers based on a set of information like location, gender, age, education etc. Consequently, this information is used to evaluate customer needs, customer satisfaction and customer interests. Furthermore, this information allows relational categorization, and, based on this additional information, categorization sentiment analysiscan be more effective and valuable for future business plans of companies.

Sub question2: To what degree are the special features of Twitter's linguistics,such as retweets, emoticons and hash-tags, number of friends etc.,contributingto better classification?

In the micro blogging service of Twitter, user creates status messages called 'tweets' (Barré&Schroeter,2011). These tweets usuallyexpressopinions about different topics. Hash- tags and emoticons have to be classified to extract user opinions and the degree of customer satisfaction. In this case we divided the user mode into three statuses: (happy case or unhappycase; positive or negative; satisfied or dissatisfied).

Sub question 3: What is the suitable visualization that explains the extractedcustomer opinions?

1.2. Motivation

A study stated that 81% of internet users of the US (60% of the population) have used the internet 60% of the time to search for a product they are interested in purchasing, as of 2007 (Horrigan, 2008).

In recent years, social networks became one of the most used applications on the web. Through social networks, information is generatedconstantly. Other services that have also emerged withWeb2.0, such as E-commerce platforms, search engines etc.;these web services' integrityenables much of the population to share their interests, opinions, influence and feedback. One example of this activity is online marketing operations, where customers tend to exchange their experienced knowledgein textual feedback via Twitter. People's feedback exchanges can occureither before or after purchasing items. From this perspective, there is a need for an automatic measurement tool based on Twitter message content to improve decision making processes, e.g.according to the latest poll, 86% ofpeople are not interested in the Microsoft Office 2012package. In this case, Microsoft company has

to produce another product or find another way to overcome this problem. Similarly, data analysis in micro-blogging content of Twitter will allow companies to make a decision through an automatic-electronic mechanism (Grosse et al., 2012). Data collected from Twitter is rich enough to represent different types of audiences such asregular users, companies, political foundations, VIP characters, celebrities, festivals, events etc. In addition to that,Twitter data structure comes in the format ofa structured database, which makes it easy to detectTwitter message attributes like author of tweet, followers, retweets and so forth.

Based on Twitter message content,our task is to extract and classify Twitter sentiment. Some researchers have called thisTwitter dialog anelectronic word of mouth 'eWOM'. Such analysisprovidesa useful tool for onlinepolling that can be used for company ranking, product reviews or as anautomated comparative mechanism for brand reputation.

This field of research has been going on intensively.During the elections in theUSA in 2008when a poll was conducted based on Twitter content, the measurement matched the real results of the USA elections, when Obama won the election. Perhaps in the future, there will be no need for high cost and time-consuming traditional surveys anymore.By enhancing the accuracyof the analytic tools, smart and automatic evaluation based on social networks can be the electronic intelligence of the next era (Horn, 2010). Jansen et al.(2009) claimed that one fifth of Tweets mention a certain brand, of which 20% included a sentiment (Jansen et al., 2009). By using sentiment analysis, market researchers will have an automated tool to observe how a product is evaluated,and consumers will take the advantage of this analytic tool when they tend to choose a popular brand. Visualization in Twitter Sentiment can help make senseof brand preference information in customer mode. In this project, I'm also developing visualization tools to explain changesinselected brand trends and polarity over time.

1.3. Hypotheses

If there is a link between Twitter message content and market classification, this means that online automated intelligence based on social networks will play a primary role of improving e-commerce and e-businessmarketing strategies.

If the accuracyof the developed classifieris successful, this can be producedon the Twitter platform as a component of Twitter to show user behaviour based on his/her sentiment.

If visualization of sentiment allows participants to locate both influential users and tweets, this will improve understanding of how and why events on Twitter stream are taking place.

1.4. Objective and Aims

Generally, the sentiment analysis approach helps marketers, business foundations, company and service providersto understand the opinions of their customers: how do customers or/and product consumers think about companies in terms ofsatisfaction, influence and quality issues? Wright stated thatfor various businesses, online feedbackhas become a kind of virtual currency that can decide the future of any product in the marketplace (Gebremeskel, 2009). As a matter offact, this kind of research can also be useful to psychologists and sociologists to understand and analyse online communities and how do they operate. Similarly, it can be used in political elections to predict the comingpresident, or for evaluation of eventssuch as football games, celebrities, moviesetc. For example,when a person or company wants to connect internet services, they will look for the best service provider in the market(Becerra,2012).

This thesis aims to:

1. Create and evaluate a sentiment classifier algorithm, which will be able to classify automatically Tweets. The developedclassifier does nothave to accept any external data, but it should rely onlyon Tweets through a proposed algorithm. The requirementfor this classifier is to have abackend where the training and test data can be handledand assessed.

2. The overall target is to build brand awareness from a highly cleaned and qualifieddataset, which will be used to classify opinions of Twitterusers at high levels ofconsistency. Once the classifier is built, it has to be evaluated. The commitmentis to achieve superior results.

3. Create a visualization of theresulting sentiment in a manner that gives clearer details than textual expressions.

1.5. Methodology andStructure of Thesis

This section illustrates and explains the methodology deployed in this thesis and the research methods found in the reading that informed my choice of methods.

1. Background and Literature Review

The background withliterature review are beingconducted on:

- ✓ The concepts ofmicro-blogs Twitter
- ✓ The characteristics of Twitter posts with an overview of Twitter API
- ✓ Consumer insight in social networks
- ✓ Text mining and natural language processing
- ✓ Concepts of sentiment analysisas a scientific problem
- ✓ The existing techniques used in data mining
- ✓ Some application done on TwitterSentiments
- ✓ Some developed tools for data mining and natural language processing

2. Design the sentiment analyser module, which will represent consumer insight from Twitter posts based on search keyword.

3. Implement the module.

4. Results and test.

This research was conducted in order to determine whether Twitter Sentiment plays a significant role in identifying and visualizing consumer feedback. The advantage and disadvantage in addition to the accuracy of the techniques used is also part of the objectives. In order to achieve the research target, data from Twitter will be collected based on company, product or brand keywords. Data gathered based on these topics will be processed to extract the required information. Along with this, natural language processing techniques will be usedto analyse the data. Finally, resulting information will be visualized.

The methodologyfollowed in this thesis throughoutis the flowchart,while the main goal isto convert electronic textual data into quantified numeric data. The researcher also opted to utilise the qualitative approach in this study because of its significant advantages. The use of qualitative and well-prepared data principles is advantageous as they will be followed to clean and prepare input data to meet the required accuracy. As soon as requirements are available from Twitter data to cleaned and labelled the data, the researcher will suggest s design for the entire module.Subsequently, data has to be input through the developed algorithm to extract the sentiments from tweets. Next, as a second stage, achieved results are going to be visualized,while evaluation of the solution has to be applied to test the accuracy and the functionality of the entire module.

On handthesis is structured as this: In Chapter 2 the background, literature review and basic concepts are explained, including the characteristics of tweeted messages and how these are importantin analysing natural language messages. In addition,the main topic that is discussed is sentiment analysis techniques. Chapter

3 provides the risk, legal, social, ethical and professional issues. Subsequently, Chapter 4 discusses the designrelevant to the structure of the proposed module. Whilein chapter 5 implementationand experiments are detailed, Chapter 6 includes analysis and results.Chapter 7 is the discussion and Chapter 8 is the conclusion.Finally,references were listed in the last Chapter of this project.

Chapter 2: Background and Literature Review

2.1. Basic Concepts

In this section, basic and relevant information to the thesis will be illustrated.

2.1.1. Micro-blogging Twitter

Twitter is a commonly known and widely used social network website. Normally, the Twitter message interface asks one question of the current subscriber: 'What's happening?' as shown inFigure 2.1.The response is limited to a maximumof 140 characters, which can be sent via web, Smartphone, Android, SMS, E-mail or third party application(Twitter.com,2013). Furthermore, the sent message of the userin Twitteris called a *tweet*. Users who are connected to the owner of this tweetis called a *follower*. In addition to that, the action that allows Twitter users to receive tweets from other Twitter users is called*follow*. Similarly, the user who sent a tweet called the *Tweet holder*. There isa feature in Twitter that enables followersto get new updates from a Tweet holder (Horn, 2010).Twitter as a micro-blogging service has some specific sentiment analysis characteristics. Limitation of Twitter posts(tweets) to only 140 charactersin length makes it very short posts compared to other social network posts. Not only does the limitationof post length make sentiment analysis different,but the nature of the language used in these limited message of Twitter is different. There is a high level of slang use and emoticonswith a high range of misspelling and errors, which increases the difficulty of analysis process. One more feature found in Twitter message is that users also usually tend to use hyperlinks and URLs to refer for external data. Therefore, the majority ofresearchers suggested using a machine learning approach to analyse sentiments on Twitter. However, there are many other sentiment analysis techniques will be listed in next

chapter. The reason why machine learning techniques are a suitable approach forTwitter sentiment analysis stems from two points:

1. The availability ofa large amount of Twitter data needed for training,

2. The ability to retrieveuser-labelled data that contains emoticons and is very helpful as test data (Pang &Lee, 2008).

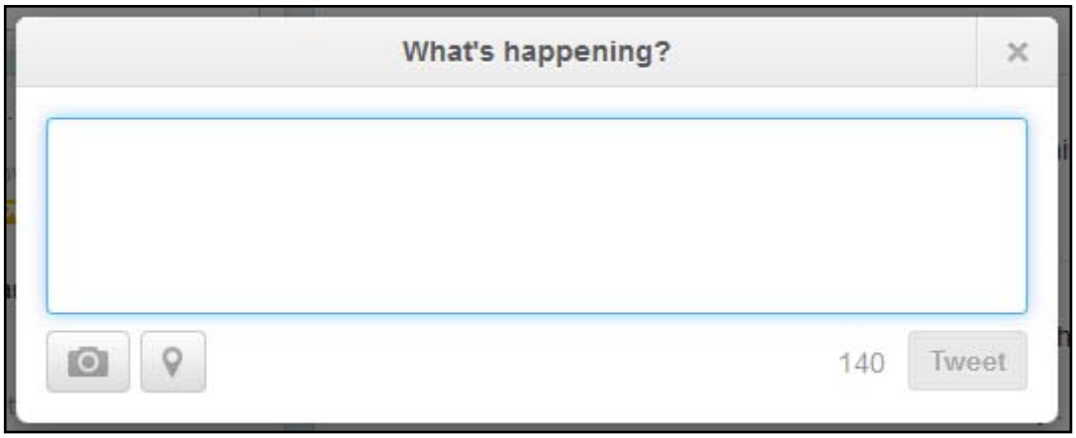


Figure 2.1 Screenshot ofTwitter Message Interface(Twitter.com)

2.1.2 Twitter Slang

The short length of 140 charactersfor Twitter messages makes writing a comprehensive conversation a challenging task. While there is no way to expand this limit, message holders tend to overcome this issue by using abbreviations, emoticons, slang and tiny URLs. Further, using the tagging feature, which is a hash sign '#', the user will be allowed to tag terms andwords and then it can be used as search keywords in Twitter searches to list all the topics including these tagged words. For example, in a Twitter message like 'I prefer #Bangor university', Bangor is the tagged word(Barré&Schroeter,2011). Twitter utilises many online server companies to support user services,such as TinyURL.com,which is the default service provider in Twitter. TinyURL allows Twitter usersabbreviate long URLs. Morerecently, Twitter has begun using bitly as a default service that provides a

package of services like tracking clicks volumes, link organizations and link sharing. These type of services makes bitly one of the big names in this field (Bitly, n.d.).

2.1.3 Twitter Message Statistics

a) A Twitter message is a short message.

b) A Twitter message is an unstructured message.

c) A Twitter message usually includesan opinion.

d) A Twitter message includes a high level of slang.

e) A Twitter message is likely to contain emoticons and URLs.

f) Retweets in Twitter always come after a hash sign.

g) A Twitter message may include spam.

2.1.4 Twitter API

Twitter as a micro blogging platform provides access to anapplication programming interface (API) to get access to Twitter data. Using HTTP protocol, the user is allowed to review all available data in the site such as search operations and subscriber status. Further, Twitter API allows researchers or interested users to get access to users' time lines and to retrieve data according to Twitter policies. API access withinvolved service is also offered by many third parties. Further, Twitter has different types ofAPIs: REST API, Streaming API and Searches API. Each provides specific functions. REST API guarantees users access to Twitter Search API core data by providing specific functions to communicate with TwitterSearch. Streaming API allowsreal-time synchronized connection to all Tweeter data (Twitter.com, 2013).In other words, Streaming API access allows users to connect to a large amount of Twitter data on a real-time basis and without filtering. Searches API retrieves only recent tweets based on specific search keys(Twitter.com, 2013). There are two access methods available to Twitter API: authenticated access and

unauthenticated access allow users to get a Twitter API Stream connection and to retrieve tweet searches respectively. Twitter API documentation gives detailed information about Twitter usability, access functions, user operations and retrieval limits(Twitter.com, 2013).Twitter API also provides the ability to use programming languages with specific libraries such as Java objected-oriented, which includes twitter4j API as part of the Java library for implementing custom applications on Twitter(Twitter4j, 2013).

2.2 Consumer Insight

According toNielson Company(2009), consumerskeeptrust the opinions of other consumers (either their contacts or opinions of different consumersexchangedonline) more than commercialadvertising. Further, visitors who observeconsumer reviews about goods have a 6% higher in terms of order average value than customerswho don't know aboutreviews (Conversation Index Vol.1). Consumer reviews are remarkablymore influentialabout twelve(12)times morethan clarificationsthat marketersmention, based ona survey of internet users from USconducted bythe electronic video review website (EXPO)(Bazaar Voice, 2010).

Social media has become a modern versionof the traditional gossip concept. Therefore, via Twitter, consumers tend to exchange experiences about daily things like what they intend to purchase (mobile phone, hair dresser services and so forth). Subsequently, such generated conversation is very valuable to marketers as it holds useful information. It conveys consumers' feelings, thinking and perceptionstoward certain products or brands (Chamlertwatet al., 2012). This type of habitual consumer feedback known as 'consumer insight' (Stone et al.,2004). In this thesis, data from Twitter is retrieved and processed to measure consumer insight about brands, products, so forth.

2.3. Text Mining and Natural Language Processing

Anne Kao and his colleaguesdefined the process ofdiscovering and extracting knowledge from free and unprepared textual data as*text mining*(Kao & Poteet, 2007; Mendes &Antunes, 2009). Text mining is a part of data miningwhere knowledgeisextractedfrom large amounts of data;this is also known as knowledge mining from data(Han &Kamber,2006). Natural languageprocessing (NLP) is one of the tools to extract a fuller meaning representation from unstructured free text. To achieve this, it makes use ofvocabulary representations such as a collectionof positive and negative words, grammatical roles of the language, punctuations and stop words lists. Texts can be found in different forms and contexts: flat files such as PDF, TXT file and CSVformat. Social media represents one of the largest sources of human textual data, where people intend to exchange information about almost every part of their daily life. NLP is not an easy task due to the unstructured use of human languages; NLP has to analyseword ambiguities.For instance, synonyms, language metaphors, abbreviations andwritingmistakes.As well astaking in account the use of slang (Kao & Poteet, 2007).

2.4. Sentiment Analysis

Generally, *sentiment analysis*, *opinion mining*, *subjectivity analyses*, *emotion analysis,review mining*etc. are actuallydifferent titles for the same topic. However, in academia, sentiment analysis and opinion analysis are most frequently used (Liu,2012). All of these different titles employ mechanismsto study and analyse data in order to achieve a quantified measurement of knowledge. Opinion mining is made up of several different issues such as: specifying whether each tokenof text (, paragraph,sentence or section) is holding opinionor not; gettingthe opinion-holder; and specifyinghow positive or negative each opinion is. In the business field,it is also useful to analyseopinionsindividuallybased on to the natureof the business or transaction it indicates: product quality, ordering, or product defects(Funk et al., 2008).

An early start of sentiment analysis was by Wysoki(1998, as cited in Brown, 2012),who claimed that there is ahigh positive linkbetween volume of messages posted on message boards overnight and trading volume and returnsof stock the day after. Sentiment analysis study started to rise in the recent years (Brown, 2012). After the dotcom crash, researchers started to consider the huge amount ofpotential information shared on the web. Liuclaims that in particularthe term *sentiment analysis* first appeared in Nasukawa and Yi (2003, as cited in Liu, 2012), while the term *opinion mining*was first mentioned in Dave, Lawrence and Pennock(2003, as cited in Liu, 2012). The term *sentiment* as a study area refers to tracking and evaluating text to analyse and predict future strategies (Pang & Lee, 2008). In 2001, Das et al. published a paper on market analysis. In 2002, Pang et al. published a paper concerningempirical methods in natural language processing (Pang et al.,2002).

Since data represents the backbone of such natural language processing, objectives from which data is processed have to be predefined. Twitter provides numerous sources of fresh data that contains features expressed by communicators in a textual format. This conveys user mode. User textual reaction might becustomer feedback or feelings like anger or happiness. It also may indicate attitude, influence, opinion toward something etc. Semantic analysis is deployed to measure such characteristics. Apart from this, obtained information from sentiment analysisis being used in different sectors.Sentiment analysis provides new strategiesformarket intelligence. Sentiment analysis defines the text's nature, whether it is subjective or objective, and provides polarity measurement of the identified subjective text (Pang & Lee, 2008).In other words, sentiment analysis is a challenge of natural language processing when texts are processed to extract opinions (Liu, 2010). Generally, semantic analysis processing can be divided into two mainparts; classification of subjectivity andclassification ofsentiment.

2.4.1. Sentiment Analysis as a Scientific Problem

It is evidentthat, from scientific point of view, to solve a problem certain requirements should be met. Formulation of the problem has to recognize certain issues,namely, necessary definitions, core concepts and related issues. At the same time, objectives from which the problem has to be solved must also be determined. Similarly, for most every inputin any process there have to be expected outputs that also have to serve a required application (Liu, 2010). A number of research papers and published works have focused on twotechniques: some researchers employ machine learning algorithms to process a set of human categorized textual data such as brand or product reviews while others used unsupervised techniques(Charalampidou, 2012).There are three main levels of analysis in sentiment: First is the document level where the entire document expressed a positive or negative opinion; some researches called this a classification ofsubjectivity (Wiebe,2000). The second level is thesentence level,wherea sentence is supposed to represent one of three statuses (positive, negative or neutral), and the third level is theentity and aspect level, which is also known as the feature level; in this technique analysis is performed based on the assumption that there is a sentiment(positive or negative) behind each opinion and each of these opinions has a target(Liu,2012).

2.4.2 Data Representation

One of the essentials in data mining is how to represent the data. In the supervised machine learning approaches, data is usually represented in a tabular format andeach datum stored in these tables is known as an instance. Along with each instance, there is an associated fixed value of measurement. At the same time, each instance in these tablesis categorized under its relevant attribute or feature as in Table 2.1. Normally, machine learning applications require two sets of these instance tables: training instances and test instances. The training instancesset is

that data that is used as learning data to produce concepts. A test set is used to examine the accuracy to the output (Hall, 1999). Table 2.1 shows an example of data with instances.

Instance	Features				Recommendation
	Camera	Battery	Memory	Screen	
1	12 MB	8 hours	4 GB	3 inches	1
2	10 MB	6 hours	2 GB	3 inches	2
3	5 MB	4 hours	1 GB	2.5 inches	3

Table 2.1.Instances

2.4.3 Sentence Extraction

Sentence extracting is a text processingtechnique thathas as a primary role insentiment classification, where text is analysed to identify both positivity and negativity. Opinions in textual data come in different ways. There are direct opinions and indirect opinions,as well as comparative opinions. *I like Bangor University*, and *I agree with students who recommend Bangor University* and *Bangor University isbetterthan Zawia University*–these three simple opinions respectively represent the mentioned types. However, there are also complex opinions, where there is no easy way todetect the sentiment.

Pang at al.(2002) applied machine learning techniques to extract sentiment from a text by employinga support vector machine(SVM) to the entire document(Pang et al.,2002). Subsequently,Pang and Lee improved their classifierby enhancingthe subjectivity extractionprocess by deciding to exclude non-subjective opinions,or

objective sentences (Pang &Lee,2004). In this thesis, SVMs will be employed in the developed classifier.

The majority of researchers have divided text extraction techniques into two approaches: the supervised approach and the unsupervised approach.

