UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

SENTIMENTAL ANALYSIS IN CRIME DETECTION: A CASE STUDY OF KENYA LAW ENFORCEMENT AGENCIES

BY

KENEDY DENDE

P58/75853/2012

SUPERVISOR

DR.LAWRENCE MUCHEMI

NOVEMBER 2014

Submitted in partial fulfillment of the requirements of the degree of Master Of Science in Computer Science.

Declaration

This project, as presented in this report, is my original work and has not been presented for any other university award.

KENEDY ONYANGO DENDE Signed: …… Date ….…….

P58/75853/2012

This project has been submitted as part fulfillment of the requirement for the award of Masters of Science in Computer Science with my approval as the University Supervisor

DR.LAWRENCE MUCHEMI Signed: ……. Date: ……

School of Computing and Informatics of the University of Nairobi

Abstract

In this work, we attempt to resolve the problem of mining text and classifying them into positive, negative and neutral to detect crime by applying Sentiment analysis on people's opinions expressed on social media. We attempt to resolve this problem by first presenting the user with a summative view of the complete data set, summarized by a label or a score, and subsequently by segmenting the opinions/sentiments into three classes (affirmative, negative and impartial).

This project developed sentiment analyzer that is appended on the browser, it then analyses the posted keywords or tweets and uses language normalizers to identify with the Kenyan context. If the post or the tweet is from Kenyan context it is given a score of one if it is not directly from the Kenyan context it is given a score of 0.5 .The keyword is then labeled with the three outputs of negative, positive and neutral after which a training is done in several rounds to ensure accuracy.

Based on the above model a prototype was developed that was based on the Naïve Bayes in which test runs were conducted on balanced corpus 460 (keywords) and unbalanced corpus 860 (keywords).

The sent analyzer was able to classify the keywords into the three categories. To evaluate the prototype three accuracy standards were applied, these were precision, recall and accuracy, the results obtained from experiments with the classifier show that the classifier is capable of performing classification with an accuracy of 77.8% for sentiments obtained from Social Media. This is near human accuracy, as apparently people agree on sentiment only around 80% of the time. Most of the sentiments in this data are expressed partly in English, Swahili, thus formal language is scarcely used. We therefore conclude that the model of classification selected is ideal for the kind of data collected from social media on Kenyan opinions.

The findings of this research will be of great importance to the researchers by adding another perspective of Naïve Bayes in opinion mining as well as the law enforcement agencies in identifying negative opinions in the social media.

Keywords: Sent analyzer, Crime analyzer, intelligent crime analyzer, web based sent analyzer.

Dedication

I dedicate this work to the memory of my lovely parents who inspired, encouraged and provided for my education. This dedication also extends to my wife for the support during my project work.

Acknowledgement

I take this opportunity to thank the School of Computing and Informatics, University of Nairobi for giving me a chance to pursue and successfully complete this course. My sincere thanks go to my supervisor, Dr. Lawrence Muchemi and the other members of staff in their various capacities for the untiring support, guidance and concern throughout my project work. I also thank my course mates for the encouragements and team work we have shared during this project. Above all, great thanks go to God for His providential care all through the time of my study.

TABLE OF CONTENTS

List of Tables

List of figures

Definitions of terms

Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks.

Twitter is an online social networking and micro blogging service that enables users to send and read short 140-character text messages.

Corpus (plural *corpora*) or **text corpus** is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

A **blog** (a truncation of the expression ***web log***) is a discussion or informational site published on the World Wide Web and consisting of discrete entries ("posts") typically displayed in reverse chronological order (the most recent post appears first).

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

An algorithm is an effective method expressed as a finite list of well-defined instructions for calculating a function. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of Well-defined successive states, eventually producing "output"

CHAPTER ONE: INTRODUCTION

1.1 Preamble:

The information age has led to materialization of the social media which pose a serious challenge on how to manage the massive information that is available online to various population worldwide as it has various advantages such as people are able to share information but at the same time the level of crime has risen in terms of people have adopted new ways of organizing in groups to spread hate speech which finally lead to crime a case study of strong mobilization of such groups has been witnessed in the Arab nations that has led to ouster of various presidents. It is therefore important for law enforcement agencies to change their tact in dealing with crimes from social sites such as Facebook, twitter, MySpace to ensure that crimes are detected before they are committed.

The sentimental analysis is commonly referred to as techniques used to determine the prejudice and schism of text, usually expressed in free text form. It is a fast emerging technology which has generated tremendous interest among academics as well as business organizations. This can be attributed to the evolution of research in the field of text analytics which has allowed researchers to devise algorithms and techniques to discover sentiments from free text more effectively than ever.

These techniques have been put to use in several practical applications such as social media monitoring, brand reputation management, online commerce, etc. to name a few. Sentiment analysis can be applied across variety of domains such as politics, social media and many others. Most sentiment analysis algorithms are not ideally suited for this task because they exploit indirect indicators of sentiment that can reflect genre or topic instead. Hence, such algorithms used to process social web texts can identify spurious sentiment patterns caused by topics rather

than affective phenomena. It is therefore important for law enforcement agencies especially the Kenya police which have been lagging behind in terms of technology to partner with the relevant people who are currently working in this area of sentimental analysis so that they can detect the usage of social media in spreading hate messages before crimes are committed.

The two most common sentiment analysis tasks are subjectivity and polarity detection. The former predicts whether a given text is subjective or not and the latter predicts whether a subjective text is positive or negative overall. Less common is sentiment strength detection, which predicts the strength of positive or negative sentiment within a text.