2.4.3.1 Supervised Approach

This approach employs a machine learning mechanism. In this approach, an agent has to be taught by means of manuallylabelled and qualified data. As soon as the machine learning algorithm is taught by inputtinglabelledtraining data that representssentiment values relevant to determined domain, then the output will be stable (Pang&Lee,2008; Wikipedia, 2013).Two essential steps are always associated with a supervised learning approached: the first step is to teach the model using training data, and the second step is to test the model using unseen test data for accuracy measurement(Liu,2012).There are several machine learning algorithms used in sentiment analysis as sentiment classifiers.These include: the naïve Bayes classifier, the max entropy classifier and the support vector machine classifier.

a) Naive Bayes Classifier

The naive Bayesian classifier is a popular and frequently used algorithm for supervised learning. It provides a flexible way to interact with any number of classes and attributes. Fundamentally, it isbuilt based on probability theory. This classifier is used for many applications, such as assigning subject categories, topics, or genres,as well as forspam detection. Moreover,the naive Bayesian classifier is used for identification of authorship,age or gender identification techniques,language detection andOCR recognition. There are different levels ofthe naive Bayesian classifier. The simplestone relies on a simple representation of text document by using the bag of words concept(Manning et al., 2011).For instance, the next conversational example clarifies how the simple naiveclassifier operates.

Ilove this Applephone! It is fantastic but with tall frame. The screen is great and the Camera is amazing. ..loool, Iphone5 is the best. I would recommend it for everyone

i) Bag of words

I**love**thisApplephone! It is**fantastic**but with different frame. The screen is**great**and thecamera is **amazing,**Iphone5 is the**best**. I would**recommend**itfor everyone.

ii) Word representation

Word	Score or weight
love	2
fantastic	2
great	2
amazing	2
recommend	1

Table 2.2. Word scoring

The general sequence of the naive Bays classifier is described by Christopher Potts as follows:

i) Computethe estimation of probability P(c) of each class c ∈ C by means ofdividing the total of words in the documents in c by the total of words in the corpus.
ii) Assumethe probability distribution P(w | c) for all words w and classes c. This willbe foundby dividing the countof tokens of w in the documents in c by the total of words in c.
iii) To specifya score to ddocument for c class, next formula has to be calculated:

$$Score(d,c) =^{def} P(c) * \prod_{i=1}^{n} p(i|c)$$

To find the most likely class label, we shoulduse the c associatedthe maximum value, and to calculate the distribution of probability:

$$P(c|d) =^{def} \left(\frac{score\ (d,c)}{\sum_{c'\in C} score\ (d,c')}\right),$$(Potts,2011).

b) Max Entropy Classifier

From Christopher Pottstheclassification derived from the maximum entropy (MaxEnt) is slightly different fromthat innaive Bayes:there is one extra feature that, instead of letting toeach feature to ownits impactindependently,search-based optimization is employed by it to associateweights withthe features that maximize the likelihood of the training croups. The characteristicsthat can be defined for a classifier of naive Bayes are simply assignedto asetting of MaxEnt.By useof MaxEnt classifier, there isa capability to handle a different types of integer, Boolean, and valued features. The general guide for building a MaxEnt classifier, covering only wordlevel features, is as follows:

i. For each word w and class $c \in C$, define a joint feature $f(w, c) = N$ where N is the count of times that $w$foundin a document in class c. (Ncanbe Boolean, makingpresence via versa. absence.)
ii. By use ofiterative optimization, specifya weight to everyjoint feature so as to obtain themaximumlog-likelihood of the training croups.
iii. Class c *its* probabilityluncheda document d and weights λ is

$$(p|c, ë) = \frac{Exp \sum_i ë_i f_i(c,d)}{\sum_{c'\in C} Exp \sum_i ë_i f_i(c',d)},$$ (Potts,2011).

c) Support Vector Machine Classifier

Support vector machine (SVM) has been identified as a highly effective technique in text classification (Chamlertwatet al., 2012). SVM classifier is both probabilistic and large-margin opposed to naive Bayes and MaxEnt classifiers (Pang et al., 2002).

SVM is a machinelearning algorithm operates tocreate a function of decision based ona training croups. Thistraining croupsis labelleddata that include predictors associated with target values, and every predictor value is weighted with a value. When the SVM algorithm start toproducesa prediction of categorical value for target attribute, thismechanismis named a classification.

For example,suppose a class holds two categorical attributes, either positive or negative with no partial values. Ifthe attribute is formatted on an numerical basis, then the process is known as a regression. SVM is a machine learning algorithm found toclassifyaninput vector Xusing a function: $F(x) = (X, W) + b$.In other wordsSVMs are hyper plane classifiers and operate by specifying on which side of the hyper plane X lies. In the previous formula, the hyper plane is orthogonal to W and at a distance$b/||W||$from the origin(Guduru,2006). SVM makes the margin around the splittinghyper planein a maximum value. The decision function is completely identified by a training croups. This collectionofinstances or vectors is known asthe support vectors(Gunn,1998). There are two types on vectors, a linear and nonlinear vector machines. Suppose that we have two features only; this means there are two dimensions in spacewith a hyper plane. In this dual directionalspace the hyper plan is a line.

One example of this is the typical situation where the classes are separable in a linear manner. In such circumstances a line can exist that separates the two classes perfectly. However, not only does a single line divide the dataset exactly, but a completeset of lines do. From these existing lines, the best is selected as the 'separation line'. The operation of maximizing the distance to the closest points of these two classes in the croups of trainingfinds the most accurateline. When maximizingof this distance is done, this transformsto asimilarproblem known as minimization problem, which ismuch simpler to solve. Support vectors are thatdata locatedon the maximal margin lines.

Figure 2.2 explainan exampleof a linearly separable dataset (de Groot, 2012).

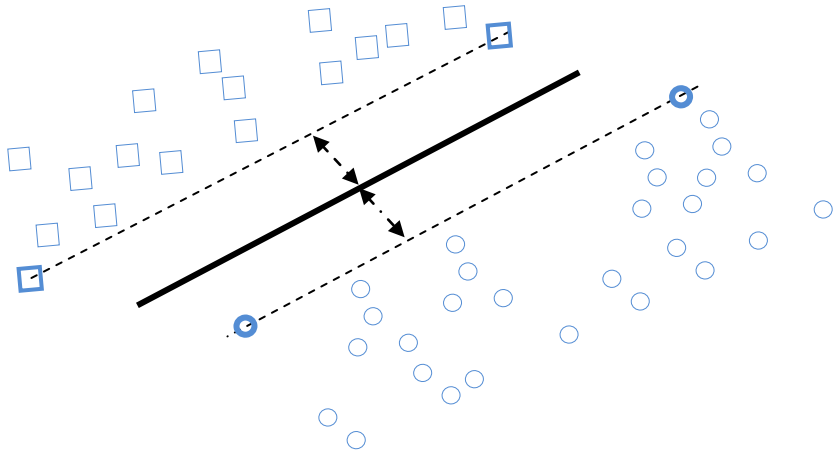


Figure 2.2. shapes that are located exactlyon margins in both sides represent vectors. Adapted from(de Groot, 2012).

2.4.3.2 Unsupervised Approach

The unsupervised approach is known as the lexicon based approach or clustering, and it is that role that operatesbased on features given by predetermined sentimental lexicon rank to measure the quality of text, whether it is positive or negative. In other words, this method filters the use of reference words, for most positive and most negative associations (Chamlertwatet al., 2012).Furthermore, data cleaning and labelling in the supervised approach is done by unsupervised approach techniques.

a) Data Preprocessing

Before startingdata processing in sentiment analysis, there are various preprocessing techniques that have to be applied on the data in order to removedata

noise and correct data inconsistencies. Preprocessing processin text miming meanstextreduction, textcleaning, texttransformation and textintegration (Han&Kamber,2006).

b) Feature Extraction

In the text preprocessing stage, convertinga text into a feature vector or other data representation is one of themost important steps of data-driven fields-to-text processing. Feature extraction is a technique where specific properties are extracted from the input data. These properties are known asfeatures and represent some characteristics and attributes of the data. In practical terms, the complete input data is too big to use in classification(Pang & Lee,2008). The features should be detected to identify the original data in multidimensional form, while these detected features contribute to reducing the data space to shrink redundancies and large dimensionality of the data.

Researchers have suggested and experimented with different features of sentiment analysis to understand which extraction method gives the best results. The following methods are used in text feature extraction: the *tf-idf* measure,*n-grams* and *part-of-speech* (POS tagging). The next example shows features included in a text.*'I love this kind of Applephones! It is fantastic but with tall frame. The screen is great and the camera is amazing, Iphone5 is the best. I would recommend it to everyone'*.

i) Features detected in the text:

Ilove this kind ofApplephones! It is fantastic but with tallframe. Thescreenis great and thecamerais amazing, Iphone5is the best. I would recommend it for everyone.

ii) Extracted features:

Feature	Brand
	Apple
Phones	
Frame	
Screen	
Camera	
	Iphone5

Table2.3. Extracted features

In the field of micro-blogs text mining, when a machine learning classifier is used with a wide range of features,it has been found that simple statistics of tweets, for example word count or readability scores, can help in sentiment analysis on Twitter text (Han et al., 2013).

c) Tokenization

From Wikipedia,tokenization is the technique ofbreaking an input of text up into its individual fragments such asphrases, words, symbols or other content that make sense and this segmentsknown as tokens. The collection of tokens can be input for further processing, for example parsing processor text mining techniques. Tokenization is usefulin linguistics, where it is a form of text fragmentation; in the area of computer science, where it play as apart of lexical process; and in text mining,it a mainpart of polarity calculation (Wikipedia,2013). In thisclassifier,tokenization is one of the preprocessing data techniques used,where

the sentenceor text is segmentedinto separatedwords to easily detect the noisy and unnecessary data such as punctuation and URLs. For example a text,'Please contact me on my Email must426@hotmail.com' is tokenized using Python white space tokenization as in Figure 2.3 to (http://text-processing.com/demo/tokonoze/).

Figure 2.3. Word Tokenization with Python NLTK Demo (NLTK 2.0.4, n.d.)

As far as text mining is concerned, there are two basic techniques that have to be applied first when extracting information.Firstly, the mechanism must be capableto detect structures in the text, such as sentences. It must detect words that are elements, and sets of words that identify some concept. It detects verbs that identify relationships among concepts, and it detects morphological variation of words. In this first step, the text is categorized into slices, dices, chunks and tags; these structured formats of text areready for extraction into a structured database. Fundamentally, this first step is a preparation of unstructured human text to be served to the machine. The second step is to extract information from the text. This can be as simple as recognizing terms like gene sequences, but usually includes recognizing namedentities such as names of authors, and fact discovery such as the relationships between terms or entities. The results are presented as facts or affirmations and collected in a database. The start-up for this step is often a template that describes the generalized form of the information to be extracted. An example would be a certain pattern of words that occur in a certain way. The goal is to design templates that will extract meaning in the form of facts from the text. A simple example from marketing is a list where two products, A and B, are found close to the phrase 'Brand name'(Clark,2013).

2.5 Twitter Sentiment Applications

One application of this sentiment technique is Twitter's Sentiment platform, which is developed based on machine learning algorithms (http://www.sentiment140.com/).

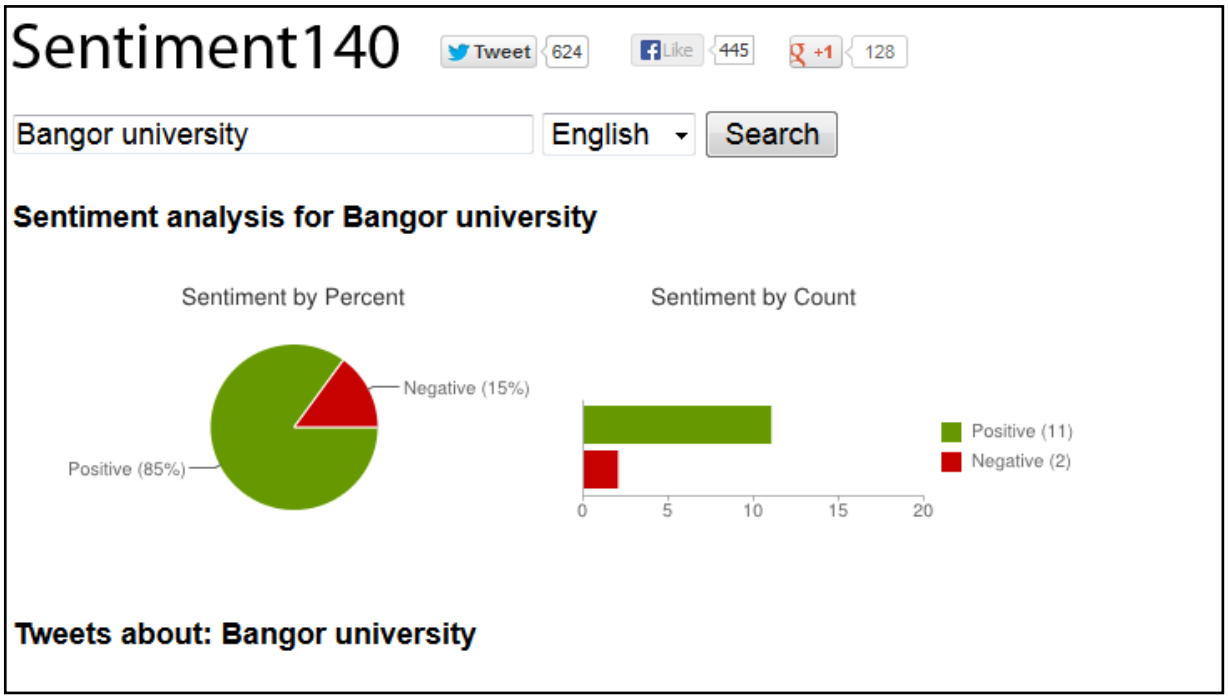


Figure 2.4Screenshot of Twitter sentiment user interface(http://www.sentiment140.com/)

Three undergraduate students,Alec Go,Richa Bhayani,and Lei Huang from Stanford University,developed thisapplication, and they called itSentiment140. It allows users to detect the sentiment of a product(Figure2.4), brand or topic on Twitter.The casesused in this project were the USA elections, planning and purchasing (http://help.sentiment140.com/ 2013).The otherapplication technique of Twitter sentiment analysisis to use a set of words with a weight attached for each word, either positively word oriented or negatively word oriented and to measure the words holding a neutral opinion. This kind of technique is unsupervised, and it classifies the provided data based on its included sentiment. An example is the Twitrratr platform, which is designed to identify brands, products and so forth. Sentiment in Twitrratris displayed as a negative, neutral or positive review (Twitrratr, 2013).See Figure 2.5.

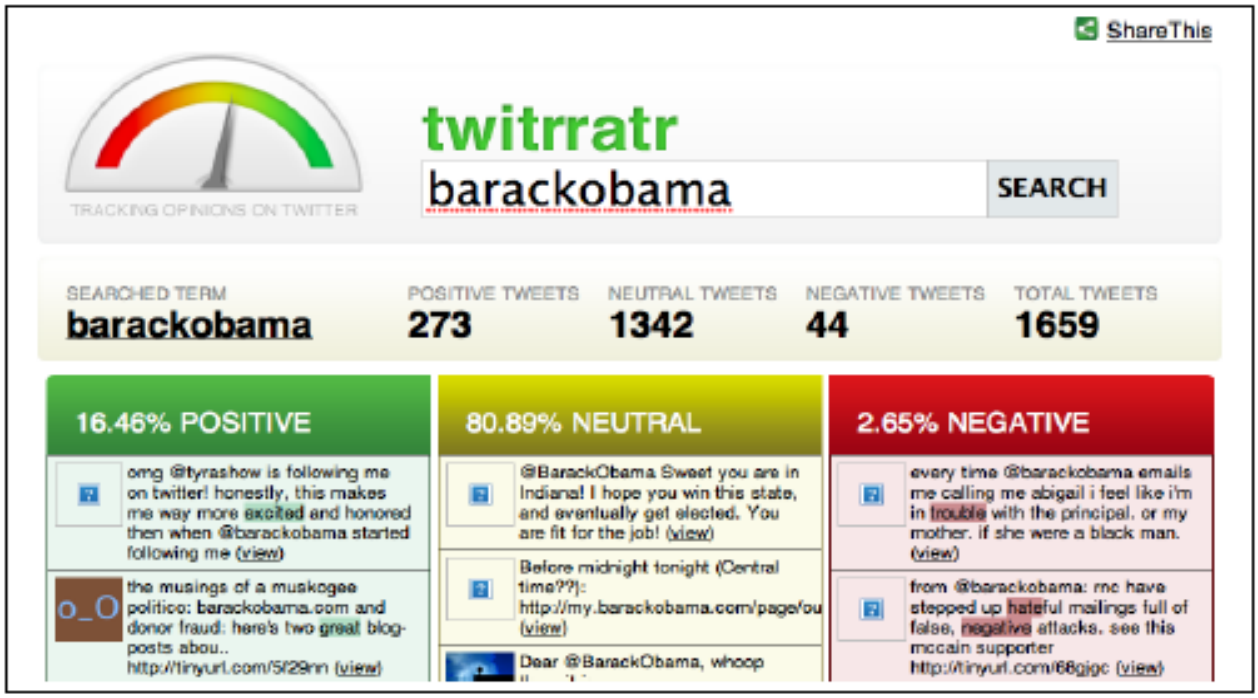


Figure 2.5. Screenshot ofTwitrratr (http://www.twtbase.com/twitrratr/).

2.6. Data Mining Tools

The application and development ofdata mining algorithms need effectivesoftware tools. Becausethe collectionof relatedtools is growing, the decision to choose the right tool hasbecome more difficult(Mikut&Reischl,2011).

It has been argued thatthe complexity of the text mining process is not only derived from the text or the data itself,but also derived from the intensive use of mathematics, probabilities and algebra for tackling the data mining process. Consequently, with the demand forboth data mining and informationdiscovery applications, helpful applications werecreatedto provide specific libraries, including data mining techniques such as machine learning algorithms implementation and data preprocessing. These applications facilitate data mining techniques in terms of coding and strengthen natural languages processing effectiveness. WEKA and NLTK are two different platforms developed to support human language processing and will be briefly overviewed.

2.6.1.WEKA

In 1993 researchers of a machine learning groupfrom University of Waikato developed an application named WEKA to support a wide range offunctions within the data mining process. The name of this application comes from a bird called the Wekathat lives in New Zealand.Furthermore, theWEKA platform was developed in Java,which makes Java one of the most-used programminglanguages in data mining. WEKA can be either used independentlyor it can be invokedfrom Java code.WEKA includesalgorithmsthat covers data mining mechanisms such aspre-processing techniques, different algorithms of classification, regression, clustering, association rules. Furthermore Wekasupports visualization of data mining tools. See WEKA GUI as in Figure 2.6(Bouckaeret al., 2010).

Figure 2.6.WEKA platform graphical user interface

2.6.1.1 WEKA Main Features

- ✓ Platform for data pre-processing tools, learning algorithms and evaluation
- ✓ Easy to use graphical user interface
- ✓ WEKA provides visualization tools
- ✓ WEKA is an environment for learning algorithms analysis (Witten & Frank,2005).

2.6.1.2 WEKA GUI

From WEKA GUI, a user is able to import data from a CVS file, ARFF file,C4.5, URL or SQL database. WEKA GUI supports command line interface,where a user is able to use WEKA commands starting from this command line prompt. In addition,WEKA GUI includes a feature of knowledge flow design, which provides visual components to build WEKA models in visual flowdiagrams. Additionally, there is an experimenter that makes it easy to compare and contrast the performance of the included learning schemes (Witten & Frank,2005).

2.6.1.3 WEKA API and Tools

Weka.jar is the full package of the WEKA application prepared to be used as a library in a Java platform such as NeatBeans 7.2.1 with JDK 7u13. As soon as WEKA jar is imported to the NetBeans libraries, the developer will be allowed to use WEKA commands in Java classes. In this thesis,WEKA API is used from Java Net Beans to build the supervised classifier **(**Witten &Frank,2005).

2.6.2 NLTK

Natural language toolkit(NLTK) is also one of the developed platforms for human language processing. It provides a smooth graphical user interface (GUI) to a variety of lexical and corpora resources such asWordNet, while providing a package of text processing language libraries for different text processes, namely, classification, praising, tokenization, stemming, tagging and sentiment reasoning. NLTK is developed in Python, which also makes Python one of the powerful programming languages in computational linguistics of human languages (NLTK, 2013).

Chapter 3. Risk, Legal, Social, Ethical and Professional Issues

This research thesis is based ona computer science approach in terms of software development and programming issues;this has no potential physical risk to the user or the reader based on HSE and risk assessment principles.

The content of this thesis is conducted according to UK rules and regulations. This thesis isprepared in accordance with criminallaws in termsof theft, piracy, espionage, fraud and sabotage. All material and content included in this thesis from adaptation tocitation content is prepared accordingto UK and international copyrights regulations. In addition, it has been writtenwith full consideration to the UKregulation of private, public and civil laws.

Data used in this module is collected from Twitter according to Twitter regulations and terms. The retrieved data contained in the tweets includes conversational content of ordinary users, politicians or companies and are pulled based on Twitter authentication.

This thesis is conducted from a scientific point of view in order to serve humanity and enhance the use of technology. This thesisis not conducted for commercial issues, personal profit or any type of business use. Additionally,it isnot focused on a particular set of attitudes, beliefs, values,or habits that a person or a group or nation displays. This thesis is available for everyone from everywhere, regardless ofethical differences across cultures. To conclude,data used in this study is not processed to lead to any kind of discrimination and does not violate any laws such as the Human Rights Act 1998 (Article1, UN 1948; United Nations, 2013).

Twitter data retrieval in this studyis performedaccording to Twitter regulations.There is noability to retrieve or to get any information fromthe Twitter database or any third party without permission from Twitter. Twitter does not

allow access to personal and security information, such as user passwords and/or any other personal credentials or attributes.

Data retrieved from Twitter is stored in a secured database using an open source MYSQLserver 6.0E.Data stored in this database is retrieved only for the specified task behind this project, which is sentiment analysis purposes. This data shall not be used in another subject incompatible with this project purpose.SQL server is only allows authorized users to connect to and usethe stored data.

Where the developed module in this research requires an internet connection to connect to Twitter API, Avast antivirus and internet security isinstalled on the PC where data is stored and the module is implemented in order to prevent any potential attacks of viruses or spam. In addition to that, a backup subroutine is implemented frequently to back up the data and store it in an external storagemedium (CD, DVD or cloud drive), in case any failure or damage occurred on the hardware.

Regarding used software licenses,all items retrieved from Twitter (tweets, creation dates, tweet, retweet, ids, number of friends etc.) are covered by Twitter's terms of service:**(**Twitter developers, 2013).

The module ofconsumer insight is redistributed and modified based on Apache license 2.0(Apache Software Foundation,2004), and all modifications done on this module are already documented. In addition, the reference to the original developers who developed the first version of this module are indicated according to Apache 0.2 license instructions. Furthermore, an email was already sent to one of the module developers informing him about the proposed activity on their model.

All contents of this thesis are aCopy rightto the author based on UK Copy rights regulations.

Chapter 4. Design

This chapter describes the design of the module of Twitter Sentiment to analyse consumer insight. The main module design is compounded from two sup-modules– sentiment classification module and sentiment visualization module –as a start to finish sequential process module. A general diagram of the main module is shown inFigure 4.1.

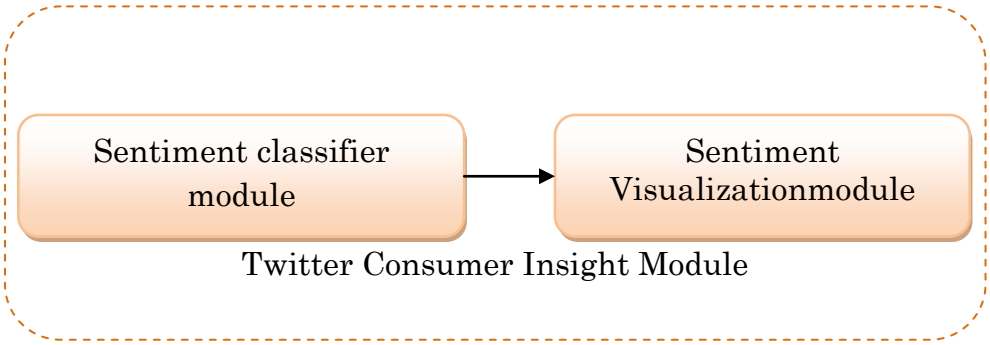


Figure 4.1 Main design module diagram

4.1.Sentiment Classification Module

To build a classifier based on machine learning, there are a general steps to be performed:

- ✓ Collect required data(Search API)
- ✓ Label Your Data
- ✓ Clean Your Data
- ✓ Upload Data
- ✓ Data store
- ✓ Train a Module

- ✓ Extract Opinions
- ✓ Polarity assessment of Extracted opinions
- ✓ Features Extraction
- ✓ Summary and results

4.1.1 Twitter API Access

In the tweets collecting stage (TCS), required messages of the study had to be collected from Twitter API. Application programming interface (API) allows programmatic access to Twitter tweets based on a specified query or keyword. Only a limited number of tweets can be collected at a time, which is 1%. Twitter posts can be retrieved either in XML orJSON format(Twitter developers,2013).

4.1.2 Data Processing Module

In this part, data is processed to extract opinions from input tweets.

4.1.2.1 Opinion ExtractingStage(OES)

To classify opinions, Tweeter messages has to be analysed to identify the subjectivity. The opinion extracting stage (OES) will classify tweets as either subjective or not subjective, which means detecting Tweets that hold opinions and tweets that do not. In this process only opinions will be considered,while tweets or sentences that are redundant or do not express opinions will be neglected. First, tweets have to be manually labelled and categorized on an opinion or non-opinion basis, and once the data is labelledwe need to make feature selection and build theclassification. Inthis model SVM is the technique used for opinion classification.

4.1.3 Sentiment Classification Module

4.1.3.1 Polarity Assessment Stage (PAS)

Twitter messages that remained after the OES stage are already filtered in terms of opinions. Consequently, in this polarity assessment stage (PAS) a sequence of processesis followed:

i) Opinions will be ranked or classified according to a scale: In particular, each word in the sentence of a tweeter's message has to treated separately by assigning a rank value to each word. For example, *just realized that download capacity in TalkTalk broadband is much higher than that in BT* is separated to 15 words.'Just''realized''that''download''capacity''in' 'TalkTalk' 'broadband''is''much' 'higher''than''that''in''BT'. This process known as word tokenization.

ii) Removal of stop-words:Exclude any stop words from the tokenized words; this allows an increase in meaningful terms and helps to avoid redundancy in dimensions of terms.

iii) URL removal: URL and external links are also removed because they have no polarity weight.

iv) Slang enhancing: In case of words like 'zooom' or 'gooooood'there is an automatic detection of letter repetition, so 'zooom' becomes'zoom' and 'gooooood' becomes'good'.

4.1.3.2 Polarity Classification

After performing all previous process polarity evaluation proceeds as follows:

1. Initialization oftotal polarity counter

2. Verifyif the token matchespositive or negative side: in case of positive polarity,sumword score to the overallof polarity,and in case ofnegative, then subtract the scoreof the word from the total of polarity.

3. Negations:For instance null, not, noetc.have to be treated as strong polarity words that have double value of polarity weight.

4.1.4- Feature Extraction Stage (FES)

To extract features from a Twitter message, predefined stored keywords based on the topic are retrieved to compare them with tokens;if a token matches brands, this token is classified as a Brand, and if a tokenmatches products or features, this will also be classified accordingly.See the schematicdesign in Figure 4.2.

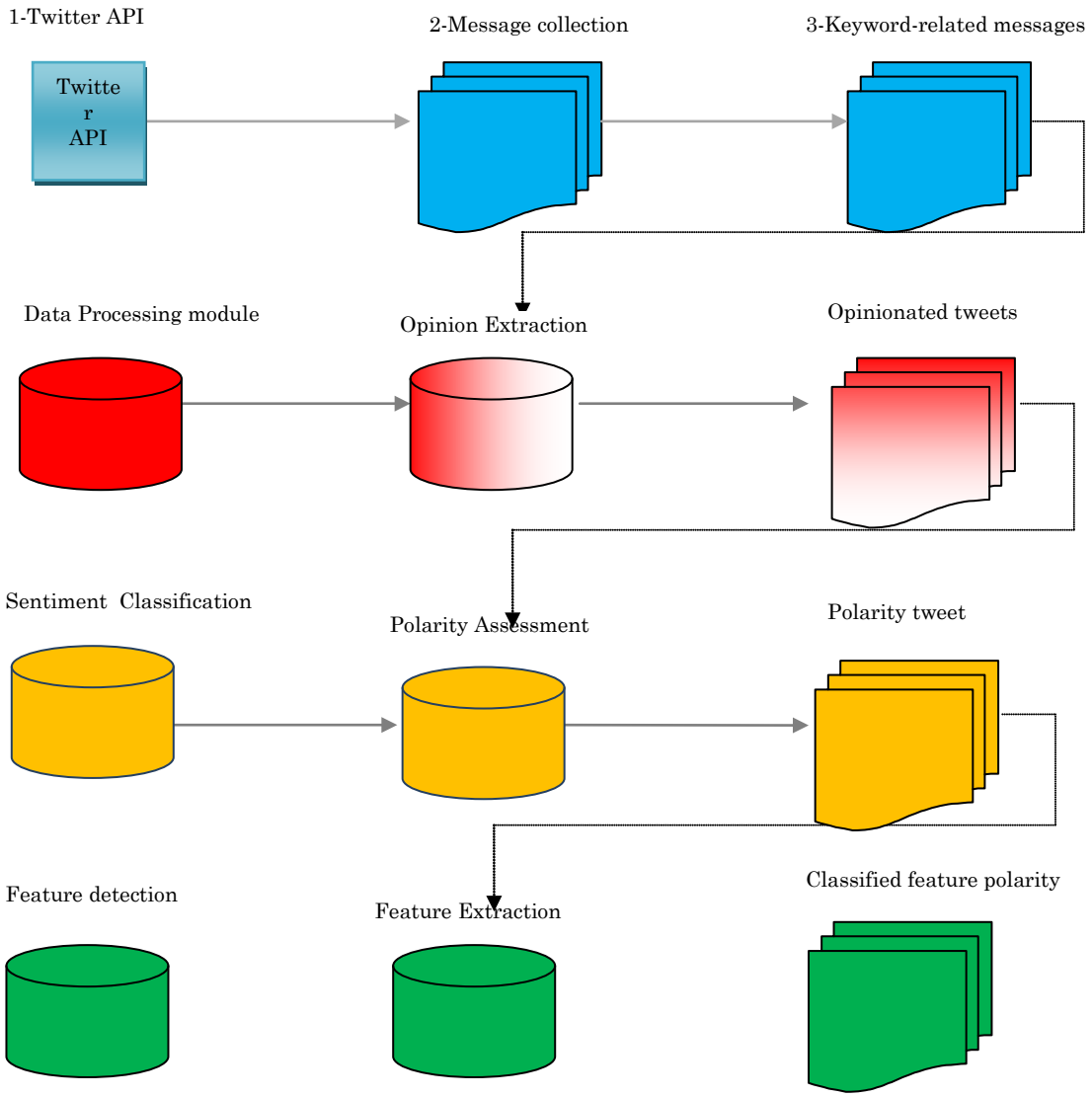


Figure 4.2. Design schematic

4.2. Sentiment Visualization Module

Visualization comes as alast stage of the main module,where the extracted opinions are already classified by developing a web-based model and using interactive charts to visualize achieved sentiment. Visualization in the next design stage identifies the proposed visualization techniques.

4.2.1 Design

In the visualization stage, the summary generated by thedeveloped classifier in the first stageis going to be visualized based on thefive design sheets principle(*FdS*),(Roberts, 2011).

4.2.1.1 Methodology

According to Jonathan Roberts from Bangor University, five design sheet(FdS) from the name has to includefive sheets with a general layout as follows:

- ✓ Employssketching and designs to ideate novel information visualization solutions.
- ✓ Fivestages: meet, brainstorm,create design-sheets, discusswithclient, generate realization design.
- ✓ Fiveoutcomes :One brain storm sheet; Three design sheets; One realization sheet.
- ✓ Five parts to brainstorm: ideate, filter, categorize, combine& refine, question.
- ✓ Five parts to each sheet: layout, meta-information, focus, operations, discussion or detail(LIFOD; Roberts, 2011).

4.2.1.2 Five Design Sheet (FdS)

1. Brainstorming sheet

To increase design possibilities with multiple suggested ideas, a brainstorm sheet is used to include different types of design principles. The brainstorm sheet contains five items: generated ideas, ideas filtration, sketches categorization,sketches combination, refining and finally the question that responds to the visualization goal or task (Roberts, 2011). See Figure 4.3

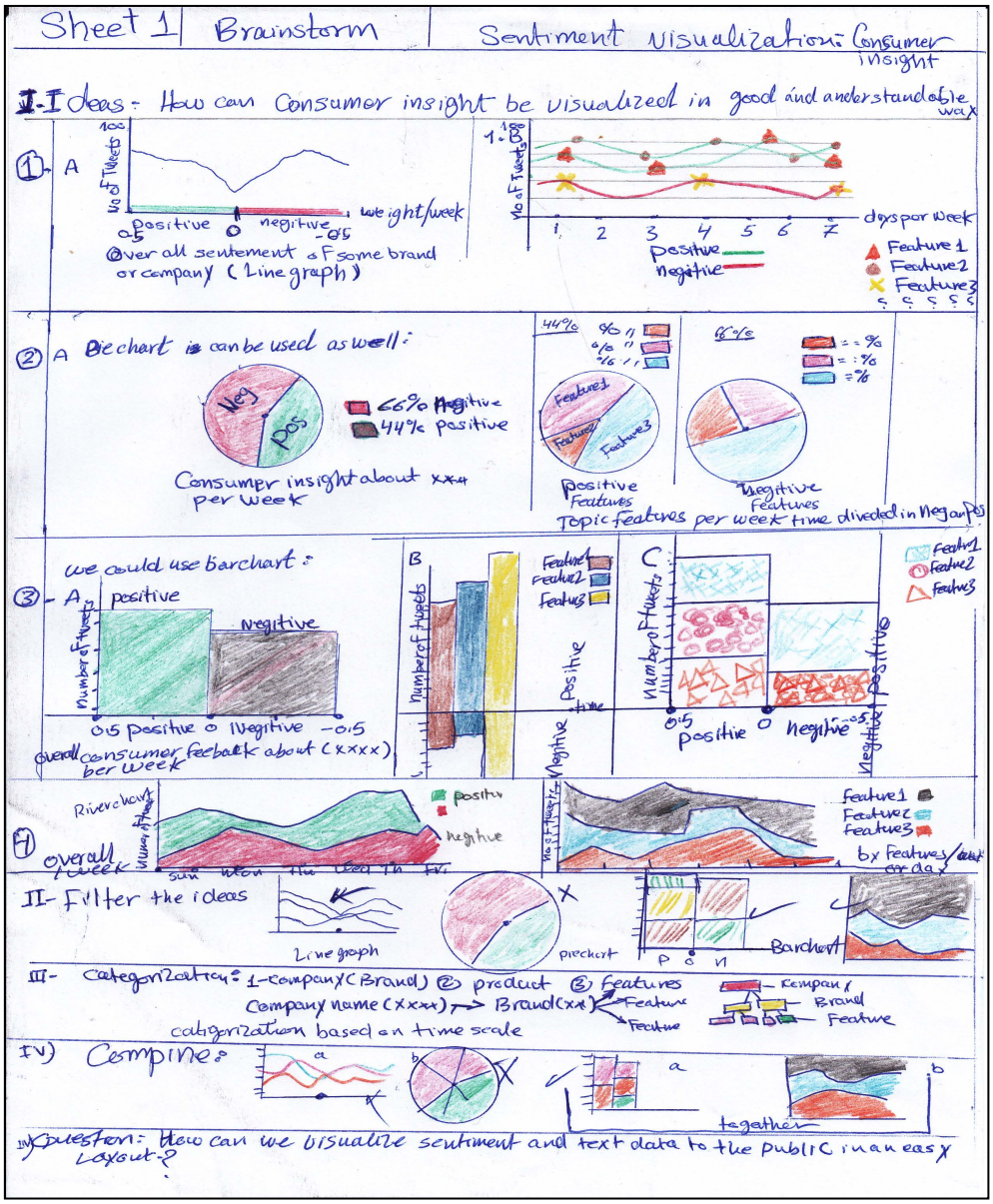


Figure 4.3 (Scanned FdS, Sheet-1)

Sheet number 2 includes the detailed sketch from one of the filtered ideas that was already stated in the brainstorm sheet. It illustrates the visualization using line graph as in Figure 4.4.

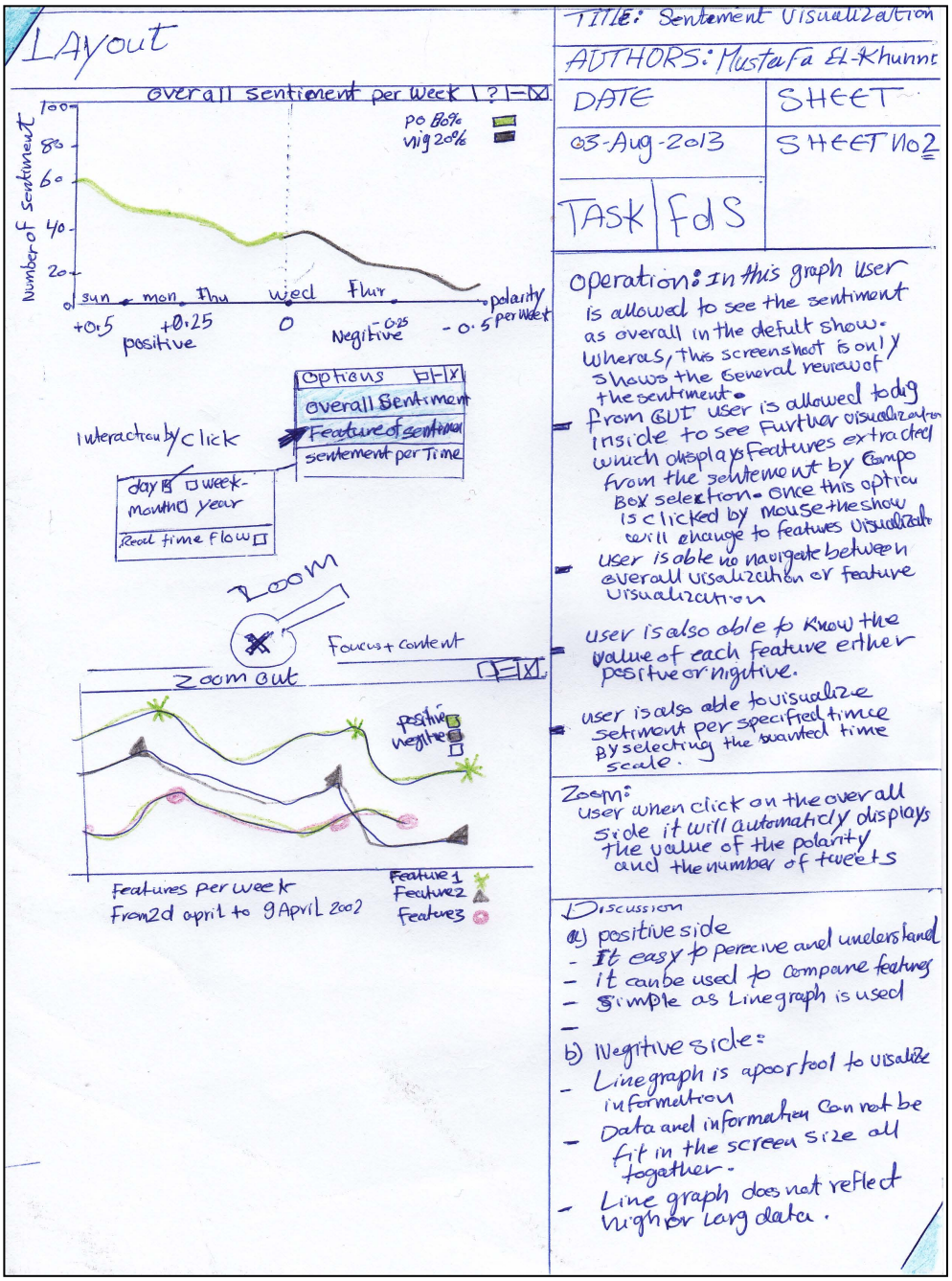


Figure 4.4 (Scanned FdS, Sheet - 2)

In the thirdsketch, another visualization technique was selected, theriver graph is described as in Figure 4.5.