This section primarily deals with polarity detection although the methods are applicable to all three tasks. A common approach for sentiment analysis is to select a machine learning algorithm and a method of extracting *features* from texts and then train the classifier with a human-coded corpus. An alternative polarity detection method is to identify the likely average polarity of words within texts by estimating how often they co-occur with a set of seed words of known and unambiguous sentiment.

Rillof et al. (2003) came up with several ways to dig out subjectivity models from subjectivity clauses and to label subjectivities of sentences. In the first method hints were separated into strongly subjective and softly subjective by the rule that “a strong subjective hint is one that is rarely used without a subjective meaning, whereas a soft subjective hint is one that frequently has both subjective and objective meanings Second, sentences were classified as subjective if they contain two or supplementary strong subjective hints, and classified as objective if they contain no strong ten subjective hints and at most one soft subjective hint in the current, previous, and next sentences.

The last was the development of a learning algorithm that was applied to learn subjective mining models using the annotated subjective and objective sentences as training corpus. The learning process contained two steps. First, instantiate the mining models in the training corpus according to the syntactic templates.

Then calculate the number of times each model occurs in subjective training corpus or objective corpus, and then ranked the mining model using the conditional probability measure. Finally, they used a bootstrapping method to apply learned mining models to classify unlabeled sentences from un-annotated text collections. The Subjective Sentence Classifier classifies a sentence as subjective if it contains at least one mining model in the training data.

Pang et al. (2002) researched on opinion analysis using movie review data. It was a document-level supervised learning and they applied Bayesian and Maximum Entropy to the attributes spaces they constructed. They found that the three machine learning methods outperformed the human conducted classifications (two students were asked to classify the corpus), and an algorithm performed better than other machine learning methods. They also found that bigrams did not perform better than unigrams with all three classification methods. To investigate performance of different weighting methods, they assigned binary attributes values that denoted presences/ absences and frequencies as attributes values.

The results showed that presence could perform better than frequencies. Gamon (2004) realized that before applying machine classification they had to get correct attributes for automatic sentiment classification it is for this reason that in order to come up with an effective way to analyse the social media by the Kenya police it will need thorough and precise classification of various words or sentences to achieve the required goal of detecting crime, though it may prove

challenging because we have forty two languages in Kenya and one word has several meanings in various tribes .The motivation for their research was pegged on the higher number of clients they received it was then necessary to propose a system that could deal with these large volume and noisy data automatically.

Due to the success of online social networking and media sharing sites and the consequent availability of a wealth of social data, social network analysis has gained significant attention in recent years. In spite of the growing interest, however, there is little understanding of the potential business applications of mining social networks (Bonchi, et al., 2011). Despite being rich in content, social media unlike traditional media, have unorganized content contributed by users, often in fragmented and sparse fashion. Users wishing to get useful information usually have to spend a lot of their time filtering useless information and yet are not able to capture the essence. Though significant research efforts have been put in sentiment classification and analysis, most of the existing techniques rely on natural language.

Computers can use machine learning, statistic, and natural language processing techniques to perform automated sentiment analysis of digital texts on large collections of texts, including web pages, online news, internet discussion groups, online reviews, web blogs and social media. Processing tools to parse and analyze sentences in a review, yet they offer poor accuracy, because the writing in online reviews tends to be fragmented and less formal than writing in news or journal articles. Many opinion sentences contain grammatical errors and unknown terms that do not exist in dictionaries.

1.1 Problem Statement

What is hard nowadays is not availability of useful information but rather extracting it in the proper context from the vast quantities of content (Yessenov and Misailovic, 2009). This information can provide some value to law enforcement agencies on how to monitor negative comments as used in the social media as doing collection manually then filtering is a cumbersome affair and hence the research problem of automatic categorization and organizing data is apparent.

It has been observed that whereas there is a large body of research on different problems and methods for social network mining, there is still a gap between the techniques developed by the research community and their deployment in real world applications (Bonchi, et al.,2011).

Lack of the model with characteristics identified has then informed this research to address the above problem by defining a sentiment analyzer based on Naïve Bayes; this is because the potential business impact of these techniques is still largely unexplored. A constant flow of information is generated as users of the social media interact with massive data.

This project proposes a model that is based on the Naïve Bayes to address the problem in the cyber space as people exchange their opinions on the social media by classifying sentiments into positive, negative or neutral and generating summaries and trends of the classified data. These summaries are intended to find applicability in supporting law enforcement agencies of crime detection before they occur; the study investigates issues that surround the development of a semantic analyzer of multilingual capability with the view of having it embedded in a web based application for use in rating opinions on different subjects.

1.2 Objectives of the project:

The main objective of this work is to design a model which is capable of categorizing data from the cyber space as positive, negative and neutral.

The specific objectives of this work are therefore summarized as follows:

I. Design a sentiment analysis model based on Naïve Bayes.

II. Develop a sentiment classifier that is able to classify sentiments into positive, negative and neutral.

III. Embed the classifier developed in a web based application for the purpose of analyzing the efficacy of the designed Naïve Bayes Model.

1.3 Justification:

The product of this research, which is mainly the system model, is likely to benefit the following three groups as explained below:

Researchers: There is going to be an addition of new knowledge on machine learning in terms of the new approach of the use of Naïve Bayes Model in sentiment analysis in crime detection.

Community of developers: they can use the results from this work (Naïve Bayes Model) to develop other systems.

Users: will benefit when the application is fully implemented and hence law enforcement agencies can use it to detect sentiments expressed in the social media this will then reduce crime that emanate from the social sites before they occur or mature, their detection and increased emphasis on co-operation and sharing intelligence means that law enforcement agencies are likely to gain access to sensitive information that are used to commit crime.

1.4 Limitations:

The words used on Facebook posts and Twitter do not fully constitute formal language, they involve acronyms, emoticons, slang and sheng. This at times makes it difficult as there is also an increase of such slang in the social media.

Identification of the hardware from which the tweets, posts are sent from is a challenge since partnership with the law enforcement agencies in providing such information is a challenge, due to the limitation of time for this study only small sizes of the corpora was used in the experiments.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Social media is an umbrella term that describes websites that connect individuals somehow. A hallmark of social media is the user generated content. This model contrasts with the editorially controlled style of old media. Social media is sometimes called Web 2.0. The best way to define social media is to break it down. Media is an instrument on communication, like a newspaper or a radio, so social media is a social instrument of communication. In Web 2.0 terms, this would be a website that does not just give you information, but interacts with you while giving you that information. This interaction can be as simple as asking for your comments or letting you vote on an article, or it can be as complex as the process of recommending movies to a user based on the ratings of other people with similar interests. Regular media is synonymous to a one-way street where you can read a newspaper or listen to a report on television, but you have very limited ability to give your thoughts on the matter. Social media, on the other hand, is a two-way street that gives you the ability to communicate too. Web 2.0 is a category of new Internet tools and technologies created around the idea that the people who consume media, access the Internet, and use the Web should not passively absorb what is available; rather, they should be active contributors, helping customize media and technology for their own purposes, as well as those of their communities. These new tools include, but are by no means limited to, blogs, social networking applications, RSS, social networking tools, and Wikis.

2.1 Sources of Social Media:

Sources or groupings of the social media are outlined as follows:

Social Photo and Video Sharing: - These sites are also known as Content Hosting Services. Content hosting or content sharing sites allow users to upload content that they have created for others to view. Two of the most popular of these sites are YouTube www.youtube.com for videos and Flickr www.flickr.com for photographs. Users can also create an individual profile and list their favorite photos or videos. Users are able to rate and comment on the videos or photos posted and provide feedback to the creator and other users.

Data harvesting in you tube.

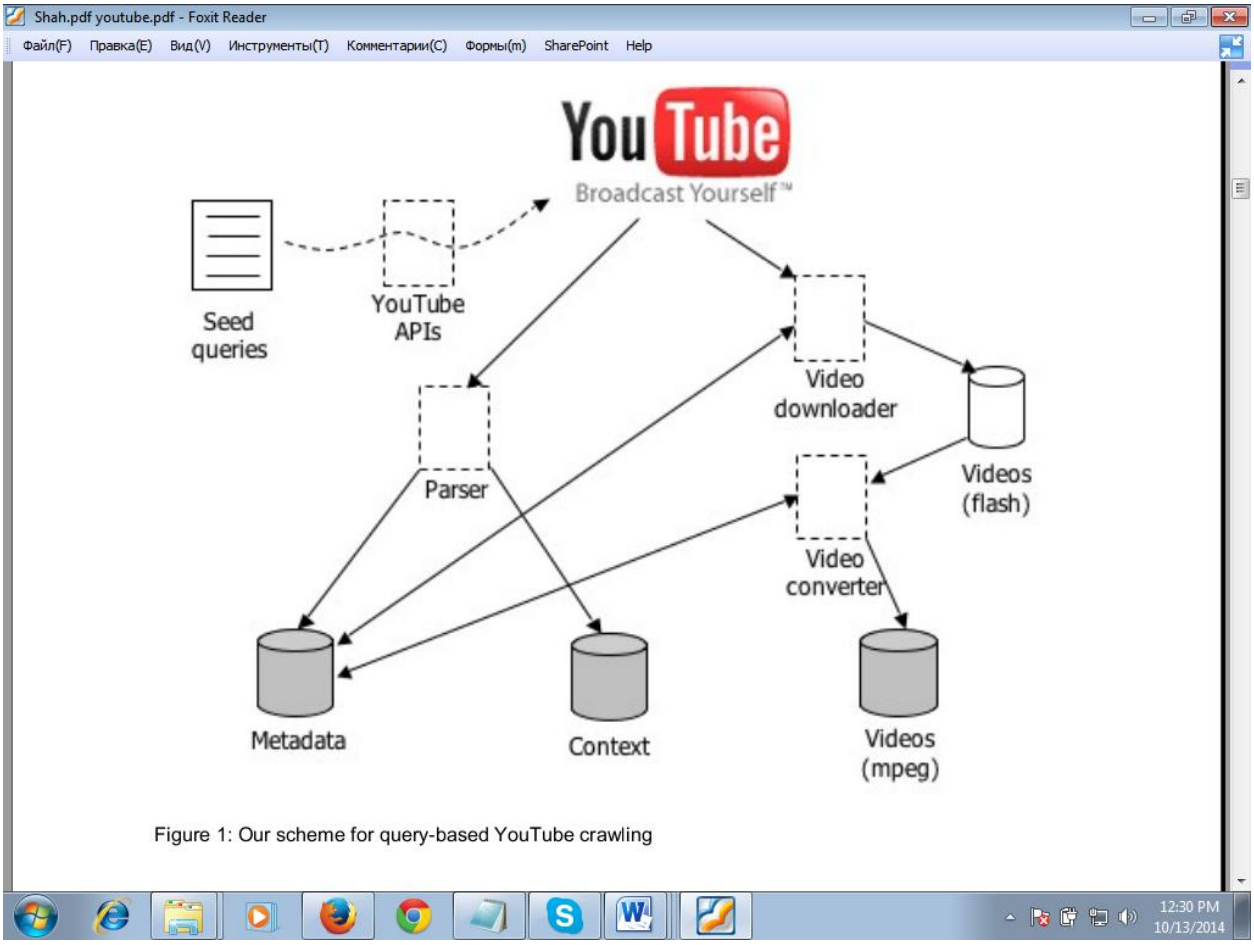
Ever since its inception in 2005, YouTube has emerged as a premium forum for hosting online videos. In this time, YouTube has become much more than posting, viewing, and sharing digital videos; it has become a platform where people express their opinions, participate in discussions, and voice their issues in many creative ways (Gomes, 2006).