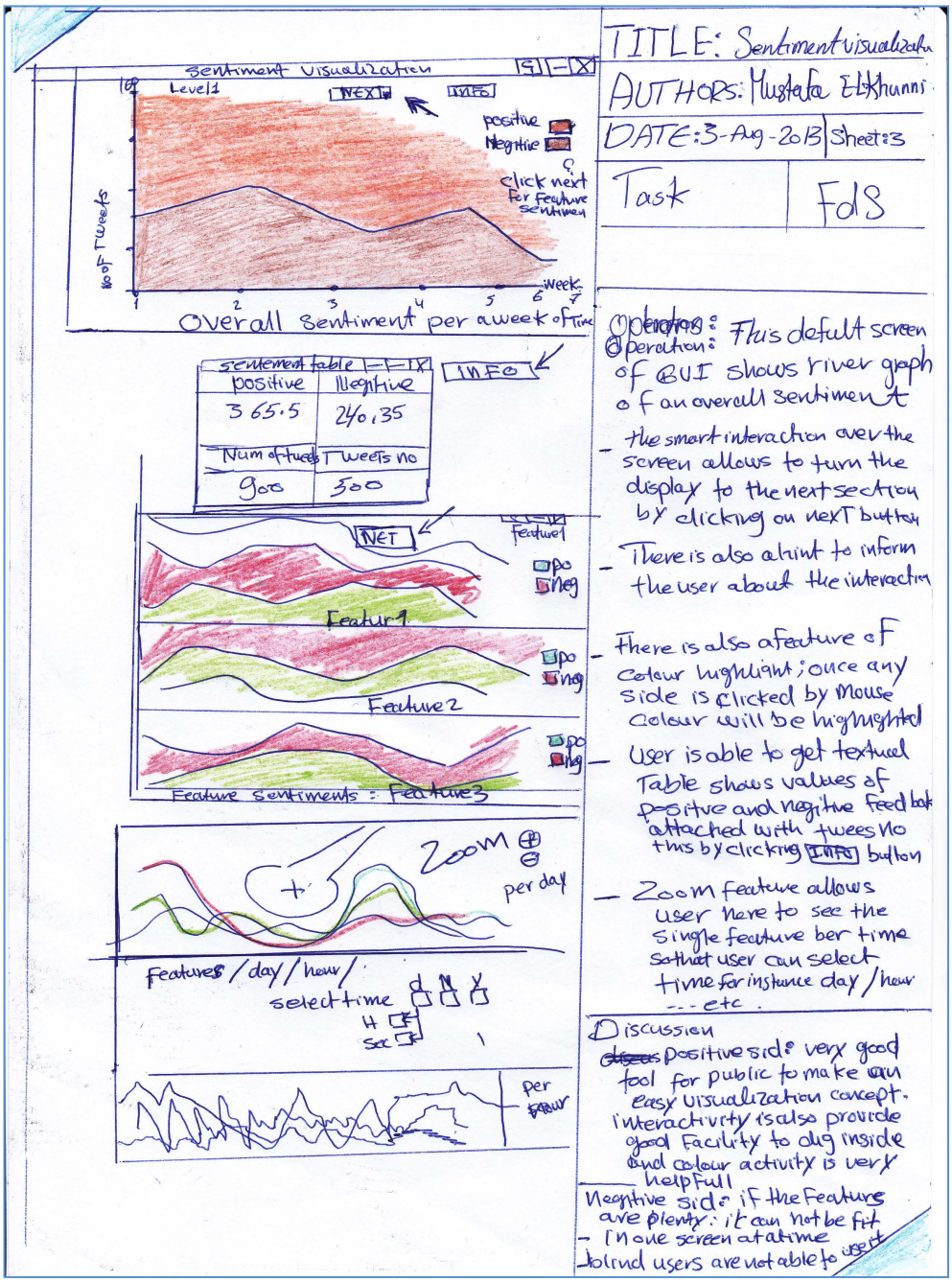


Figure 4.5 (Scanned FdS, Sheet - 3)

The fourth sheetshows visualization technique using a histogram, illustrating how sentiment can be visualized using negative column graphs as in Figure 4.6.

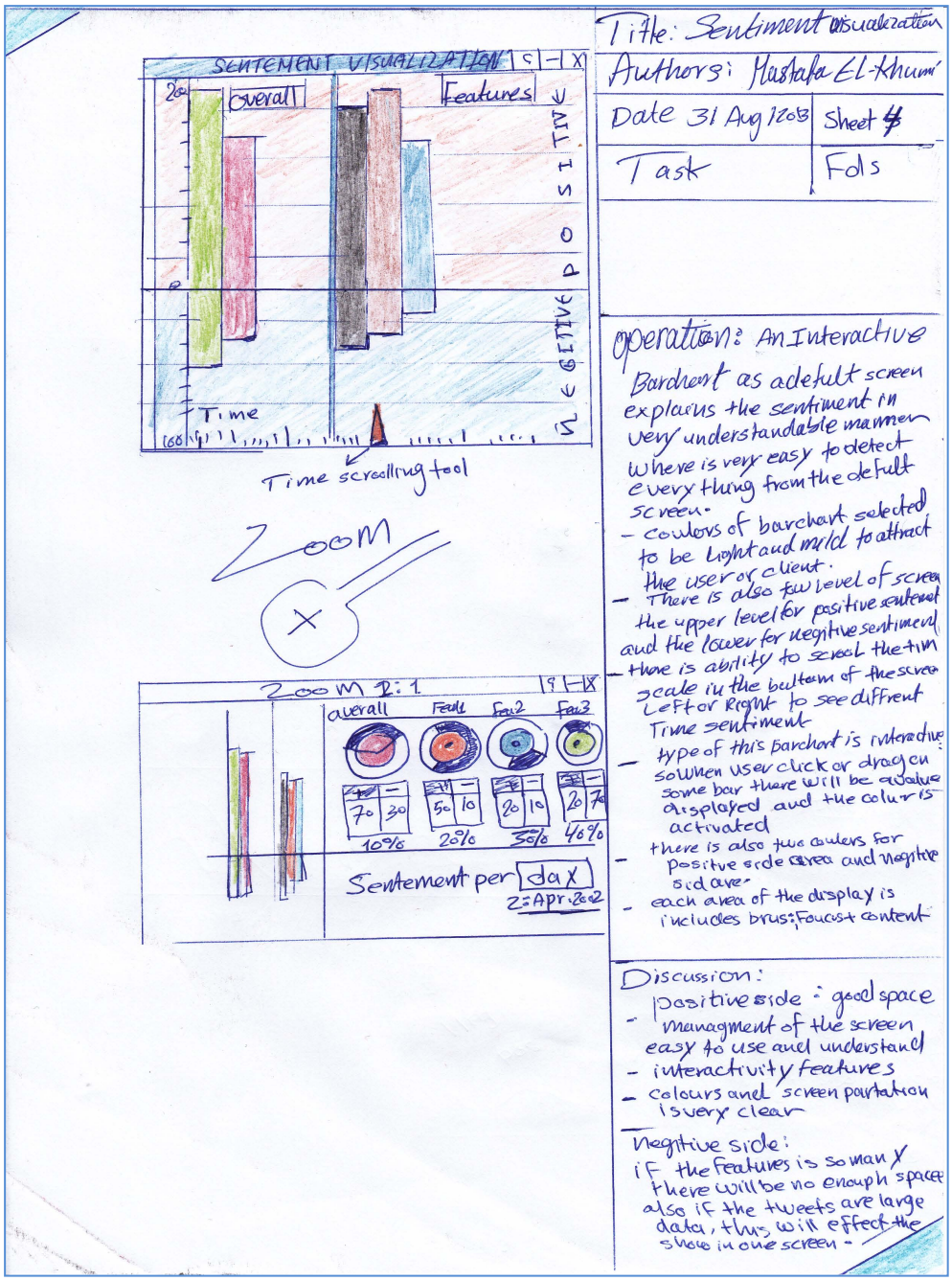


Figure 4.6. Scanned FdS, Sheet - 4

In FdS, sketchingthe principle is the last sheet, which represents the final choice or decision for the proposed visualization strategy, as in Figure 4.7.

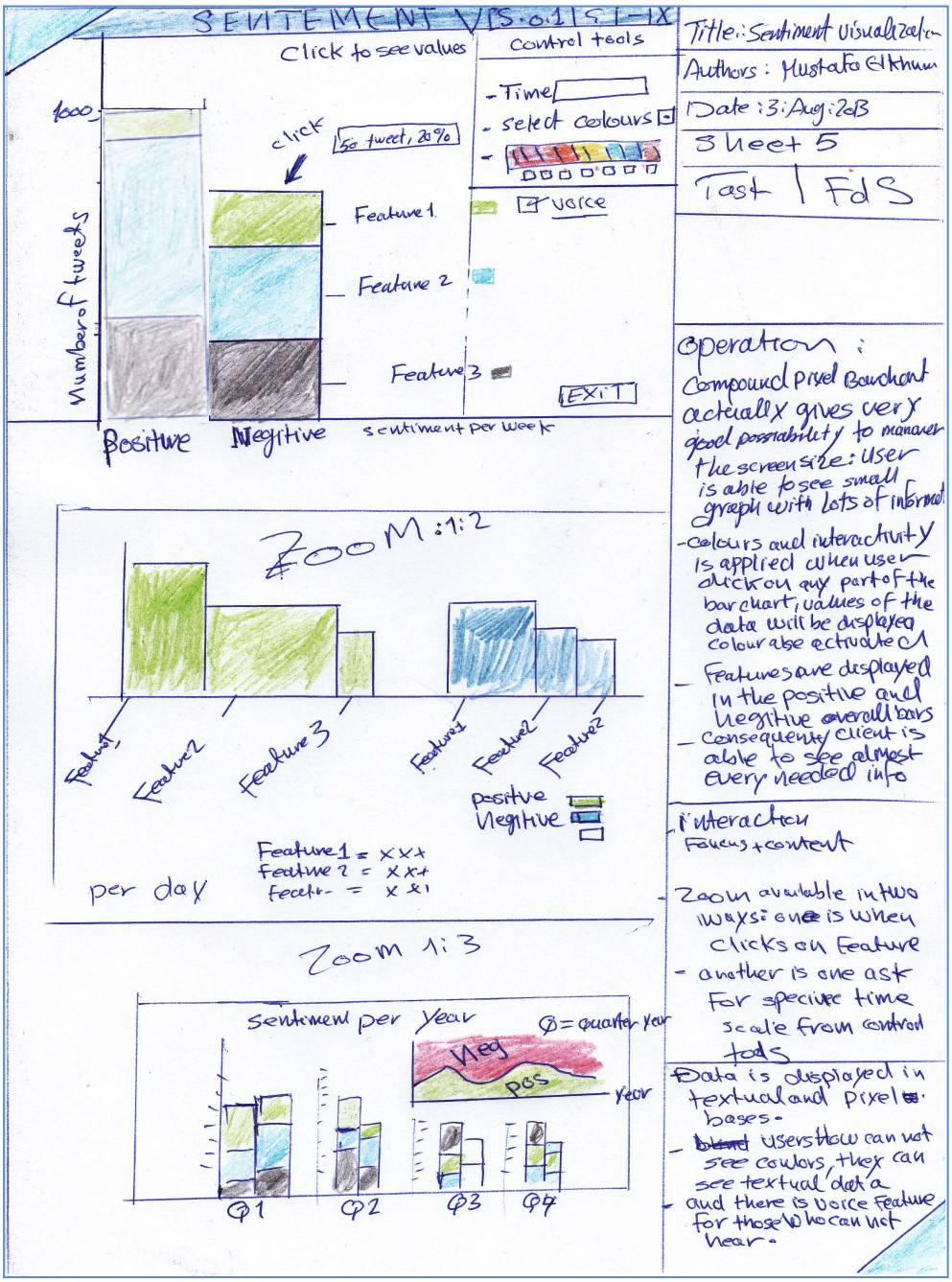


Figure 4.7. Scanned FdS, Sheet - 5

4.3 Project Requirement Specification

Starting from the design point of view, there is a need to use a suitable programming language to conduct this module code. There are some languages that support data mining techniques.Based on my brief review of programming languages, I found two languages supporting related libraries of natural processing language: Java and Python, with WEKA and NLTK librariesrespectively. Due to the necessity to use SQLdatabase and GUI in an objected oriented platform, Java wasselected. Furthermore, visualization techniques in Java can be done usingprocessing language techniques, Google charts or Highcharts, which were found as third party libraries.

4.3.1 Scope of the Project

This project is designed to provide a database for visualization, relying on text analysis of every text record retrieved from the Twitter database (API). It provides structured and quantified data after being processed based on natural language processing, for a front-end visualization model. Therefore, it must be platform independent, easy to use and easy to maintain. To fulfilfunctionalityof the final outcome of this developed module, Javaplays as the object oriented programming language with integratedplatform independence. The choose of open source MYSQL database allows costlessusage of databasewith a high level of performance. BothAPIs and libraries are open source with easyaccess. This project isbackend principle. Polarity calculation,annotation extraction, and the sentiment analyser run over the retrieveddata from Twitter and thensavedin a database.

4.3.2System Requirements

1. JDK 7.3(whichsupports HTML 5 projects,maven, GlassFish 3)
2. MYSQL server 6.0E.
3. Twitter API authentication and libraries
5. Highcharts free library

JDK 7.3 was selected because it is an independent platform providing a suitable environment for object oriented programming, in addition to its flexibility of integration with other systems, which allows a wide range of code reusability. Other listed requirementswere chosen due to availability with free cost and ease of use.

Chapter 5. Implementation and Experiments

Experimentation either with search keywords or supervised techniques based on WEKA machine learningwas implemented. The keyword-based approach does not require training data. The work is only on the test data. Moreover, it does not require the test data to be represented in a special way and can be done by string matching operations,whereas the supervised approach requires data cleaningand preprocessing on the training data and has to have both the training data and the test data represented in a specific way. The representation can be either bag-of-words or representation-based representation, or both. Basically, the supervised approach is a bit complex in terms of the further use of mathematics and probabilities. The keyword based approach was included for the sake of comparison.First of all, data has to be collected from Twitter API. The open source twitter4j library is used to implement Java/REST connector and to validate accessing to REST API.

5.1 Developing an Existing Model

The model usedwas developed by a group of students and published under an open source Apache 2.0 license.This model is created for company reputation based on NPLand machine learning techniques that match the target of theprojectat hand (Damaschinet al.,2012). To get this model working, the old Twitter API connector was completely modified into a newer version by using twitter4j library,anda completely new GUI interface using HTML and Java script was developed instead of using swing in Java. By use ofa servletlibrary,a web application based on this module was created. Furthermore, for visualizing the sentiment, theHighchartslibrary was employed;this is found online (http://www.highcharts.com) and is a free library with different types of

visualization chartsavailable to usefree for personal and non-commercial purposes. The reason why the Highcharts library was selected is because it is easy to use, provides a high level of interaction andis coded with HTML5 and Javascript, which can be renderedby different browsers(Highcharts, 2013).This Existing modelavailable online on Google code (http://code.google.com/p/twitter-sentiments).

5.2 Project Platform Preparation

There are non-functional preparations required to be carried out.

5.2.1 Register for an Application with Twitter

Aftercreating a new application in Twitter API, there are credentials given that have to be used as an authentication to retrieve Twitter data. Set of credentialsrepresent in consumer secret, consumerkey, access token andaccess token secret are codedinJava/REST connector code to start the key word search in Twitter content.

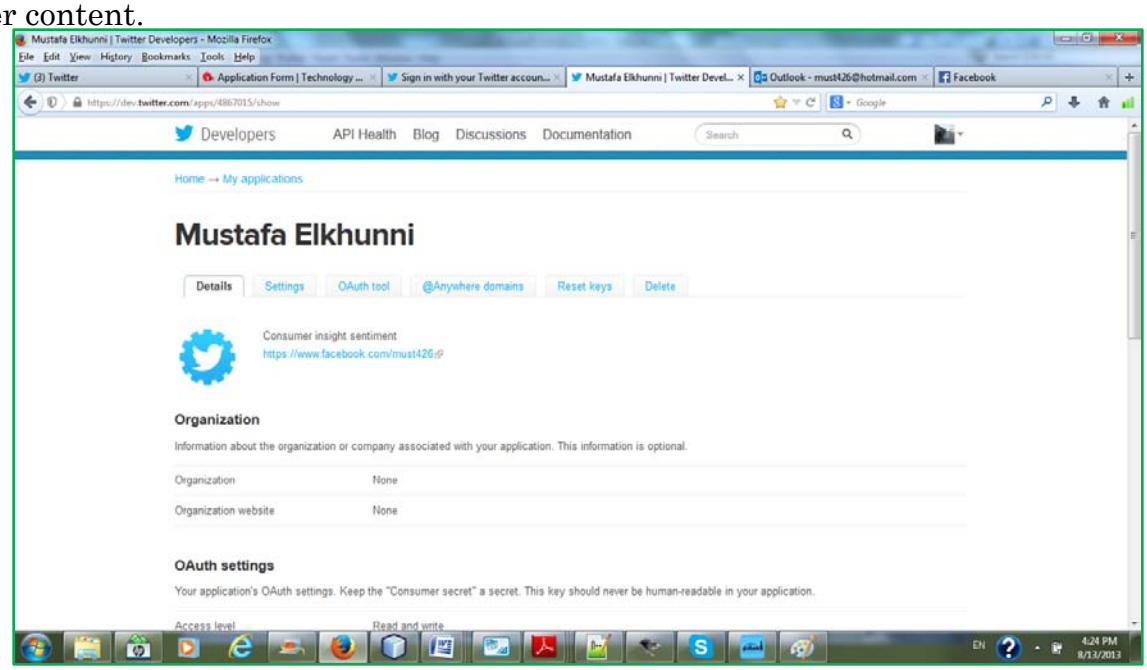


Figure 5.1. Registration confirmation on Twitter API application

5.2.2 Global Variables and Authentication

Configuration ofthe MYSQL Database serverhas to prepared, making sure that query cache size is suitable, such as 16 MB or higher.

5.2.3 Workstation or PC Specification

As this project is a server client application, the server side needs to be a high performance PC with high performance processor, 3.0 GHz speed or higher and 4 GB ofphysical RAM or higher.

5.3 Project Flowchart

As indicated in the design chapter, the proposed system includes multi-server usage.The system acts as a clientsystem to the server of Twitterduring the session ofaccessing tweets; it requesting and get respondedvia the internet, fulfilling client credentials of the TwitterAPI. Whereas, This proposed system behaves asa server side by dealingwith components of text analysing,while the user is interacting with the graphical user interface of the system by searching and browsing graphs as a client for integrated web application. The project flowchart is shown in Figure 5.2.

5.4 Twitter Connector

The tweet connector module(TwitterOperator) is actually providing both connectivity to Twitter API and to MYSQL database. To avoid unsuccessful data retrieval from Twitter API, three query options have been deployed in this algorithm. This way,the user when placing a search keyword will not feel a delay between each search session in case eitherthe SQLdata connection or the Twitter APIis down. Further, the option list allows managing the use of the required database, such as when only the database is needed as in the code below: 1,2,3 static variables representingoptions of query (3=local SQL database and online Twitter data, 1= SQL database onlyand 2=Twitter online data only).This

algorithmimplements a user search session from Twitter online search to the required queries on the stored database.

```
private Operator db;
private int mode;
public static final int DB_AND_ONLINE = 3;
public static final int DB_ONLY = 1;
public static final int ONLINE_ONLY = 2;
private static final String CONSUMER_KEY = "XwEagf0oSlyPE59Z86fQ";
private static final String CONSUMER_SECRET = "8nNHG4Tbz8JH57YecP87vwheX464dWM9wufJnMdRLY";
private static final String ACCESS_TOKEN = "89405802-ez85Nkr7r2mn2EfFdS0MXPadMsw1E5YCx6oixdbIg";
private static final String ACCESS_TOKEN_SECRET = "2ykkYdU5qiloVML43CRaffb79RF7MIC9Q3mb9Y7o";
Twitter twitter;

public TweetOperator(int mode) {
    try {
        this.mode = mode;
        db = Operator.getInstance();
    } catch (Exception ex) {
        ex.printStackTrace();
    }
    if (mode == 3 || mode == 2) {
        TwitterFactory twitterFactory = new TwitterFactory(buildConfig());
        twitter = twitterFactory.getInstance();
    }
}

private Configuration buildConfig() {
    ConfigurationBuilder cb = new ConfigurationBuilder();
    cb.setDebugEnabled(true)
            /**
```

```
public LinkedList<LocalTweet> getTweets(String queryString) throws TwitterException, IOException {
    LinkedList<LocalTweet> results;
    if (mode == DB_AND_ONLINE) {
        results = getTweetsFromDB(queryString);
        results.addAll(getTweetsWithAPI(queryString));
    } else if (mode == DB_ONLY) {
        results = getTweetsFromDB(queryString);
    } else if (mode == ONLINE_ONLY) {
        results = getTweetsWithAPI(queryString);
    } else {
        results = new LinkedList<LocalTweet>();
    }
    return results;
}
```

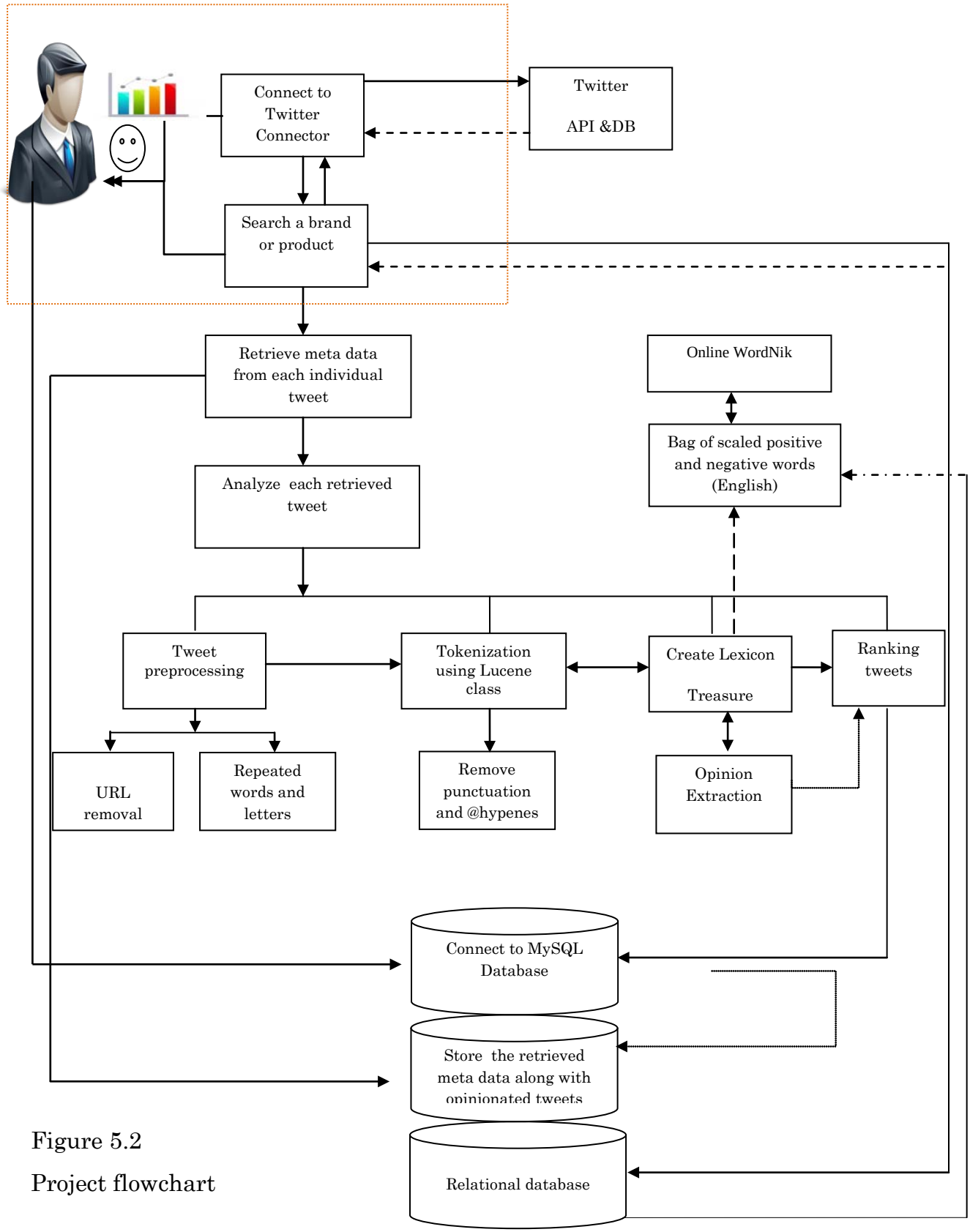


Figure 5.2

Project flowchart

5.5 Preprocessing Techniques

Before sentiment analysis processing takingplace, tweets are tokenizedusing white spaces as delimitersin a simple tokenization technique. Consequently, text clean up on collected tweetsneed to be carried out. Text cleaning is applied in order to clean tweets from noisy and unwanted data. A list of stop words as belowwas created to iterate tweets based and remove any matches from tweet tokens. Next list shows some the stop words used in preprocessing stage.

```
String[]stopWords={"a","about","above","across","after","afterwords",
"again","against","all","almost","alone","along","already","also","although","always
","am","been","but","by","","","could","dear","did","do","does","either","among","amo
ngst","amoungst","amount","an","and","another","any","anyhow","anyone","anythi
ng","anyway","anywhere","are","around","as","at","back","be","became","because","
become","becomes","becoming","been","before","beforehand","behind","being","belo
w","beside","besides","between","beyond","bill","both","bottom","but","by","call","ca
n","often"
```

```
/**
 *
 * Start cleaning tweet using stop words list
 *
 */
StringTokenizer tokens = new StringTokenizer(Tweet, " ");
        String newTweet = "";
        int i=0;

        while (tokens.hasMoreTokens() && (i<stopWords.length))
        {

             String temp = tokens.nextToken().toLowerCase();

             if (!temp.contains(stopWords[i]) ) {

                 newTweet += temp + " ";

                 i++;
            }
        }

        tweet = "";
        tweet = newTweet;
        System.out.println(" \n"+tweet);
        return tweet;

        }
```

As it can be seen from the above piece of code from the preprocessing algorithm, each tweet is tokenized first by use of stringtokenization 'StringTokenizer tokens = new StringTokenizer(Tweet, "");' and then cleaned using the stop words list. One more technique is also deployed in preprocessing to remove duplications in the tokens such as correcting some commonly used slang (looool==>lool, Greeeeat==>Greeat and so forth). The mechanism to remove duplication is used to reduce repeated letters in each token to make it closer to the right form of the original term that can be enhanced to be more effective if bag of words of English

slang is employed in this technique. Another preprocessing technique employed in preprocessing model is called N-grams;this can be used for different tasks oftext extraction. Here an N-gram simple algorithm is also used as a tokenization mechanism to detect repeated words. Its algorithm is listed inthe following pseudo code (Okazaki &Tsujii, 2010).

```
Input: X: array of features of the query string
Input: τ: minimum number of overlaps
Input: V: collection of strings
Input: l: size of target strings
Output: R: list of strings similar to the query
M ← {};
R ← [];
foreach q ∈ X do
    foreach i ∈ get (V, l, q) do
        M[i] ← M[i] + 1;
        if τ ≤ M[i] then
            append i to R;
        end
    end
end
return R;
```

N-gram (Okazaki &Tsujii, 2010)

```
public static ArrayList<String> getNGrams(String s,int minNgram, int maxNgram){
    ArrayList<String> ngrams = new ArrayList<String>();
    for (int n = minNgram; n <= maxNgram; n++) {
        for (String ngram : ngrams(n, s))
        {  ngrams.add(ngram);}

    }

return ngrams;
}
```

In the above code N-gram function returnsa list of N-grams as shown next.

Begin

-Initialize max and min parameters

-input array of string

-if min = 1 and max = 3 will return

-nGrams with 1 token, ngrams with 2 tokens and ngrams with 3 tokens

-if min =2 and max = 3 will return

-nGrams with 2 tokens and 3 tokens

-Where is minNgram represents minimum number of token for a Ngram

-MaxNgram represents maximum number of token for a Ngram

- s is the tokenized String

- return arraylist with results

-End

This is a test of N-gram function:

```
    public static void main(String[] args) {
        ArrayList<String> test = getNGrams("Hello Bangor, I am about finishing my MSc project!",2,2);
            for (String ngram : test )
                System.out.println(ngram);


    }
}
```

Results from the N-gram function test:

Hello Bangor,

Bangor, I

I am

am about

about finishing

finishing my

my MSc

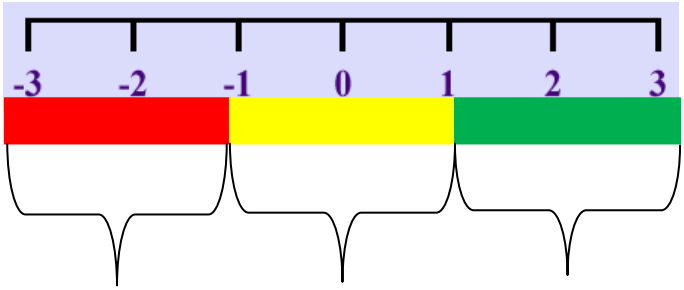
MSc project!

5.6 Sentiment Extraction and Polarity Scores

As soon as pulled tweets are cleaned in the preprocessing algorithm, the sentiment analysis process will start in order to extract opinion and assign weights of polarity for individual tweets.

5.6.1 Lexicon

Three labels of polarity are addressed as a scale forextracted opinions:positive, negative and neutral, where positive tweets are assigned to the range ofgreater than zero (positive tweet ====> (If total scores of tweet is (less than> 0)), (negative tweet====>(If total scores of tweet is(< 0)) and neutral tweet is a tweet with 0 scores (neutral tweet===>(if total scores of tweet is(=0)) as shown in the polarityscale in Figure 5.3.



NegativeNeutralPositive

Figure 5.3. Polarity scale

Simplistically, this algorithm is responsible for assigning a score to each token of a tweet and aggregating all scores of weighted tokens to find total scores of these tokens. It overall has to determine the sentiment of a tweet, which must match one of the three sentiment statuses listed in the polarity scale inFigure 5.3. To dig inside this algorithm, SQL Database has been used to store lexicon words, modifiers and negations in database tables. In the lexicon table, each word is stored with its score as described in the example Table 5.1.

Id	Word	Score(Sentiment)
1	abandoned	-0.5
2	abandonment	-0.5
3	abase	-0.5
...	...	...

Table 5.1. Lexicon database

By use of this table, each word in tweet tokens are scored accordingly. Similarly, we employed an onlinetreasury for synonyms or related words to expand the query session by employing WordNik after getting a key from http://developer.wordnik.com.Inaddition,Wikipediawas also used in case of no

treasury found to the query(MihaiDamaschin et al., 2012).There is also a function to detect smiley faces with the scoring mechanism.

5.6.2 Negations and Modifiers

English language,like any other language, often uses some linguistic terms either to confirm a positive opinion or to make it strongly negative.On MYSQL database, both modifiers and negation lists are stored to be used by the sentiment analysis algorithm in order compute polarity of tweets. For instance, if we got a tweet that contains 'Iphone5 is not good', the score of good is +1 and because of negation (not) the applied mechanism is to multiply the two words that come after *not* by (+1) × (-1); the final sentiment of the tweet is(-1), which is negative. Modifiers are words like 'really', 'quite', 'slightly' and so forth and are called modifiers because they modify the sentiment score ofthe word that comes after it. Each modifier word has its strengthto be multiplied by a factor; for instance 'very' has a strength factor 2, so that following words will be multiplied by 2, while 'really' will be multiplied by 0.5. Thenext piece of code ofthe sentiment analysis algorithm shows how scores of modifiers are computed(.

```
 */
private double getSentiment (LocalTweet tweet) throws IOException {
    LinkedList<String> tokens = tokenizer.getTokens(tweet);
    Double score              = 0.0;
    Double negation           = 1.0;
    Double modifier           = 1.0;
    int range                 = 0;
    Double gauss              = 1.0;
    int i                     = 0;
    for (String token : tokens) {
        gauss = getDistance(parent,token,tokens,i,this.gaussian);
        score += (negation) * (modifier) * (gauss) * lexicon.get(token);

        // Deal with negations
        if (isNegation(token)) {
            negation = -1.0;
            range = 2;
        }
        else if (range > 1)
            range--;
        else
            negation = 1.0;

        // Deal with modifiers (like 'very')
        modifier = getModifierValue(token, negation);

        i++;
    }
```

In case of a comparative or contradicting opinions such as 'I like Sony, but I hate Samsung', Gaussian distribution with maximum distance is employed to calculate the frequencies based on the distance between words according to the following formula:

$G(\mu.x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\pi)2}{2\sigma^2}\right)$where μ ismean and σ is variance(Eugene Weinstein).

This function will increase and decrease the sentiment scores according to the distance ofwords from the object word.In the above example, Sony is the object and other words 'I' and'Like' are words to be scored based on their distance from 'Sony'.The function is shown in the next code,after performing all sentiment analysis:

```
/
blic double[] CalculateDistance() {
    double sigma = 1.0;
    double gaussian_max = Math.exp(-(1/(sigma*sigma*2)))  / (sigma * Math.sqrt(2 * Math.PI));
    for (int i=0;i<=5;i++) {
        gaussian[i] = (1/gaussian_max)*Math.exp(-(((i)/sigma)*((i)/sigma)) / 2) / (sigma * Math.sqrt(2 * Math.PI)
    }
    return gaussian;


ivate double getDistance(Search parent, String token, LinkedList<String> tokens, int i, double gaussian[]) {
    String entity = parent.getQuery().getRequest();
    int x = Math.abs(tokens.indexOf(entity));
    double height = Math.abs(lexicon.get(token));
    if (Math.abs(x-i)<=5)
        return height*gaussian[Math.abs(x-i)];
    else
        return 0.0;
```

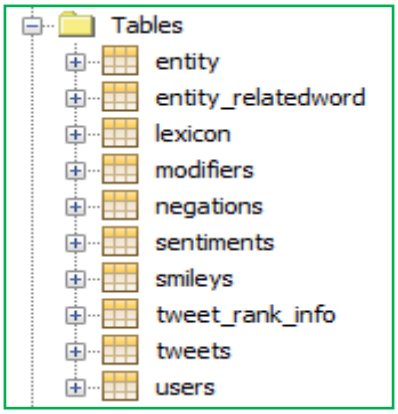
5.6.3 Tweets Ranking

Database found on Twitterprovides many characteristics of the tweet holder, such as number of followers, number of tweets, retweets number and number of

friends;theseare not important for sentiment analysis, but it can give feedback about tweets effectiveness, importance, popularity and so forth. Some of these features cannot be retrieved by Twitter API, so that the ranking procedure is limited and implemented in this module to represent retweet count, number of friendsand favourites and to store them in the MYSQL database. Furthermore, there is an attempt to make page ranking, but Twitter API still restricts many functions.

5.7 Data Storage

MySQL server 6.0E is configuredand database tables are created to be a data warehouse to store meta data and generated sentiments of retrieved tweets in addition to tables of tweets content, lexicon, negations, modifiers ,smileyand so forth as in Figure(5.4). The database plays an essential search engine role for user queries and at the same time presents the row data for visualization. MYSQL provides a complete text-search index and because Twitter search for keywords is permitted with limited retrieval, each search sessionincreases the tweets subset in the database, so the user is allowed to retrieve data from both Twitter search and database search. Sentiment from each search session is computed by adding the on-goinganalysed tweets pulled from Twitter API to the relevant tweets already stored in the database. storing tweets is coded to avoided repeated tweets. Here is how overall polarity is computed:Sentiment of search keyword(X)=(Computed sentiment from analysing and pulled tweets of keyword(X) from twitter API+Computed sentiment of already analysed tweets of the same Keyword(X) stored on database).



Figure(5.4)MYSQL database tables

5.8 Consumer Insight Visualization and User Interface

Automated consumer insight tools are relatively new and still not widely used. The majority of users will be somewhat new to consumer insight tools. Therefore, designing a smooth GUtokeep it simple and studied interface is a priority, as this graphical user interfacewill be dedicated for users from the public.According to Smith,five design maximswere considered to design the GU(Tex font, justifiedtext, font size, regularsentences and black text colour. Consumer insight visualization also needs to be an interactive and simple tool to allow an easy to use and provides a wide range of interactivity (Smith,1998).

5.8.1 Web Application Configuration

Today there is wide use of web-based applications due to simplicity and browsing ability from different browsing platforms. A web page application for consumer insightis designed as indicated in the design chapter and created to provide both user interface and visualization graphs of consumer insight.

By use of a servlet class, which is a Java programming languageclass used to extend the capabilities of a server,the system can accept any type of requests(Wikipedia, 2013). In this system servlet is used to extend the sentiment analysis application hosted by a local host (8080). The GlassFish web server 3.0 server in the NetBeans platform was used to create the web application project by defining a local domain and a remote domain to create a client and server model. Furthermore, GlassFish provides a connection pool to enhance the performance of implementing commands on a database in order to speed up the response time created on the database from the client web page while keeping the connection to database active all the time with no need to establish a new connection for each request session.GlassFish is mostly used for dynamic database on web applications. The pool connection requires MYSQL database and MYSQL Java driver to be implemented on NetBeans

IDE(Wikipedia, 2013).Because weneed to use JavaScript, a JavaScript engine has been included in the model to provide Java scripting capability into theJava environment from the server side to client side. The JavaScriptengine codeis shown below.

```
public class SentimentController extends HttpServlet {

    SentimentScriptEngine jsEngine = new SentimentScriptEngine();

    protected void processRequest(HttpServletRequest request, HttpServletResponse response)
            throws ServletException, IOException {
        response.setContentType("text/html;charset=UTF-8");
        /* If Compare button pressed use DB to pull scores from sentients table */
        if (request.getParameter("compare") != null) {
            compare(request, response);
        RequestDispatcher dispatch = getServletConfig().getServletContext().getRequestDispatcher("/Comparison.jsp");
            dispatch.forward(request, response);
        /* Otherwise pull tweets from TwitterAPI and DB, analyze and visualize */
        } else {
            RequestDispatcher dispatch = getServletConfig()
                    .getServletContext().getRequestDispatcher("/Result.jsp");
            String q = request.getParameter("token");
            Search s = new Search(q, 3);
            Results r = new Results(s);
            String result[] = {r.getEntity(), "" + r.getPositive(), "" + r.getNegative(), "" + s.getTweets().size
            request.setAttribute("result", result);
            request.setAttribute("timeline", r.getDates());
            request.setAttribute("timelineValues", getTimeLineData(s.getTweets(), s.getQuery().getRequest()));
            dispatch.forward(request, response);
        }
    }
```

Aftertesting the created web application, data has to be sent from the server side to the client side based on the client side request; theservlet class will organize theserequests to and from theclient side. The target data available for the client from the server side is controlled by a model played as a controller or a bridge between the client and server for catchingclient queries and satisfying them. In this model,JavaScript engine capabilities are used to script data to web application in the JavaScript environment as shown in the next code. Client queries placed by the user are fulfilled, stored in arrays and sent again to the client.

```
public class SentimentController extends HttpServlet {

    SentimentScriptEngine jsEngine = new SentimentScriptEngine();

    protected void processRequest(HttpServletRequest request, HttpServletResponse response)
            throws ServletException, IOException {
        response.setContentType("text/html;charset=UTF-8");
        /* If Compare button pressed use DB to pull scores from sentients table */
        if (request.getParameter("compare") != null) {
            compare(request, response);
        RequestDispatcher dispatch = getServletConfig().getServletContext().getRequestDispatcher("/Comparison.jsp");
            dispatch.forward(request, response);
        /* Otherwise pull tweets from TwitterAPI and DB, analyze and visualize */
        } else {
            RequestDispatcher dispatch = getServletConfig()
                    .getServletContext().getRequestDispatcher("/Result.jsp");
            String q = request.getParameter("token");
            Search s = new Search(q, 3);
            Results r = new Results(s);
            String result[] = {r.getEntity(), "" + r.getPositive(), "" + r.getNegative(), "" + s.getTweets().size
            request.setAttribute("result", result);
            request.setAttribute("timeline", r.getDates());
            request.setAttribute("timelineValues", getTimeLineData(s.getTweets(), s.getQuery().getRequest()));
            dispatch.forward(request, response);
        }
    }
```

5.8.2 Developing Visualization Tools

As long as the web application is already established and the project URL is created (http://localhost:8080/SentimentAnalysis4Tweets/sentiment), the client side is able to render JavaScript, html and CSS. Therefore, the coding environment becomes a web-based development techniques. The JS/HTML5 Highcharts library is employed to provide visualization graphs using both pixel bar charts and negative histograms for brand or product sentiment analysis and brand or product comparison respectively. Data on charts is provided from the server side using the JavaScript engine and servlet according to each user request, as described in the web application section 5.7.1. It provides interaction of the type focus + context.The sentiment trend informationcontainssingle or multiple category. Each category is prepared as a JSON object. Each object holds all the required information to display a session in the sentiment trendview. Two main JNSON are prepared to lunch graphs of search and compare actions.