The 2008 US presidential election was unique in that it was the first election where a tool like YouTube was used very extensively, creatively, and methodically for the first time (Dalton, 2007; Jarvis, 2007; Seelye, 2007). Due to its large impact on political movements and public opinions, it became essential for anyone - political and social scientists, archivists, curators, information scientists, journalists, and librarians - interested in studying the elections to monitor and analyze YouTube activities around the elections.

We present Tube Kit, a query-based YouTube crawling toolkit. This software is a collection of tools that allows one to build one's own crawler that can crawl YouTube based on a set of seed queries and collect up to 17 different attributes. Tube Kit assists in the phases of this process starting with database creation to finally giving access to the collected data with browsing and searching interfaces. We further demonstrate how we used this toolkit to collect elections related

data from YouTube for nearly two years. Some analysis of the collected data relating to the elections is also given. Keywords: YouTube crawling, video data collection, presidential elections 2008 hence this in comparison can be used to collect data related to crime that are reported in you tube.

Figure 2.1 Data extraction in youtube



A **blog:** (a truncation of the expression ***web log***) is a discussion or informational site published on the World Wide Web and consisting of discrete entries ("posts") typically displayed in reverse chronological order (the most recent post appears first). Until 2009 blogs were usually the work of a single individual, occasionally of a small group, and often covered a single subject. More recently "multi-author blogs" (MABs) have developed, with posts written by large numbers of authors and professionally edited. MABs from newspapers, other media outlets, universities,

think tanks, advocacy groups and similar institutions account for an increasing quantity of blog traffic. The rise of Twitter and other "micro blogging" systems helps integrate MABs and single-author blogs into societal new streams. *Blog* can also be used as a verb, meaning *to maintain or add content to a blog*. Many blogs provide commentary on a particular subject; others function as more personal online diaries; others function more as online brand advertising of a particular individual or company. A typical blog combines text, images, and links to other blogs, Web pages, and other media related to its topic. The ability of readers to leave comments in an interactive format is an important contribution to the popularity of many blogs. Most blogs are primarily textual, although some focus on art (art blogs), photographs (photoblogs), videos (video blogs or "vlogs"), music (MP3 blogs), and audio (podcasts). Micro blogging is another type of blogging, featuring very short posts. In education, blogs can be used as instructional resources. These blogs are referred to as edublogs.

Data harvesting in blogs, sample code. The code shown below can help get lattest posts from blogs

```
function get_recent_posts($no_posts = 1, $before = '<li>', $after = '</li>', $show_pass_post = false,
$skip_posts = 0) {
global $wpdb, $tableposts, $tablepost2cat;
$request = "SELECT ID, post_title, post_content, category_id FROM $tableposts, $tablepost2cat WHERE
post_status = 'publish' AND (post_id = ID AND category_id != '5')";
if(!$show_pass_post) { $request .= "AND post_password =" "; }
$request .= "ORDER BY post_date DESC LIMIT $skip_posts, $no_posts";
$posts = $wpdb->get_results($request);
$output = '';
```

Wikis: - This is a collaborative website that anyone within the community of users can contribute to or edit. A wiki can be open to a global audience or can be restricted to a select network or community. Wikis can cover a specific topic or subject area. Wikis also make it easy to search or browse for information. Although primarily text, wikis can also include images, sound recordings & films. Wikipedia http://en.wikipedia.org the free internet encyclopedia is the most well-known wiki.

Social News: Sites under this category allow a user to interact by voting for articles and commenting on them it includes all the digital newspapers. Examples under this category include http://reddit.com and http://www.digg.com.

Computer-supported collaboration (CSC): The research focuses on technology that affects groups, organizations, communities and societies, e.g., voice mail and text chat. It grew from cooperative work study of supporting people's work activities and working relationships. As net technology increasingly supported a wide range of recreational and social activities, consumer markets expanded the user base, enabling more and more people to connect online to create what researchers have called a computer supported cooperative work, which includes "all contexts in which technology is used to mediate human activities such as communication, coordination, cooperation, competition, entertainment, games, art, and music" (from CSCW 2004).

Micro blogging: is a broadcast medium that exists in the form of blogging. A micro blog differs from a traditional blog in that its content is typically smaller in both actual and aggregated file size. Micro blogs "allow users to exchange small elements of content such as short sentences, individual images, or video links". These small messages are sometimes called *micro posts*.

As with traditional blogging, micro bloggers post about topics ranging from the simple, such as "what I'm doing right now," to the thematic, such as "sports cars." Commercial micro blogs also exist to promote websites, services and products, and to promote collaboration within an organization.

Some microblogging services offer features such as privacy settings, which allow users to control who can read their micro blogs, or alternative ways of publishing entries besides the web-based interface. These may include text messaging, instant messaging, E-mail, digital audio or digital video. Among the most notable services are Twitter, Tumblr, Friend Feed, Cif2.net, Plurk, Jaiku and identi.ca. Different versions of services and software with micro blogging features have been developed. Plurk has a timeline view that integrates video and picture sharing. Flipter uses microblogging as a platform for people to post topics and gather audience's opinions. Emote.in has a concept of sharing emotions, built over micro blogging, with a timeline.

Wikis: - This is a collaborative website that anyone within the community of users can contribute to or edit. A wiki can be open to a global audience or can be restricted to a select network or community. Wikis can cover a specific topic or subject area. Wikis also make it easy to search or browse for information. Although primarily text, wikis can also include images, sound recordings & films. Wikipedia http://en.wikipedia.org the free internet encyclopedia is the most well-known wiki.