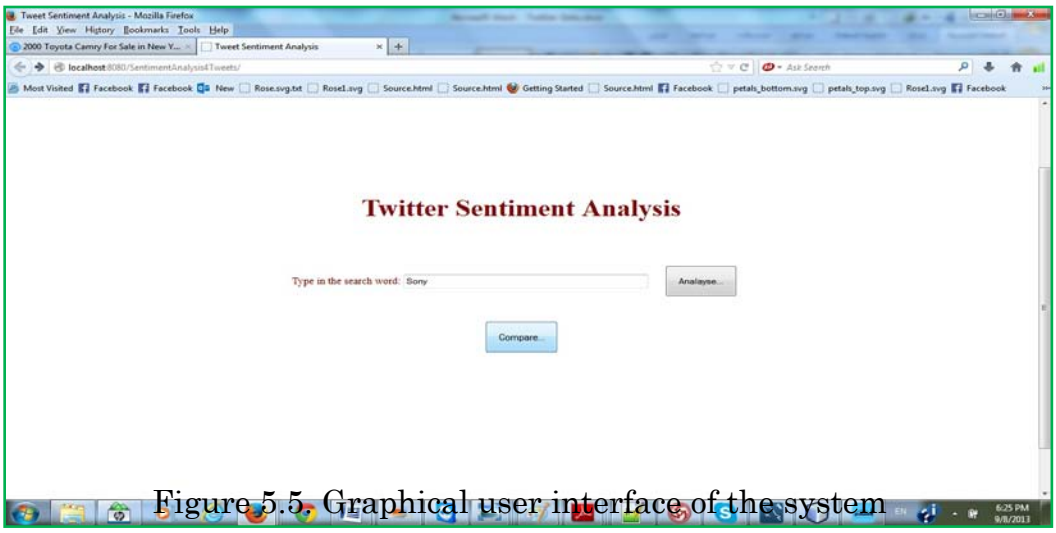


Figure 5.5. Graphical user interface of the system

Once the system is run and shows a user interface as in Figure5.5, and as soon as the keyword search is placed by the user and the analyse button is clicked, visualization of polarity will be displayed. In this case, data will be pulled from Twitter API and analysed, then stored and added to the classified tweets inthe database. The visualization graph of a pixel bar chart shows the analysed tweets in

the database in addition to the tweets already analysed from the Twitter API, which should match the search keyword. See the resulting bar chart in Figure 5.6.

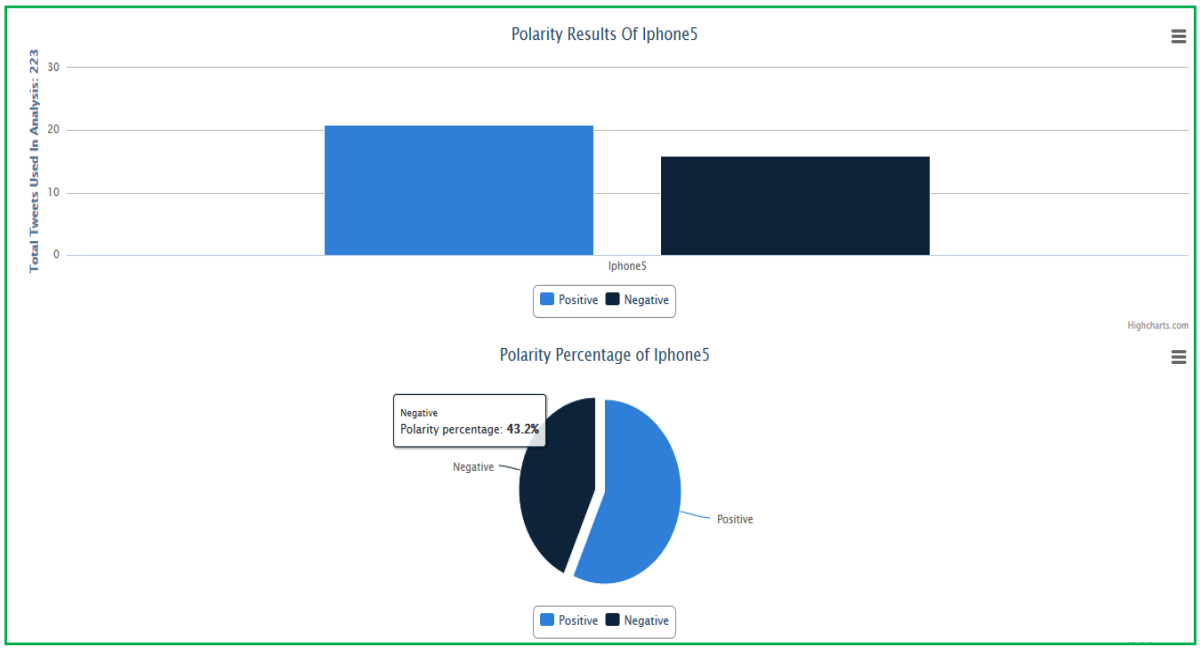


Figure5.6. Sentiment visualization resulted from user keyword search action

From the above graphs, the user is able to see brand reviews, which is a reputation based on sentiment analysis model in two types of graphs. The bar chart shows negative and positive values,and the pie chart shows sentiment percentage. With two different colours and interaction over both the bar chart and pie chart, the user is able to distinguish both classifications and interact with the graphs.

5.9 Implementation Challenges and Technical Issues

There were several technical issues experienced within the implementation stage:

i) Carrying out web application techniques and howto implementweb applications in addition to determining the right technique to perform this task was one of the challenges because there are many solutions offered for web application creation. Socket, GlassFish server and servlet and Apache tomcat server are different tools to create web applications.

GlassFish and servlet technique were selected because it is easy to implement and provides good techniques to deal with the database as the module has to interact with a largedatabase.

ii) The wide range of dependencies needed to implement thissystem have to be applied to use theNetBeans project platform and without this,the system will not run. See used dependencies in Figure5.7.

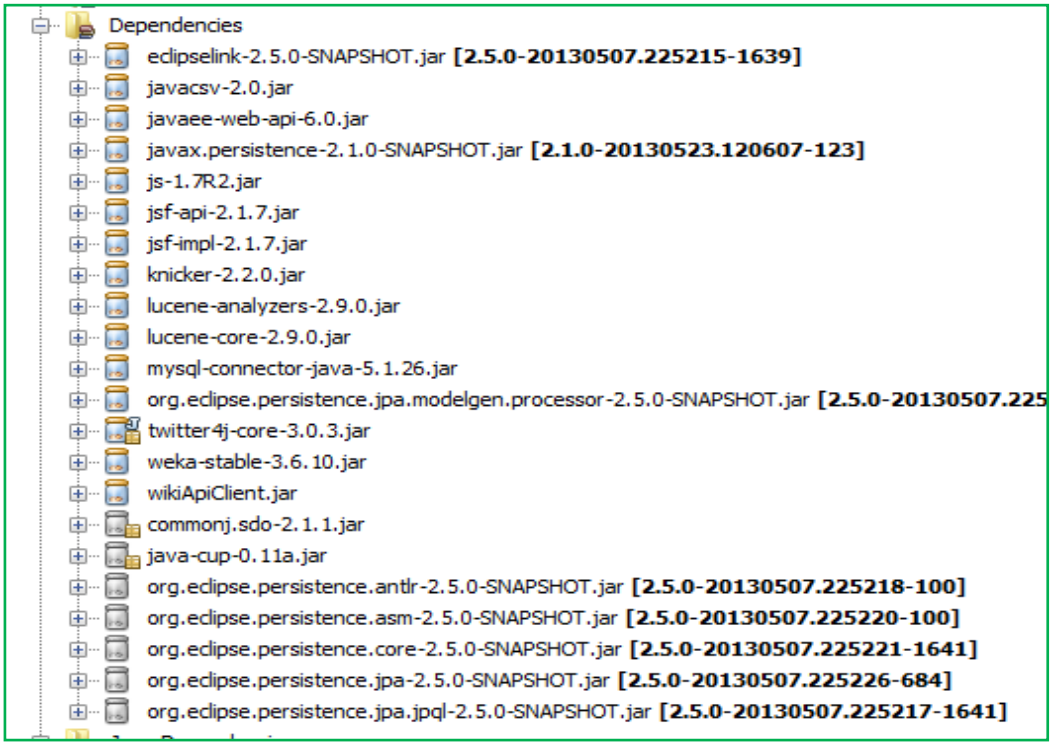


Figure5.7. Used dependencies

iii)The duration time which isthe response time between client and server is always considered a priority. The user should not feel any delay when the system is used. Actually this issue was not solved completely, but it was slightly overcome by improving the search over the database by focusing only on the required table and row from the database tables.Domadarhas also suggested some solutions to this issue by use ofKISS Which is Keep It Simple and Straightforward Principles for OR-Mapping Products from Software Tree, Inc which can be considered in future(Domadar, 2005).

Chapter 6. Analysis and Results

6.1 Evaluation

In this chapter, results from the system will be analysed and evaluated based on text mining metrics. Furthermore, a complete software package of the project with class diagram an UMLdocuments generated by Visual Paradigm forUML professional is available on the enclosed CD with this project copy.

6.1.1 Case Study

Two companies, Sony and Apple,were selected as a case study to test the system and to compare both results in terms of sentiment. About 1500 tweets each were collected in 15 days' time, and because of the type of access to Twitter API,the average of collected tweets within an individual keyword search session is about 15 tweets. It is possible to make about 150 search sessions per hour. In the graphical user interface there is a compare button used to show comparative figures of polarity using negative column graphs for both companies. Figure 6.6 shows the results when the compare button is clicked by the user.

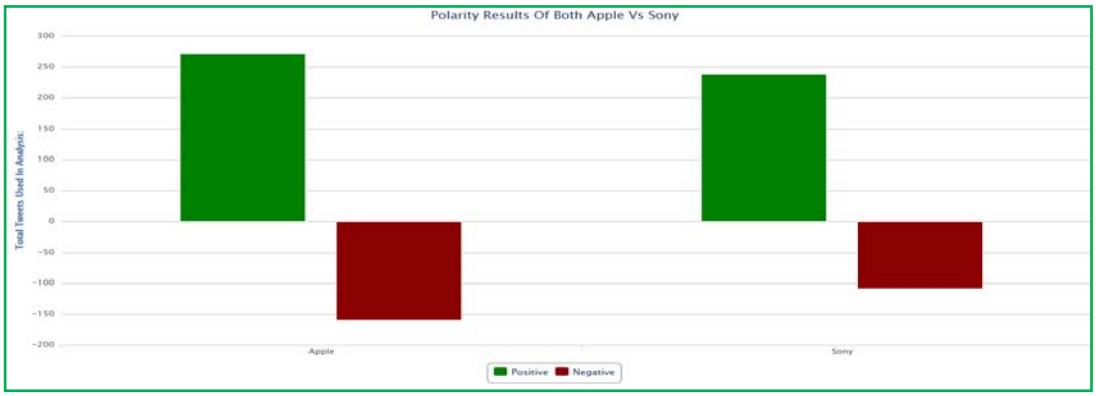


Figure 6.1. Comparative column graph for Apple and Sony

Table 6.2 shows sentiment classification including neutral tweets, which have been neglected in the visualization due to the confusion that it might cause for the user as those are not subjective results.

Company	Positive	Negative	Neutral	Total tweets
Sony	239	108	1117	1464
Apple	275	160	1028	1463

Table 6.2. Sentiment classification per both companies, Apple and Sony

To show how the sentiment changed over timeforboth companies,Figure 6.7 shows both Sony and Apple sentiments over time respectively from 21/08/2013 to 12/09/2013.

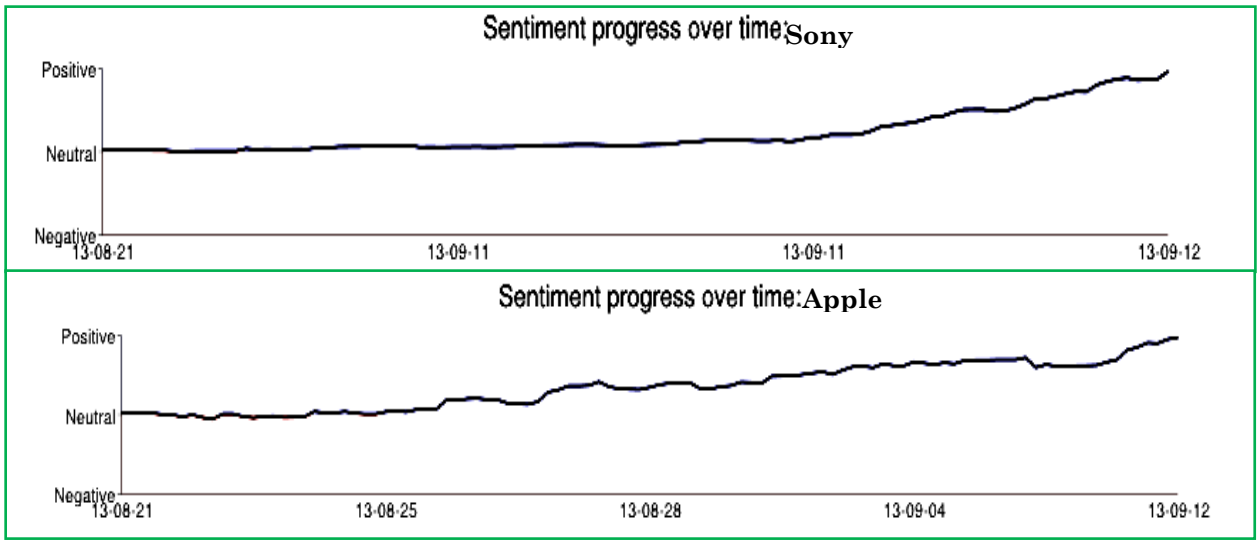


Figure 6.2. Sentiment time line of both Apple and Sony.

From the above results, it can be observed that there is a significant change of sentiment over time, and the reputation of Apple is much higher than Sony.

6.2 Measuring Algorithm Effectiveness

Sentiment analysis based on unsupervised text processing has many metrics methods. Because there is more than one method already built into the classifier, the accuracy is the results of all those methods,which aretokenization, text normalization, noisy clean up algorithm and the scoring algorithm based on negations, smiley and modifiers.

Precision and recall are one of the metrics methodsused to measure the effectiveness ofthe algorithm. While *precision* is the ratio of the totalof relevant instances(correctly classified)restoredto both the aggregation of relevant and irrelevant instances retrieved, *recall*known asthe fraction of relevant instances were retrieved(Wikipedia, 2013).See Figure 6.3

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)}$$

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)}$$

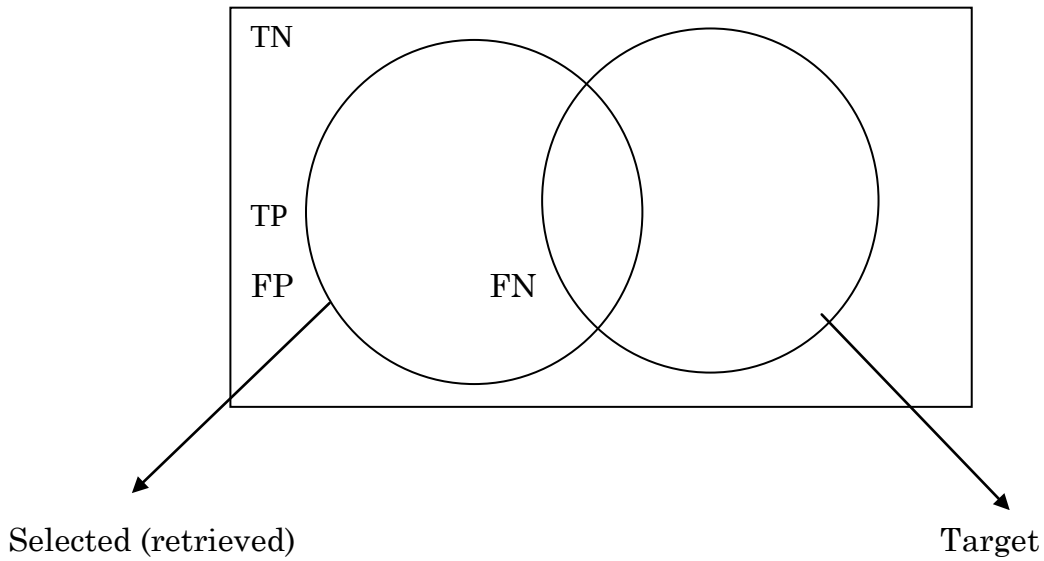


Figure 6.3. Evaluation Concept diagram (Manning et al.,2011)

To measure the accuracy ofthe algorithm, a machine learning naive Bayes classifier for evaluation is used. Carefully classified data are used to train the classifier, and some of it is used as test data. Sanders Twitter sentiment groups have been downloaded and used to train the naive Bayes classifier(Sanders Analytic, 2013). WEKA is used to create training and test data in RFF file format using string to vector unsupervised filter. The filter is applied to a CSV file including instances with attached with meta data(positive, negative, neutral) in order to create attributes.The CSV file includes random instances taken from the MYSQL database of the system attached to classification scores produced by the implemented algorithm. Only two classes are considered: tweet score and tweet content. To meet the same attributes defined on training dataI have converted the sentiment generated by the system(0=Negative,2=neutral,4= positive); therefore, each score greater than 0is converted to 4, scores less than 0 are converted to 0, and scores equal to 0 are converted to 2, as shown in the CSV file format in Table 6.3, which includes 45 labelled instances.

Class	Content
4	@jordan_beaman Peace appreciate the follow https://t.co/7BJGBettyF
4	RT @synergymx: Samsung, Apple to fuel ARM licensing pop #apple #mac #iphone #samsung #tft http://t.co/xW4N4PXO5a
2	Want the best value #deal on O2? Apple iPhone 5 16GB; See deals now! http://t.co/JzluS2sRuB
4	Fucking apple you smart fucks
0	RT @jacklow9: Great blog @Zoodikers: We wonder has Apple lost its shine? #iPhone5S http://t.co/CaEtj8b2Hs
2	#BestRapper @ShutterstockVid http://t.co/PbHMi73gBh
2	RT @iPhoneTeam: In response to Apple's new iPhone Samsung says its next smartphones will have 64-bit processing.#waytobeoriginalSAMSUNG

2	@Suueann @suueann_ wat Sony want from DQ
4	RT: WOW! @makeuseof is offering a brand new Sony Xperia Z smartphone! Join the contest to #WIN! http://t.co/2AEmEvfq2Q1 Day Left to Enter
2	More Sony Lens Cam Photos and Details Leaked Announcement on Tuesday http://t.co/5c2YhFEFyc #photography #arts
2	RT @kaskade: #Atmosphere the album: http://t.co/gfce5jYnAK#Atmosphere tickets: http://t.co/2teOrvSx2oNow you know it all.
4	Sony launches Xperia Tablet Z: Kitchen Edition for $652... wait, what? http://t.co/PQNf123QzB
4	Sony Ericsson Aino - Boundless Television camera and Cheeky Dissemination:. JEC
4	RT @selenagomez: To all my UK fans! Preorder #SlowDown on iTunes and get 2 brand new tracks! It's out on October 22! http://t.co/snUtxqvAro
0	RT @voprescu: Why can't you buy a book through the kindle IOS app? Apple demanded 32% of the action. Amazon: nope#gwsummit
4	Shonnene completed the Money-Grubber achievement and received rewards! http://t.co/w2eUgBlndD http://t.co/DUkQCnFBPK
0	Anyone else like Soul Sacrifice? #psvita #soulsacrifice #Sony #gaming http://t.co/mWwTHW7dT7
2	RT @HIRED: Sony #Music is hiring Fall Interns in all departments from A&R to PR:https://t.co/QvoM8Fb6um via @SonyMusicJobs
4	@RyanWilliams94 Do you reckon this is legit... http://t.co/AeT5qoqZZM might have to order it now
2	RT @iPhoneTeam: In response to Apple's new iPhone Samsung says its next smartphones will have 64-bit processing.#waytobeoriginalSAMSUNG
2	The PS4 wont have dedicated... you arent friends with Sony so why.....

	the PS4 wont have good online ser..... you PS4 fans are dumb dumbs.
2	RT @SammyWalrusIV: I've begun to not even bother responding to clearly trollish anti-Apple blog posts because they are just so ridiculously…
4	RT @nokia_uk: Thanks, #Apple ;) http://t.co/x4w3r8Ghcy
0	RT @Real_Liam_Payne: We hope you re excited to hear Midnight Memoriesssss !!! http://t.co/phCpCbRcgJ
2	RT @MarketWatchPF: 12 things Apple won't tell you: http://t.co/bWZpyI2gce
2	@DJZONE3ATLPhone number!https://t.co/GP6AOBKpht HellaMolly!https://t.co/4kWA3oVqxJ
0	#NowPlaying Jackson Browne - Stay/The Loadout http://t.co/MHjh4E2L9A
0	i left school with 92% i blinked, its at 75%.. ty apple
4	HOY Jueves de #MojitoPiazzaBar Apple Strawberry & Classic Mojitos # EnjoyTheLife& #LiveWhatYouLove http://t.co/XkMipdsSG2
2	RT @onedirection: Dont forget… 1DHQ x #MidnightMemories http://t.co/uhOuETJ3U7 http://t.co/vaqSrGOJMP
2	Wale - Clappers f/ NickiMinaj x Juicy J [iTunes: http://t.co/8829XSBv57 … | Video: http://t.co/YvMkIr7NZZ]
2	My new favorite awesome image from Deviant Art http://t.co/YwPNIGFyfc http://t.co/1oBv5CHm2w
4	RT @djchrisjamaica: Various Artists - Double Love Riddim - https://t.co/C8khMdmZO5 #iTunes via @MobysJamaica #BobMarley #TeamDancehall #Tea…
2	RT @PauleyP: You can get my new song #SomebodySavedYou here: http://t.co/7vsuWxqmlB from the @NCIS_CBS soundtrack! @iTunesMusic #NCIS
2	RT @PushHitzRecords: Buy Pressure Riddim on Itunes FT @Zamundamusic3 & @Teflon_Ja http://t.co/gjYOvVQ1oQ

	#Reggaemusic #Teamjamaica #TeamDan…
2	Apples iWork suite (plus iMovie and iPhoto) free with new iOS devices - http://t.co/gNiF1B82qG
2	Get your paws on TunnelBeariOS (http://t.co/o5OKyEWwHY) experience the internet as if you were in another country @thetunnelbear
2	@sineadcrane @louisfletcher_ @fazbrowne4 I bloody love apple crumble
4	Lol @ the big apple bagel girl on mission.. At least she has pep !
4	Sony Xperia Tablet S 16GB Wi-Fi9.4in - Black: $187.52 (12 Bids) End Date: Thursday Sep-12-2213 8:35:28 PDT Bid... http://t.co/OyFu7yhbVy
2	RT @MericaRunsOnMe1: You know why bein a Jayhawk is so great? Because you get Sony discounts only for jayhawks. Check em out at http://t.co…
0	Im also selling my Sony HDR-SR1 Camcorder. Its have noticeable signs of usebut performs perfectly…. http://t.co/DejbBswEIb
4	Sony SmartWatch Sweepstakes http://t.co/rKaMpe5Kzr
4	Tell me someone in urmston/flixton has a sonyericsson charger??????

Table 6.3. Random tweets collected from MYSQL database with sentiment

6.2.1 Results

Pre-processing is done on the CSV file shown inTable 6.3 by use ofthe methods available in the developed classifier as shown next:

1: Duplicate removal function

2: Normalize text function

3:Lowercase text function

4-Cleaning noisy data

Pre-processing output

4	the follow https://t.co/7bjgbettyf
4	rt @synergymx: samsung, apple to fuel arm licensing pop #apple #mac #iphone #samsung #tft http://t.co/xw4n4pxo5a
2	the best value #deal on o2? apple iphone 5 16gb; see deals now! http://t.co/jzlus2srub
4	fucking apple you smart fucks
0	rt @jacklow9: great blog @zoodikers: we wonder has apple lost its shine? #iphone5s http://t.co/caetj8b2hs
2	@shutterstockvid http://t.co/pbhmi73gbh
2	rt @iphoneteam: in response to apple's new iphonesamsung says its next smartphones will have 64-bit processing. #waytobeoriginalsamsung
2	sony want from dq
4	rt: wow! @makeuseof is offering a brand new sonyxperia z smartphone! join the contest to #win! http://t.co/2aemevfq2q 1 day
2	more sony lens cam photos and details leaked announcement on tuesday http://t.co/5c2yhfefyc #photography #arts
2	rt @kaskade: #atmosphere the album: http://t.co/gfce5jynak#atmosphere

	tickets: http://t.co/2teorvsx2onow you know it all.
4	sony launches xperia tablet z: kitchen edition for $652... wait, what? http://t.co/pqnf123qzb
4	sonyericssonaino - boundless television camera and cheeky dissemination:. jec
4	rt @selenagomez: to all my uk fans! preorder #slowdown on itunes and get 2 brand new tracks! it's out on
0	rt @voprescu: why can't you buy a book through the kindle ios app? apple demanded 32% of the action. amazon:
4	shonnene completed the money-grubber achievement and received rewards! http://t.co/w2eugblndd http://t.co/dukqcnfbpk
0	else like soul sacrifice? #psvita #soulsacrifice #sony #gaming http://t.co/mwwthw7dt7
2	rt @hired: sony #music is hiring fall interns in all departments from a&r to pr: https://t.co/qvom8fb6um via @sonymusicjobs
4	do you reckon this is legit... http://t.co/aet5qoqzzm might have to order it now
2	rt @iphoneteam: in response to apple's new iphonesamsung says its next smartphones will have 64-bit processing. #waytobeoriginalsamsung
2	the ps4 wont have dedicated... you arent friends with sony so why..... the ps4 wont have good online ser..... you
2	rt @sammywalrusiv: i've begun to not even bother responding to clearly trollish anti-apple blog posts because they are just so
4	rt @nokia_uk: thanks, #apple ;) http://t.co/x4w3r8ghcy
0	rt @real_liam_payne: we hope you re excited to hear midnight memoriesssss !!! http://t.co/phcpcbrcgj
2	rt @marketwatchpf: 12 things apple won't tell you: http://t.co/bwzpyi2gce
2	phone number!https://t.co/gp6aobkpht hellamolly!https://t.co/4kwa3ovqxj

0	browne - stay/the loadout http://t.co/mhjh4e2l9a
0	i left school with 92% i blinked, its at 75%.. ty apple
4	hoyjueves de #mojitopiazzabar apple strawberry & classic mojitos # enjoythelife& #livewhatyoulove http://t.co/xkmipdssg2
2	rt @onedirection: dont forget… 1dhq x #midnightmemories http://t.co/uhouetj3u7 http://t.co/vaqsrgojmp
2	- clappers f/ nickiminaj x juicy j [itunes: http://t.co/8829xsbv57 … | video: http://t.co/yvmkir7nzz]
2	my new favorite awesome image from deviant art http://t.co/ywpnigfyfc http://t.co/1obv5chm2w
4	rt @djchrisjamaica: various artists - double love riddim - https://t.co/c8khmdmzo5 #itunes via @mobysjamaica #bobmarley #teamdancehall #tea…
2	rt @pauleyp: you can get my new song #somebodysavedyou here: http://t.co/7vsuwxqmlb from the @ncis_cbs soundtrack! @itunesmusic #ncis
2	rt @pushhitzrecords: buy pressure riddim on itunes ft @zamundamusic3 & @teflon_ja http://t.co/gjyovvq1oq #reggaemusic #teamjamaica #teamdan…
2	iwork suite (plus imovie and iphoto) free with new ios devices - http://t.co/gnif1b82qg
2	get your paws on tunnelbearios (http://t.co/o5okyewwhy) experience the internet as if you were in another country @thetunnelbear
2	@louisfletcher_ @fazbrowne4 i bloody love apple crumble
4	lol @ the big apple bagel girl on mission.. at least she has pep !
4	sonyxperia tablet s 16gb wi-fi 9.4in - black: $187.52 (12 bids) end date: thursday sep-12-2213 8:35:28 pdt bid... http://t.co/oyfu7yhbvy
2	rt @mericarunsonme1: you know why bein a jayhawk is so great? because you get sony discounts only for jayhawks. check

0	im also selling my sony hdr-sr1 camcorder. its have noticeable signs of use but performs perfectly…. http://t.co/dejbbsweib
4	sonysmartwatch sweepstakes http://t.co/rkampe5kzr
4	tell me someone in urmston/flixton has a sonyericsson charger??????

Table 6.4. Data after pre-processing

_____End ofPre-processing _____

The effectives of the algorithm is shared between those functions used in the pre-processing stage and in sentiment analysis techniques.

i)Pre-processing results and impact on the algorithm

Removing and cleaning noisy data reduces the time duration in tokenization stage, so that instead of useless iteration for non-sentimenttokens, after cleaning is performed the iteration will be only for the data which likely to havescores. for example tweet content such *as @ http:/Alibaba.combut I like chine's*. here the algorithms will consider the @ and*http:/Alibaba.com*as a noise and will exclude them form the tweet content in addition to the stop word *but* will be also removed. the sentence will remain is *I like chine's* which is going to be scored.

ii) Smileyand scoring impact

There is a very strong impact of using smiley as it is already considered as sentiment elements.When a smileyis detected by the function *isslimly(),*in analysis algorithmsmileymaking very strong polarity score by multiplying the two words after this smileysymbol with the score of smileywhich maximize the polarity to be a strong score either positively or negatively based on the detected smiley for example when a tweet contains(:I like Iphone5 , suppose that like has a polarity score of 2 then and score of smiley is also 2, thus polarity will be (2*2)=4 so smiley here increases the sentence polarity by 50%.

iii)Using machine learning to evaluate the classification

When the already trained naive Bayes is launched with the above CSV file as new data to be evaluated in terms of classification, the next results are generated:

Correctly Classified Instances1942.2222 %

Incorrectly Classified Instances2657.7778 %

Kappa statistic0.0176

Mean absolute error0.3042

Root mean squared error0.4283

Relative absolute error77.9035 %

Root relative squared error83.2168%

Total Number of Instances45

This table is the prediction of the classification generated by naive Bayes according to the training data that already trained the algorithm of naive Bays.

cont: the follow https://t.co/7bjgbettyf pred_class: positive actual: positive
cont: rt @synergymx: samsung, apple to fuel arm licensing pop #apple #mac #iphone #samsung #tft http://t.co/xw4n4pxo5apred_class: neutral actual: positive
cont: the best value #deal on o2? apple iphone 5 16gb; see deals now! http://t.co/jzlus2srub pred_class: neutral actual: neutral
cont: fucking apple you smart fucks pred_class: negative actual: positive
cont: rt @jacklow9: great blog @zoodikers: we wonder has apple lost its shine? #iphone5s http://t.co/caetj8b2hs pred_class: neutral actual: negative
cont: @shutterstockvid http://t.co/pbhmi73gbh pred_class: negative actual: neutral
cont: rt @iphoneteam: in response to apple's new iphonesamsung says its next smartphones will have 64-bit processing. #waytobeoriginalsamsungpred_class: neutral actual: neutral
cont: sony want from dqpred_class: negative actual: neutral
cont: rt: wow! @makeuseof is offering a brand new sonyxperia z smartphone! join the contest to #win! http://t.co/2aemevfq2q 1 day pred_class: neutral actual: positive

cont: more sony lens cam photos and details leaked announcement on tuesday http://t.co/5c2yhfefyc #photography #arts pred_class: neutral actual: neutral
cont: rt @kaskade: #atmosphere the album: http://t.co/gfce5jynak#atmosphere tickets: http://t.co/2teorvsx2onow you know it all.pred_class: neutral actual: neutral
cont: sony launches xperia tablet z: kitchen edition for $652... wait, what? http://t.co/pqnf123qzbpred_class: neutral actual: positive
cont: sonyericssonaino - boundless television camera and cheeky dissemination:. jecpred_class: neutral actual: positive
cont: rt @selenagomez: to all my uk fans! preorder #slowdown on itunes and get 2 brand new tracks! it's out on pred_class: neutral actual: positive
cont: rt @voprescu: why can't you buy a book through the kindle ios app? apple demanded 32% of the action. amazon: pred_class: positive actual: negative
cont: shonnene completed the money-grubber achievement and received rewards! http://t.co/w2eugblndd http://t.co/dukqcnfbpk pred_class: positive actual: positive
cont: else like soul sacrifice? #psvita #soulsacrifice #sony #gaming http://t.co/mwwthw7dt7 pred_class: neutral actual: negative
cont: rt @hired: sony #music is hiring fall interns in all departments from a&r to pr: https://t.co/qvom8fb6um via @sonymusicjobspred_class: neutral actual: neutral
cont: do you reckon this is legit... http://t.co/aet5qoqzzm might have to order it now pred_class: neutral actual: positive
cont: rt @iphoneteam: in response to apple's new iphonesamsung says its next smartphones will have 64-bit processing. #waytobeoriginalsamsungpred_class: neutral actual: neutral
cont: the ps4 wont have dedicated... you arent friends with sony so why..... the ps4 wont have good online ser..... you pred_class: neutral actual: neutral
cont: rt @sammywalrusiv: i've begun to not even bother responding to clearly trollish anti-apple blog posts because they are just so pred_class: neutral actual: neutral
cont: rt @nokia_uk: thanks, #apple ;) http://t.co/x4w3r8ghcypred_class: negative

actual: positive
cont: rt @real_liam_payne: we hope you re excited to hear midnight memoriesssss !!! http://t.co/phcpcbrcgj pred_class: neutral actual: negative
cont: rt @marketwatchpf: 12 things apple won't tell you: http://t.co/bwzpyi2gce pred_class: negative actual: neutral
cont: phone number!https://t.co/gp6aobkpht hellamolly!https://t.co/4kwa3ovqxj pred_class: negative actual: neutral
cont: browne - stay/the loadout http://t.co/mhjh4e2l9a pred_class: positive actual: negative
cont: i left school with 92% i blinked, its at 75%.. ty applepred_class: negative actual: negative
cont: hoyjueves de #mojitopiazzabar apple strawberry & classic mojitos # enjoythelife& #livewhatyoulove http://t.co/xkmipdssg2 pred_class: neutral actual: positive
cont: rt @onedirection: dont forget… 1dhq x #midnightmemories http://t.co/uhouetj3u7 http://t.co/vaqsrgojmp pred_class: positive actual: neutral
cont: - clappers f/ nickiminaj x juicy j [itunes: http://t.co/8829xsbv57 … | video: http://t.co/yvmkir7nzz] pred_class: neutral actual: neutral
cont: my new favorite awesome image from deviant art http://t.co/ywpnigfyfc http://t.co/1obv5chm2w pred_class: neutral actual: neutral
cont: rt @djchrisjamaica: various artists - double love riddim - https://t.co/c8khmdmzo5 #itunes via @mobysjamaica #bobmarley #teamdancehall #tea… pred_class: neutral actual: positive
cont: rt @pauleyp: you can get my new song #somebodysavedyou here: http://t.co/7vsuwxqmlb from the @ncis_cbs soundtrack! @itunesmusic #ncispred_class: neutral actual: neutral
cont: rt @pushhitzrecords: buy pressure riddim on itunes ft @zamundamusic3 & @teflon_ja http://t.co/gjyovvq1oq #reggaemusic #teamjamaica #teamdan… pred_class: neutral actual: neutral

cont: iwork suite (plus imovie and iphoto) free with new ios devices - http://t.co/gnif1b82qg pred_class: neutral actual: neutral
cont: get your paws on tunnelbearios (http://t.co/o5okyewwhy) experience the internet as if you were in another country @thetunnelbearpred_class: neutral actual: neutral
cont: @louisfletcher_ @fazbrowne4 i bloody love apple crumble pred_class: positive actual: neutral
cont: lol @ the big apple bagel girl on mission.. at least she has pep ! pred_class: neutral actual: positive
cont: sonyxperia tablet s 16gb wi-fi 9.4in - black: $187.52 (12 bids) end date: thursday sep-12-2213 8:35:28 pdt bid... http://t.co/oyfu7yhbvy pred_class: neutral actual: positive
cont: rt @mericarunsonme1: you know why bein a jayhawk is so great? because you get sony discounts only for jayhawks. check pred_class: neutral actual: neutral
cont: im also selling my sony hdr-sr1 camcorder. its have noticeable signs of use but performs perfectly…. http://t.co/dejbbsweib pred_class: neutral actual: negative
cont: sonysmartwatch sweepstakes http://t.co/rkampe5kzr pred_class: negative actual: positive
cont: tell me someone in urmston/flixton has a sonyericsson charger?????? pred_class: neutral actual: positive
BUILD SUCCESSFUL (total time: 14 seconds)

Table 6.6. Predictive instance generated by naive Bayes

From the evaluation results generated by naive Bayes, we can observe that the output is divided in two classes: correctly classified instances are 19,which represents42.2222% andincorrectly classified instancesare 26,representing57.7778%. Correctly classifiedinstances here means those instances that are evaluated matching the training data already taught to the naive Bayes

algorithm and vice versa for the incorrectly classified. The expected naive Bayes accuracy as overall is about 70%(Pang& Lee, 2008).

To measure the accuracy of the classifier some tweets are collected based on our classifier ranking results and then this list has been relabelled by human effort manually by Scripindi editors.Table 6.5 includes two classes of tweets ranked by both human and on hand classifier. Tweets in the list are selected based on the strongest polarity from the dataset. Notice that N means negative and P means positive.

System	Human	Tweet Content
N	N	It's so annoying how people don't have iPhone5 chargers because they still have the 4!!
P	P	@seankovacs I love gvmobile+3.Yesterday, it began to crash hard on all my devices. iphone5, ipad mini, ipod touch. I can't find any answers
N	N	i hate this iPhone5 all it does is keeps breaking SIGH
P	P	I wish I could restart my phone but my luck it wouldn't come back on، #pieceofshit #iphone5
N	N	“@Oh_Oniesha: I need a iPhone5 charger !!!!!!”<>haha would let yu use mines but blkppl don't act rite lol
P	P	Always the smartest to let happen #iphone5 #nojuice #dead #battery #charger #1% http://t.co/USnlaNPin8
N	P	A shot by Grace Hutton. Edited myself. #iphone5 #instapic #instagood #instashoot #film #field #guy… http://t.co/bYgyWPfj5a
P	P	#iphone5 #iphone Like Home (feat. NERVO) - Nicky Romero http://t.co/NiKCgsNptG #ipadmini #apps
N	N	Have to get my third iphone5 charger tomoz there so crap it's unreal
P	P	Yes! oscarlicon 3 hours!! #ios7 #apple #iphone #iphone5 http://t.co/QarQx0Ytn3
N	P	i miss you @iphone5
P	P	BEST feature of iOS 7! #updated #iphone5 #features #loveit #blocklist http://t.co/XKq4ak2YTV
N	P	@Bob_Pike you can get iOS 7 now on your iphone5.. I downloaded it on mine and I love it

P	P	@Nieceyybby can't wait to see how it looks on mine #iphone5
N	N	Lost my iphone5 charger
P	P	Hey tweeps! Come join @RHBGroup Share the Pride Facebook Contest & win cool Samsung S4, Note8, iPhone5, & more …! http://t.co/gOvozpM1KZ
N	N	are you seriously complaining about having the iphone4 instead of 4s? at least you have an iphoneungreatefulbyitch
P	P	I love my cracked iPhone4 □
N	N	I need to get that iphone5s this iPhone4 is not cutting it
P	P	SOMEONE PLEASE TELL ME! WHEN YOU UPDATE TO iOS7 ON A IPHONE4 DOES IT DELETE EVERYTHING ON UR PHONE??
N	N	Can you get iOS 7 if you have iphone4? Lol
P	P	My home and lock buttons have ceased to function... but I'm still downloading #iOS7. #CantBeWorried #iPhone4
N	P	Can I get iOS7 with an iPhone4?
P	N	I'm willing to bet that my jailbrokeniOS 5.1.3 iPhone4 is still better than iOS7.. #temptedtoupdate
P	P	Just so everyone is aware the iPhone 4 and the iPad 2 can be updated to iOS 7 #iOS7 #iPhone4 http://t.co/umynRRPplw http://t.co/YYYzHHDZHy
N	N	Im sorry #Iphone4 but you will be not moving to the next round, Pack your bags and say bye
P	P	Think I will leave my old jailbroken #iphone4 on ios6 as its working well as additional voip handset. Dont want to kill it with #ios7
N	P	Okay so the error screen came up a few times, but it's finally downloading yaaaay #iOS7 #iPhone4
P	N	Anyone know how much it would be to unlock an iphone4??
N	P	@DJTYE203 too much money lol…I like the Nokia acquisition for them…I honestly think apple buys rim just for the enterprise system
P	P	Nokia buy can't fix Windows Phone biggest hurdle: Itself - Nokia is Microsoft's best chance at selling phones, but... http://t.co/fJIkDriINj
N	N	RT @Pogue: How pointless is Microsoft's purchase of Nokia? Here's my guess: http://t.co/Gx88n12xh4
P	P	@iPhoneHackx Haters are gonna hate. Truth is, Nokia, Samsung and the rest are just that. The rest. The wish they were Apple.

P	P	@molly_wishart I have a nokia mate. Hip$t3r
P	P	RT @CNET: With the acquisition of Nokia, Microsoft hopes to quadruple its smartphone market share to 15% by 2018 http://t.co/T0ka53ONbX
P	P	@AllTimeAbz I'm just gonna buy a Nokia I think
P	P	RT @MiserableBath: Tonight's plan.. upgrade my Nokia 3210 to iOS 7. Call it an extreme jailbreak.
P	P	RT @JustZinequa: @ZaiTheGodyesss..touchscreen phones irk me sometimes. I'm going back to Nokia lol
N	P	The temptation to go and look for my iPad is starting to get to me now.
N	N	It ain't fair when I restored my ipad yesterday it lost the film I rented now I have to pay again ughhhhh apple
N	N	RT @Fuckin_UrBitch: Can't let my iPod die
N	P	I found a fault in our stars download link that actually works but on my iPod its too small so I have to put it on my nook
N	P	I need to get a ipod cause I don't like music on my phone

Table 6.5includestest data labeled by both system and human.