2.2. Objective of Sentiment analysis

Liu (2010) sums up his observations that given an opinionated document d, the Objectives of

Sentiment analyses on direct opinions are:

To discover all opinion quintuples (oj, fjk, ooijkl, hi, tl) in d, and

To Identify all synonyms (Wjk) of each feature fjk in d.

Liu further observes that not all five pieces of information in the quintuple need to be discovered for every application because some of them may be known or not needed. For example, in the context of online forums, the time when a post is submitted and the opinion holder are all known as the site typically displays such information.

A Sentiment can be identified at several levels which include the overall document (e.g., product review, blog, forum post), a sentence or a specific object attribute. For each level, search and analysis operates under somewhat different assumptions.

2.3 Types of opinions

An opinion can be either one of the following two types:-

Direct opinion: According to Damer, T. Edward (2008) a direct opinion is a quintuple (oj, fjk, ooijkl, hi, tl), where oj is an object,fjk is a feature of the object oj, ooijkl is the orientation of the opinion on feature fjk of object oj, hi is the opinion holder and tl is the time when the opinion is expressed by hi. The opinion orientation ooijkl can be positive, negative or neutral.

Comparative opinion: A comparative opinion expresses a preference relation of two or more objects based on some of their shared features. It is usually conveyed using the comparative or superlative form of an adjective or adverb, e.g., “Jaguar is better than Mercedes.

Other types of opinions

Public opinion In contemporary usage, public opinion is the aggregate of individual attitudes or beliefs held by a population (e.g., a city, state, or country), while consumer opinion is the similar aggregate collected as part of marketing research (e.g., opinions of users of a particular product or service). Typically, because the process of gathering opinions from all individuals are difficult, expensive, or impossible to obtain, public opinion (or consumer opinion) is estimated using survey sampling (e.g., with a representative sample of a population).

Group opinion In some social sciences, especially political science and psychology, group opinion refers to the aggregation of opinions collected from a group of subjects, such as members of a jury, legislature, committee, or other collective decision-making body. In these situations, researchers are often interested in questions related to social choice, conformity, and group polarization.

Scientific opinion "The scientific opinion" (or scientific consensus) can be compared to "the public opinion" and generally refers to the collection of the opinions of many different scientific organizations and entities and individual scientists in the relevant field. Science may often, however, be "partial, temporally contingent, conflicting, and uncertain"so that there may be no accepted consensus for a particular situation. In other circumstances, a particular scientific opinion may be at odds with consensus. Scientific literacy, also called public understanding of science, is an educational goal concerned with providing the public with the necessary tools to benefit from scientific opinion.

Legal opinion A "legal opinion" or "closing opinion" is a type of professional opinion, usually contained in a formal legal-opinion letter, given by an attorney to a client or a third party. Most legal opinions are given in connection with business transactions. The opinion expresses the attorney's professional judgment regarding the legal matters addressed. A legal opinion is not a

guarantee that a court will reach any particular result . However, a mistaken or incomplete legal opinion may be grounds for a professional malpractice claim against the attorney, pursuant to which the attorney may be required to pay the claimant damages incurred as a result of relying on the faulty opinion.

Judicial opinion A "judicial opinion" or "opinion of the court" is an opinion of a judge or group of judges that accompanies and explains an order or ruling in a controversy before the court, laying out the rationale and legal principles the court relied on in reaching its decision. Judges in the United States are usually required to provide a well-reasoned basis for their decisions and the contents of their judicial opinions may contain the grounds for appealing and reversing of their decision by a higher court.

Editorial opinion An "editorial_opinion" is the stated opinion of a newspaper or of its publisher, as conveyed on the editorial_page.