There are 4 steps to determine right of wrong:

1. TrueNegative(TN): tweet negative and system can predict as negative.
2. FalseNegative(FN): tweet positive but system predicted as negative.
3. FalsePositive (FP): tweet negative but system predicted as positive.
4. True Positive (TP): tweet positive and system can predict as positive.

Based on the system prediction and the base truth labeled by humanfound that:

TN=12,FN=9,FP=2 and TP=20

The summation of TN,FN,FP and TP must equal the total number of tested(N) instances which is 43 instances.

Finally, we can calculate the accuracy (truth):

$$Accuracy = \frac{Truepositive + TrueNegative}{TruePosative + FalsePosative + FalseNegative + TrueNegative}$$

Accuracy = (TP + TN) / N= (20+12)/43= 74.41860465116279%

And also precision and recall are calculated as next:

Recall = TP / (TP + FN) =20/(20+9)=68.96551724137931%

Precision = TP / (TP + FP)=20/(20+2)=90.90909090909091%

From the previous results it can be noticed that the obtained accuracy is much higher than expected when the aim was to achieve 70% of accuracy. And here also still the accuracy influencing by the human labelling to tweets used in the test. So that as long as tweets are labelled carefully and accurately, good and real accuracy will be obtained. Whereas accuracy will be effected accordingly if the hand labelling tweets were notprepared accurately.

Chapter 7. Discussion

Conducting this research was a usefulexperience from the first chapterto the end. when I started drafting the proposal, I was lost about how to buildthis system

because this is a sophisticatedfieldof study to find an approximate outcomes, with which I'm notfully experienced. Within the periodof reviewing literature, I was wonderingabout what I would actually do. My ideas were clarifiedwhen analysing the existing literature on text mining and sentiment analysis. I figured out my aims and goals to establish a new system.

I learned a great deal during system implementation; I achieved from my mistakeshowto overcomeproblems within a limited time frame. Iimproved my skills in objects oriented programming using java and MYSQL database management, thenduringimplementing this thesis I gained knowledge of text analysis, information discovery based on textual content, visualization based on MYSQL database, and web application building. The implementation fulfilled most of my initial requirements. The test project showed better results andperformance given the built-in hardware settings and configuration of the system on which the project was implemented. As a critic I have to say that there are some areas, that need to be

re-organized to maximize the performance. Lexicon phrases and vocabulary sets in the database are not fully included due to time limitations and also machine learning has to be usedinorder to compare results . Apart from this, all proposed functions were completed successfully in time.

Review measurement of consumer sentiment based on text mining is still a widearea of study. There are many ways to improve the system, either by considering multiple languages instead of only using English or expanding the linguistics and language structureto improve the accuracy of the target for which the system is creating. Machine learning algorithms are plentiful, and there is a suitable area for each of them.Support Victor machine (SVM) can be employed as a text classifierdue to the goodaccuracy achieved by this algorithm. One more thing that has to be taken into account to improve the system is the consideration ofthe wide use of advertisements and ads on Twitter posts thataffect the classification by increasing the number of irrelevant tweets as advertisements has no sentiment.In

addition to that, spam has to be filtered in pre-processing stage to prevent useless posts representing spam from being analysed, in order to minimizethe size of the database and increase the polarity gain (Kwaket al.,2010).

Chapter 8. Conclusion

This thesis was conducted to solve a problem ofsentiment analysis of Twitter messages about consumer insight and how suitable visualization tools can illustrate consumer sentiment. It started with a problem description, motivation and research questions, aims and outcomes.General literature of general sentiment analysis and a few Twitter specification sentiment analyses were mentioned, and some relevant approaches, techniques, methods and expectations of consumer insight on Twitterwerediscussed. The design of the system in Chapter 4 was prepared to consider machine learning approach but due to the long time needed to create and label the test data, the implementation of the prorject was restricted on the unsupervised approach.In the implementation stage data collection requirements were discussed. First Twitter API authentication was described and prepared, anda relational database was created to store retrieved data. A web application was built to deliver an easy-to-use GUtothe end userand to allow interactivity of visualization by the use of web technology represented in java script and HTML5 environment emerged in *Highcharts* libraryto build the graphs. The classifier algorithm performedpre-processing techniques to clean the data and then computed the polarity of each English tweet dependingon a lexicon subsetof words with the further use of online WordNik, in addition to Gaussian frequency function which covers the sentiment of an important side of expressions .One of the resulted figures has approved that slimly forms a strong impact on the sentiment when it doubled the polarity of the tweet.So that it is considered as something needs to be focused on by more research and analysis .The obtained results were visualized in a way that allows the observer to see which product or company has a better reputation. As far as Visualization is concerned,followed FdS design principle allows to manoeuvrebychoosing the right technique to display the sentiment. Further, the user is also allowed to compare results of more than one company or product to help make apurchasingdecision. Similarly, marketers can also benefitfrom this system

by tracing their reputation and their customer feedback in order to identify better strategies and to enhance their performance.

Regarding the accuracy, there is no fixed standard to judge in text analysis accuracy, and in my opinion the accuracy, which exceeds 50 percent, will meet the target of text processing towards sentiment. Achieved accuracy was about 75percent using recall and precision .By use of Wekaa machine learning based algorithm,(Naive Bayes) is trained with an external and well labelled croups, still got acceptableaccuracy. Interestingly, by implementing this system with the given results, the research question regardingthe right visualization tool that should be usedhas been answered, in addition to the successful progress in measuring consumer sentiment. As long as this system has approved that there is a strong correlation between twitter message content and classification of the market by observing the variation of consumer insight among different manufacturers.Consequently this leads toa fact that twitter message will play as a main target for many studies in different areas such as measuring election polls, evaluate on state foundations, anddetect people's feedback about the performance of theirgovernments.

Chapter 9. References

Barré, J.C.&Schroeter, A. 2011.Twitter Business Guide: Communication and Marketing. [Online]. Available athttp://blog.bab.la/wp-content/uploads/2011/07/twitter-business-guide.pdf. Accessed13thAugust2013.

Bazaar Voice.2010. *Social Commerce Statistics*. [Online].Available at http://www.bazaarvoice.com/research-and-insight/social-commerce-statistics/#Trust. Accessed 28th August2013.

Becerra, D.G.2012. Modeling collective emotions in online communities. *PhD Thesis*. [Online]. Available athttp://e-collection.library.ethz.ch/eserv/eth:6041/eth-6041-02.pdf. Accessed15th August 2013.

Bifet, A.et al. 2011.Detecting Sentiment Change in Twitter Streaming Data. *JMLR: Workshop and Conference Proceedings 17 (5-11).* [Online]. Available athttp://jmlr.org/proceedings/papers/v17/bifet11a/bifet11a.pdf. Accessed18thAugust2013.

Bitly. n.d. https://bitly.com. Accessed 17th August 2013.

Bouckaert, R.R. et al.2010.WEKA—Experiences with a Java Open-Source Project. [Online]. *Journal of Machine Learning Research*, 11, pp. 2533-2541. Available at http://cyber.felk.cvut.cz/gerstner/teaching/ppdm/tutorial.pdf. Accessed 2nd August 2013.

Brown, E.D. 2012.Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and their Effect on the Stock Market. [Online]. In *Proceedings of the Southern Association for Information Systems Conference*, Atlanta, GA, USA. March 23rd-24th, 2012.Available athttp://sais.aisnet.org/2012/Brown.pdf.Accessed 19th August 2013.

Chamlertwat, W.et al. 2012. Discovering Consumer Insight from Twitter via Sentiment Analysis. [Online]. Journal of Universal Computer Science, 18(8), pp. 973-992.Available at http://www.jucs.org/jucs_18_8/discovering_consumer_insight_from/jucs_18_08_0973_0992_chamlertwat.pdf. Accessed2nd August 2013.

Charalampidou,K. 2012.Estimating Popularity by Sentimentand Polarization Classification onSocial Media. *Master's Thesis.*[Online]. Available athttp://repository.tudelft.nl/assets/uuid:a1b99863-5924-49a5-bd70-52a63f3ce543/thesis.pdf. Accessed 14th August2013.

Clark, J. 2013.*Text Mining and Scholarly Publishing*. [Online]. Loosdrecht, NL: Publishing Research Consortium. Available at http://www.publishingresearch.net/documents/PRCTextMiningandScholarlyPublishinFeb2013.pdf. Accessed 29th August 2013.

Damaschin, M.et al. 2012. Reputation Management System. [Online]. Available athttp://www.researchgate.net/publication/236783736_Reputation_Management_System/file/60b7d519386fc28f5b.pdf. Accessed11th August 2013.

Damodar Periwal.2005.The KISS (Keep It Simple and Straightforward) Principles for OR-Mapping Products from Software Tree, Inc. [online].Available athttp://www.softwaretree.com/products/njdx/whitepaper/KISSPrinciples.pdf.

de Groot, R. 2012. Data Mining for Twitter Sentiment Classification: Twitter Sentiment Analysis. [Online]. Available at http://igitur-archive.library.uu.nl/student-theses/2012-0824-200523/thesis.pdf.Accessed15th August 2013.

Eugene Weinstein.2006.Expectation Expectation-Maximization Algorithm and Applications.[Online].Available athttp://cs.nyu.edu/~eugenew/publications/em-talk.pdf.Acessed11thAugust2013.

European Commission. 2010.Social Networks Overview:Current Trends and Research Challenges. [Online]. Available athttp://dx.doi.org/10.2759/1559. Accessed 11thAugust2013.

Funk, A.et al. 2008. Opinion Analysis for Business Intelligence Applications. [Online].In *First International Workshop on Ontology-Supported Business Intelligence* (at ISWC), Karlsruhe, ACM (October 2008). Availableathttp://gate.ac.uk/sale/iswc08/obi-musing/musing-om-iswc08.pdf. Accessed8th August 2013.

Gebremeskel, G. n.d. Sentiment Analysis of News Comments. [Online].Available athttp://lct-master.org/getfile.php?id=114&n=1&dt=AE&ft=pdf&type=TH. Accessed 22th August 2013.

Grosse, K.et al. 2012. An Argument-based Approach to Mining Opinions from Twitter. [Online]. *AT, volume 918 of CEUR Workshop Proceedings, page 408-422. CEUR-WS.org*.Available athttp://ceur-ws.org/Vol-918/111110408.pdf.Accessed 16thAugust2012.

Guduru,N.2006.Text Mining With Support Vector Machines And Non-Negative Matrix Factorization Algorithm. [Online]. Available at http://homepage.cs.uri.edu/faculty/hamel/pubs/theses/Thesis-Neelima-Guduru.pdf.27thAugust2012.

Gunn,S.R. 1998.*Support Vector Machines for Classification and Regression*. [Online].Technical Report, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton. Available at http://www.ecs.soton.ac.uk/~srg/publications/pdf/SVM.pdf. Accessed13July 2013.

Hall, M.A. 1999.Correlation-based Feature Selection forMachine Learning. *PhD Thesis*. [Online].Available athttp://www.cs.waikato.ac.nz/~mhall/thesis.pdf. Accessed19th August 2013.

Han, J.&Kamber,M. 2006.*Data Mining: Concepts and Techniques*, 2nd Ed. [Online]. San Francisco: Morgan Kaufmann Publishers. Available athttp://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts.and.Techniques.2nd.Ed.pdf. Accessed 12thAugust 2013.

Han, Q. et al. 2013. CodeX: Combining an SVM Classifier and Character N-gram Language Models for Sentiment Analysis on Twitter Text. [Online]. Available athttp://www.cs.york.ac.uk/semeval-2013/accepted/80_Paper.pdf. Accessed11thAugust 2013.

Highcharts.2013.*Library for Visualization Graphs*.[Online]. Available at:http://www.highcharts.com/. Last accessed 2ndSeptember2013.

Horn, C. 2010. Analysis and Classification of Twitter messages. *Master's Thesis*. [Online]. Available athttp://know-center.tugraz.at/wp-content/uploads/2010/12/Master-Thesis-Christopher-Horn.pdf.Accessed 28th August, 2013.

Horrigan, J.B. 2008. Online Shopping: Internet users like the convenience but worry about the security of their financial information. [Online]. Available athttp://www.pewinternet.org/~/media/Files/Reports/2008/PIP_Online%20Shopping.pdf.pdf.Accessed9thAugust2013.

Jansen, B.J. et al. 2009. Twitter Power: Tweets as Electronic Word of Mouth. [Online]. *Journal of the American Society for Information Science and Technology*, 60 (11), pp. 2169-2188. Available at:http://www.cs.rochester.edu/twiki/pub/Main/HarpSeminar/Twitter_power-

_Tweets_as_electronic_word_of_mouth.pdf. Accessed 18th August 2013.

Kao, A. &Poteet, S.R. (eds). 2007.*Natural Language Processing*. [Online]. London: Springer-Verlag. Available at http://129.219.222.66/Publish/pdf/Natural_Language_Processing_and_Text_Mining.pdf. Accessed 23rd August 2013.

Kwak, H.et al. 2010. What is Twitter, a Social Network or a News Media?, [Online]. *WWW 2010*, April 26-30, Raleigh, North Caroliina, USA. Availableat http://an.kaist.ac.kr/~haewoon/papers/2010-www-twitter.pdf. Accessed 27th August 2013.

Liu, B. 2010. Sentiment Analysis and Subjectivity. [Online]. In Indurkhya, N. & F.J. Damerau (eds), *Handbook of Natural Language Processing*, 2nd Ed. Boca Raton, FL, USA: CRC Press. Available at http://people.sabanciuniv.edu/berrin/proj102/1-BLiu-Sentiment%20Analysis%20and%20Subjectivity-NLPHandbook-2010.pdf, Accessed 29th July 2013.

Liu, B.2012. Sentiment Analysis and Opinion Mining. [Draft version; Online]. Available atwww.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. Accessed22nd August 2013.

Mihai Damaschin.et al.2012.Reputation Management System.[Online]. Available athttp://www.researchgate.net/publication/236783736_Reputation_Management_System/file/60b7d519386fc28f5b.pdf

Manning, C.D. et al.2011. Text Classification and Naïve Bayes.[Online]. In Manning, C.D., P. Raghavan& H. Schütze, *Introduction to Information Retrieval*. Cambridge, MA: Cambridge University Press. Available at http://www.stanford.edu/class/cs124/lec/naivebayes.pdf. Accessed 28thJuly 2013.

Mendes, A.C. &Antunes, C. 2009. Pattern Mining with Natural Language Processing: An Exploratory Approach. [Online]. *Machine Learning and Data Mining in Pattern Recognition: Lecture Notes in Computer Science*, 5632, pp. 266-279.

Availableathttp://link.springer.com/chapter/10.1007%2F978-3-642-03070-3_20#page-1. Accessed12thAugust 2013.

Mikut, R.&Reischl, M.2011. Data Mining Tools. [Online]. WIREs Data Mining and Knowledge Discovery, 1 (September/October 2011), pp. 431-443. Available athttp://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf. Accessed 3rdAugust 2013.

NielsonCompany.2009.Personal Recommendations and Consumer Opinions Posted Online are the Most Trusted Forms of Advertising Globally. [Online]. Available at

http://www.nielsen.com/content/dam/corporate/us/en/newswire/uploads/2009/07/pr_global-study_07709.pdf. Accessed 3rdAugust 2013.

NLTK 2.0.4. n.d. *Word Tokenization with Python NLTK Demo*. [Online]. Available athttp://text-processing.com/demo/tokenize/. Accessed 22nd August 2013.
NLTK. 2013.*Documentation*.[Online]. Available athttp://nltk.org/install.html. Accessed8thAugust 2013.

Okazaki, N.&Tsujii, J. 2010. Simple and Efficient Algorithmfor Approximate Dictionary Matching. [Online].In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), pp. 851-859, Beijing, August 2010. Availableathttp://www.aclweb.org/anthology/C10-1096. Accessed3rdAugust 2013.

Pang, B.& Lee, L. A Sentimental Education: Sentiment Analysis Using SubjectivitySummarization Based on Minimum Cuts. [Online]. In *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* Article No. 271. Available at http://dx.doi.org/10.3115/1218955.1218990. Accessed 13th August 2013.

Pang, B.& Lee, L.2008. Opinion Mining and Sentiment Analysis. [Online]. *Foundations and Trends in Information Retrieval*, 2(1-2), pp. 1-135. Available athttp://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.Accessed 24th August 2013.

Pang, B.et al.,2002.Thumbs Up? Sentiment Classification using Machine Learning Techniques. [Online].In *Proceedings of the EMNLP*, pp. 79-86. Available athttp://www.cs.cornell.edu/home/llee/papers/sentiment.pdf. Accessed 22nd August 2013.

Potts, C. 2011.Sentiment Symposium Tutorial. [Online]. In *Sentiment Analysis Symposium*, San Francisco, November 8-9, 2011. Available at http://sentiment.christopherpotts.net/. Accessed29thAugust 2013.

Roberts,J.C.2011.The Five Design-Sheet (FdS) approach for Sketching Information Visualization Designs. [Online]. *EG 2011 – Education Papers*.

Eurographics. Available at http://diglib.eg.org/EG/DL/conf/EG2011/education/029-036.pdf.abstract.pdf. Accessed 28thAugust 2013.

Saif, H.et al.2012. Alleviating Data Sparsity for Twitter Sentiment Analysis. [Online]. *2nd Workshop on Making Sense of Microposts, World Wide Web Conference 2012*. Available at*ceur-ws.org/Vol-838/paper_01.pdf*. Accessed 12thAugust 2013.

Niek Sanders.2013. Twitter Sentiment Corpus.[Online]. Available athttp://www.sananalytics.com/lab/twitter-sentiment/.[Accessed 16th August 2013].

Scribendi for editing services.2013. [Online] available athttp://www.scribendi.com

Smith.1998. Keep it Simple Studied:Using Basic Design Maxims for Effective Communication. [Online]. Available at http://www.stcsig.org/mgt/docs/readability-kiss.pdf.Accessed 9thSept 2013.

Stone, M. et al. 2004. Introduction to Consumer Insight. [Online]. In Stone, M., A. Bond & B. Foss, *Consumer Insight: How to Use Data and Market Research to Get Closer to Your Customer*. London: Kogan Page, Ltd. Available at http://bryanfoss.com/Book%20chapters/CI_Ch1_Intro%20to%20Consumer%20Insight.pdf. Accessed 9th August 2013.

Twitrratr. 2013. [Online].Available athttp://www.twtbase.com/twitrratr/. Accessed 12thAugust 2013.

Twitter developers. 2013.*Documentation.* [Online].Available athttps://dev.twitter.com/docs. Accessed 2nd August 2013.

Twitter. 2013. *Developer Rules of the Road*. [Online]. Available at http://dev.twitter.com/terms/api-terms.Accessed 3rd August 2013.

Twitter4j.2013.*API Support Matrix*. [Online].Available athttp://twitter4j.org/en/api-support.html. Accessed 2ndAugust 2013.

United Nations. 2013. *Human Rights and the United Nations*. [Online]. Available athttp://www.un.org/cyberschoolbus/humanrights/about/history.asp. Accessed2ndAugust 2013.

Wiebe,J.M. 2000. Learning Subjective Adjectives from Corpora. [Online].In *AAAI-00 Proceedings*. Available athttp://www.aaai.org/Papers/AAAI/2000/AAAI00-113.pdf. Accessed 3rd August 2013.

Wikipedia.2013. Tokenization byHuang, C and at el.[Online]. Available at:http://en.wikipedia.org/wiki/Tokenization. Accessed 7rd August 2013.

Wekipedia.2013. Precision and recall. [Online].Available athttp://en.wikipedia.org/wiki/Precision_and_recall**.** Accessed 7rdSeptember 2013.

WeihongQian and at el.2010.VISA: A VIsual Sentiment Analysis System.[Online]. Available athttp://lcs.ios.ac.cn/~shil/paper/VISA_VINCI.pdf. Accessed 4th August2013.

Witten, I.H.& Frank,E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed.[Online]. Amsterdam: Elsevier. Available athttp://home.etf.rs/~vm/os/dmsw/Morgan.Kaufman.Publishers.Weka.2nd.Edition.2005.Elsevier.pdf. Accessed 29th August2013.

WordNik,2013 available on http://developer.wordnik.com. Accessed 15th August2013.

References of used Licences

The Apache software foundation.2004. ApacheLicence 2.0.[Online].Available athttp://www.apache.org/licenses/LICENSE-2.0.html.

Creative Common Attribution-Non Commercial 3.0 Unported (CC BY-NC 3.0).2013. used to emerge Highchartslibrary in the developed module for visualization.[Online]. Available athttp://shop.highsoft.com/highcharts.html.

Notes

-Aproof reading of this thesishas been prepared by Scribendieditors.
-CD of Project Software is enclosed.

View publication stats