2.4 Approaches to Sentimental Analysis

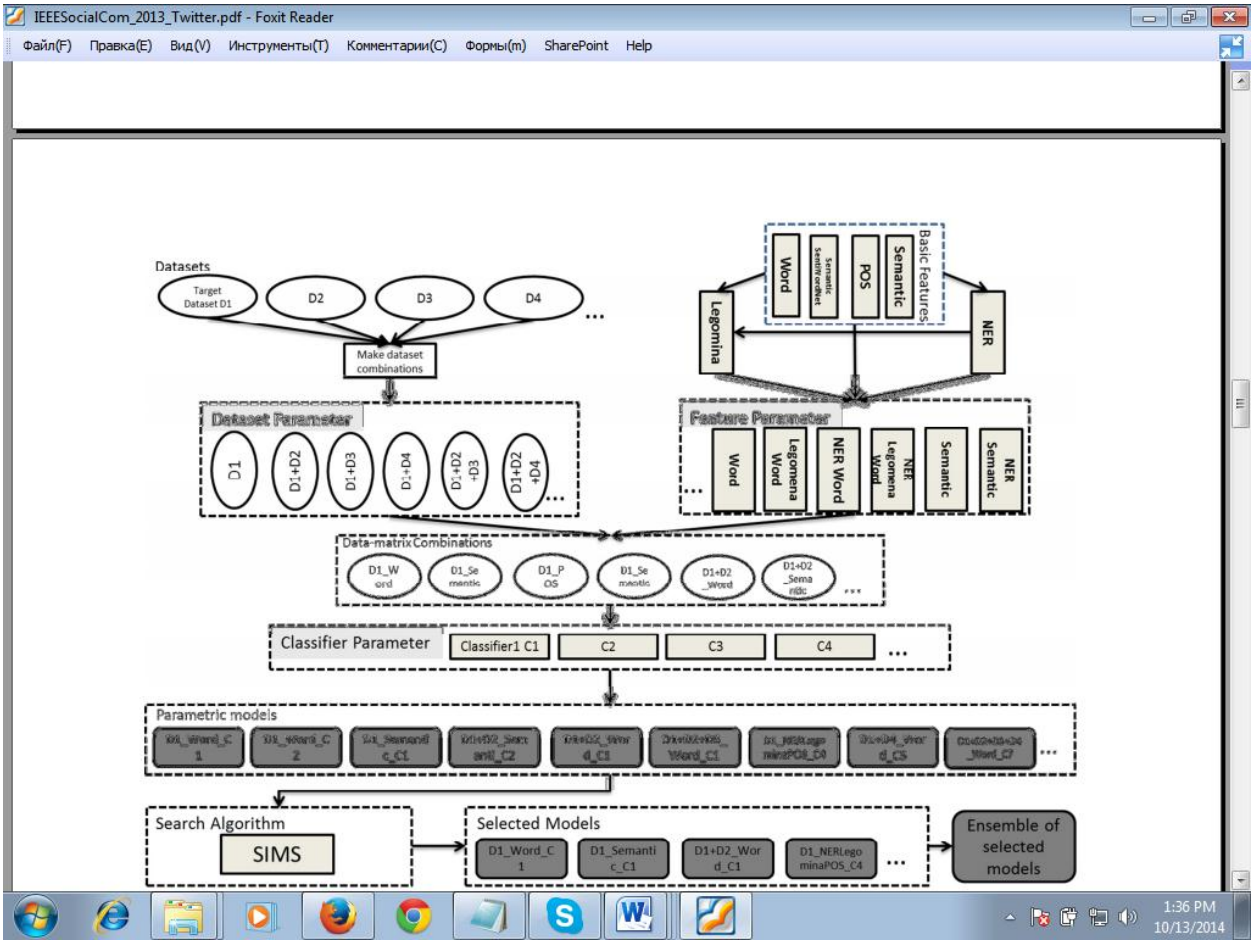
A lot of study has been done on sentiment analysis by using supervised learning and unsupervised learning techniques. This has been made possible because of the development in artificial intelligence as a distinct area of study. It is for this reason that this study tend to provide an algorithm that may help in analysis of words that may lead to crime detection especially in social sites. For supervised learning, attributes selection methods are important to classification performance. Before the recognition of polarity classification studies, which is to identify positive or negative polarities of a material or sentences, many studies were on prejudice classification, which is used to categorize whether sentences or words are subjective or objective.

2.4.1 Bootstrapping

In machine learning bootstrapping is iterative training on a known set **.**As revealed in Riloff et al. (2003), slanted terms include opinions, rants, allegations, accusations, suspicions, and speculations. Riloff et al.(2003) presented a bootstrapping process that learned linguistically rich

mining model for biased expressions. The learned blueprints were then used to routinely identify whether a sentence was subjective or objective. The results showed that the mining models had better output than other older models. Rillof et al. (2003) came up with several ways to dig out subjectivity models from subjectivity clauses and to label subjectivities of sentences. In the first method hints were separated into strongly subjective and softly subjective by the rule that "a strong subjective hint is one that is rarely used without a subjective meaning, whereas a soft subjective hint is one that frequently has both subjective and objective meanings Second, sentences were classified as subjective if they contain two or supplementary strong subjective hints, and classified as objective if they contain no strong ten subjective hints and at most one soft subjective hint in the current, previous, and next sentences. The last was the development of a learning algorithm that was applied to learn subjective mining models using the annotated subjective and objective sentences as training corpus. The learning process contained two steps. First, instantiate the mining models in the training corpus according to the syntactic templates. Then calculate the number of times each model occurs in subjective training corpus or objective corpus, and then ranked the mining model using the conditional probability measure. Finally, they used a bootstrapping method to apply learned mining models to classify unlabeled sentences from un-annotated text collections.

Figure 2.2 Bootstrapping model



The Subjective Sentence Classifier classifies a sentence as subjective if it contains at least one mining model in the training data. Pang et al. (2002) researched on opinion analysis using movie review data. It was a document-level supervised learning and they applied Bayesian and Maximum Entropy to the attributes spaces they constructed. They found that the three machine learning methods outperformed the human conducted classifications (two students were asked to classify the corpus), and an algorithm performed better than other machine learning methods. They also found that bigrams did not perform better than unigrams with all three classification methods.

To investigate performance of different weighting methods, they assigned binary attributes values that denoted presences/ absences and frequencies as attributes values. The results showed that presence could perform better than frequencies.Gamon (2004) realized that before applying machine classification they had to get correct attributes for automatic sentiment classification it is for this reason that in order to come up with an effective way to analysis the social media by the Kenya police it will need thorough and precise classification of various words or sentences to achieve the required goal of detecting crime, though it may prove challenging because we have forty two languages in Kenya and one word has several meanings in various tribes .The motivation for their research was pegged on the higher number of clients they received it was then necessary to propose a system that could deal with these large volume and noisy data automatically.

Gamon (2004) tested with a variety of different attributes sets, from deep linguistic analyses based attributes to surface-based attributes. The surface-based attributes contain unigrams, bigrams, and trigrams. The linguistic attributes contain part-of-speech trigrams, length measures (e.g., length of sentences), structure models (e.g., DECL::NP VERB NP denotes a declarative sentence consisting of a noun phrase, a verbal head, and a second noun phrase), and tags coupled with semantic relations (e.g., "Verb-Subject-Noun" indicates a nominal subject to a verbal predicate). Binary attributes weighting values were assigned to the attributes. The outcome showed that the usage of linguistic analysis based attributes consistently contributed to higher classification accuracy in sentiment classifications.

Apart from focusing on the right attributes and conveying right attributes weighting values, the application of attributes selection methods is also important. Yang et al. (1997) pointed out that a major feature of text categorization problem is the high dimensionality of attributes spaces.

Attributes used in text categorizations are usually category on word attributes such as unigrams or n-grams in the corpus, the size of which are usually decided by the size of vocabularies contained in the corpus. A big corpus usually contains tens of thousands vocabularies. The high dimensionalities in machine learning process could result in the curse of dimensionality, which refers to various phenomena that arise when analyzing and organizing high dimensional spaces (Wikipedia). High dimensions could cause a attributes space to contain many sparse values. Yang et al. (1997) focused on evaluating and comparing several attributes selection methods that can reduce dimensions of attributes spaces in text categorizations. Attributes selection methods that were compared in their studies included DF, IG, x , Mutual Information (MI), and term strength (TS).

2.4.2 K-nearest neighbor classifier

The K-nearest neighbor (KNN) is a typical example-based classifier that does not build an explicit, declarative representation of the category ci, but rely on the category labels attached to the training documents similar to the test document. As a result, KNN has been called lazy learners, since it defers the decision on how to generalize beyond the training data until each new query instance is encountered. Given a test document d, the system finds the k nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. They used classification methods k-nearest-neighborhood (kNN) and Linear Least Squares Fit (LLSF) mapping. Using both of them could reduce the possibility of classifier bias. Each of the attributes selection method was evaluated using a number of different term-removal thresholds. The results showed that IG, DF and x could eliminate up to 90% or more unique attributes with either an improved or no loss in categorization accuracy under kNN and LLSF Forman (2003) presented an

empirical method to compare twelve attributes selection methods to investigate which attributes selection method or combination of methods was most likely to produce the best performance. They found that Information Gain could get highest precision among the twelve selection methods. Except supervised learning, unsupervised learning is also used often for sentiment analysis.

2.4.3 Unsupervised Learning.

Unsupervised learning involves the calculation of the opinion polarities of opinion words, and classifies the documents or sentences by aggregating the orientation of opinion words. Turney (2002) presented a simple unsupervised learning algorithm to classify the reviews based on recommended or not recommended reviews online. The sentiment classification of a review is predicted by the average semantic orientation of adjective or adverb phrases in the review. Opinions are usually expressed by adjectives and adverbs. They used Point-wise Mutual Information and Information Retrieval to measure the similarity of pairs of words or phrases, which is to calculate semantic orientation of a word or phrase by subtracting mutual information between the word or phrase and the reference word “excellent” from the mutual information between the word or phrase and the reference word “poor”.

The mutual information is the co-occurrence of the two words or phrase among millions of online documents. Using such as “bad scene” or “good scene” which are not sentiment words it will be therefore important to select that capture feelings.

Although they received a decent result, the way they calculated the semantic orientation of phrases was not efficient enough as it involved retrieving millions of online documents to get the co-occurrence of two words. In sentiment analysis, especially in an unsupervised learning

process, opinion word lexicons are usually created first. An opinion word lexicon is a list of opinion words with interpreted opinion polarities. Then opinion word lexicons could be used to deduce the polarities of other words in the context, or be treated as attributes in attributes spaces for supervised learning. The methodology that I am proposing to adopt is related to Martin and White's Appraisal Theory, Whitelaw et al.(2005) which presented a method for extraction to formulate a lexicon. "An appraisal group is a set of attribute values in several task-independent semantic taxonomies.

This will focus on mining and analysis of adjectival appraisal optionally modified by a sequence of modifiers (such as 'very', 'sort of', or 'not').

Different domains or contexts usually need different opinion lexicons because opinion words are context dependent. One positive opinion word in one domain may be neutral in another domain or context. This is the challenge that many organizations that may want to use this kind of analysis may experience since we live in a world that has various groups speaking different languages a case study of Kenya which has above 42 languages the already polarity of an opinion word in one lexicon is usually called prior polarity. Wilson et al. (2005) proposed a method to automatically distinguish prior polarity from contextual polarity of a phrase. Because categorizations that are based on prior polarities of opinion words are not precise enough as argued earlier, Wilson et al. (2005) conducted classification experiments by developing attributes such as word attributes, modification attributes, and structure attributes to identify contextual polarities of phrases. It is therefore important to use more than twenty meanings from a word so as to capture more hidden meanings it may have.

The developed attributes that took into account the contextual polarities produced high classification performance. Eguchi et al. (2006) proposed a method based on the assumption that sentiment expressions are related to topics. For example, negative reviews for some selection events may contain kinds of indicator word “flaw”. They combined topic relevance models and sentiment relevance models with parameters that were estimated from training data using retrieval models. Sentence-level analysis was conducted, and one sentence was treated as one statement. Each statement consisted topic bearing and sentiment bearing words. They trained the model by annotating S (sentiment) and T (topic) to sentiment words and topic words. Then, S, T, and polarities of the sentiment words formed a triangular relationship, which was trained by a generative model. The classification obtained high performance using the trained models.

2.4.4 Semi supervised learning

Semi-supervised learning approach is an approach to reduce the need for labeled data by taking advantage of unlabeled data. In general, there are two kinds of semi-supervised learning approaches. One is to bootstrap class labels using techniques like self-training, Expectation Maximization (EM) and co-training.

Self-training trains a classifier and uses it to classify unlabeled data, and then add the most confident data to the training data and repeat the process. EM approach can be viewed as a special case of”soft” self-training. It assumes the data is generated according to some known parametric models, and then iteratively estimates the expectation of hidden class variables and update the model parameters. Co-training splits features into two sets and trains two classifiers. Each classifier picks its most confident data and retrains with the additional labeled data provided by each other. One can imagine that classification mistakes can reinforce it by using this kind of methods. Another category is structural learning methods which learn good functional structures using unlabeled data. It proposed a graph-based method which constructs a

graph with labeled and unlabeled examples as nodes and their similarity relationships as edges. It makes the assumption of label smoothness over the graph proposed a framework to learn predictive structures on hypothesis spaces using unlabeled data, and then use these structures to enhance learning. The performance of such methods is influenced by how much the structure characterizes the underlying hypothesis.

2.5. Common Issues

(a) Technical challenges

A number of technical challenges have been observed in sentiment analysis (Liu, 2010). The first is Object identification. A blog or a tweet may have sentiments expressed on different objects. In such a case the problem lies in identifying the object on which a sentiment has been expressed without which the opinion is of little use. In a typical opinion mining application, the user wants to find opinions on some competing objects (e.g., products). The system thus needs to separate relevant objects and irrelevant objects.

The second is feature extraction and synonym grouping. Current research mainly finds nouns and noun phrases. Although the recall may be good, the precision can be low. Furthermore, verb features are common as well but harder to identify. To produce a good summary there is need to group synonym features as people often use different words or phrases to describe the same feature (e.g., "voice" and "sound" refer to the same feature). This problem is also very hard. A great deal of research is still needed.

The third challenge is on opinion orientation classification. The task here is determination of whether there is opinion on a feature in a sentence, and if so, whether it is positive or negative.

Existing approaches are based on supervised and unsupervised methods. One of the

key issues is to identify opinion words and phrases (e.g., good, bad, poor, great), which are instrumental to sentiment analysis.

The problem is that there are seemingly an unlimited number of expressions that people use to express opinions, and in different domains they can be significantly different. Even in the same domain, the same word may indicate different opinions in different contexts. For example, in the sentence, "The battery life is long" "long"indicates a positive opinion on the "battery life" feature. However, in the sentence, "This camera takes a long time to focus", "long" indicates a negative opinion.

Fourthly integration of the five pieces of information in the quintuple so as to get a match is a complex task. That is, an opinion must be given by an opinion holder on a certain feature of an object at a certain time. To make matters worse, a sentence may not explicitly mention some pieces of information, but they are implied due to pronouns, language conventions, and the context. To deal with these problems, we need to apply NLP techniques such as parsing, word sense disambiguation and coreference resolution in the opinion mining context. Other challenges include domain dependency, irony and sarcasm and social elements. Awareness of the multifaceted nature of opinions gives us a foundation upon which to conduct our search and analysis.

(b) Privacy Concerns

In recent years, social network research has been carried out using data collected from online interactions and from explicit relationship links in online social network platforms such as Facebook, LinkedIn, Flickr and Instant Messenger (Bonchi, et al., 2011).

The mining of this data makes it difficult for an individual to autonomously control the unveiling and dissemination of data about his/her private life (Wel & Royakkers, 2004).

Sentiment analysis or opinion mining involves the use of personal data of some kind and can lead to the disruption of some important normative values. One of the most obvious ethical objections lies in the possible violation of peoples' informational privacy. Protecting the privacy of users of the Internet is an important issue.

Informational privacy mainly concerns the control of information about oneself. It refers to the ability of the individual to protect information about oneself. The privacy can be violated when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. Wel and Royakkers note that the privacy issues due to web mining often fall within this category of informational privacy. They further state that the value of peoples' individualism is violated when they are judged and treated based on patterns resulting from web-data mining.

When data obtained from the web is made anonymous the discovered information no longer links to individual persons, and there is no direct sense of privacy violation because the data is not then directly linked to a person.

This study intends to use data from Facebook and Twitter that is specified as public by the users who generated it. The data will be made anonymous to protect the privacy of individuals by picking attributes of posts and tweets that exclude the author's identity.

We will therefore collect public messages, the corresponding message identifiers and the time the messages were created for purposes of this study. Facebook clarifies on its privacy statement that information made public on its site can be viewed by anyone even off Facebook. Such information also shows up when someone does a search on Facebook or on a

public search engine. The information can also be accessed by the games, applications, and websites one visits with friends and is also accessible to anyone who uses its APIs such as the Graph API.

2.6 Features of Sentiment Analyzer

Sentiment classification is a very challenging task, the high dimensionalities in machine learning process could result in the curse of dimensionality, which refers to various phenomena that arise when analyzing and organizing high dimensional spaces. On one hand, traditional text classification techniques usually do not work well on this task, since they tend to view frequent-occurring words as good indicators of the class labels, while in opinionated text; sentiment words are usually ambiguous and infrequent. On the other hand, acquiring human-labeled data for sentiment classification is very difficult. Opinions are hidden in a huge amount of online resources like forums and blogs. Manual annotation is very expensive and time-consuming.

The goal of this project is to design a model based on Naïve Bayes to address the above challenges.

2.7 Feature based Sentiment Analysis Model

With the concepts in section above in mind, Liu (2010) defines a model of an object, a model of an opinionated text, and the mining objective, which are collectively called the feature-based sentiment analysis model.

Model of an object: An object o is represented with a finite set of features, F = {f1, f2, …,fn}, which includes the object itself as a special feature. Each feature fi of F can be expressed with any one of a finite set of words or phrases Wi ={wi1, wi2, …, wim}, which are synonyms of the feature.

Model of an opinionated document: A opinionated document d contains opinions on a set of objects {o1, o2, …, or} from a set of opinion holders {h1, h2, …, hp}. The opinions on each object oj are expressed on a subset Fj of features of oj.

2.8 Conceptual Model

Based on the insight obtained from the relevant literature cited above, this paper draws upon three approaches:-

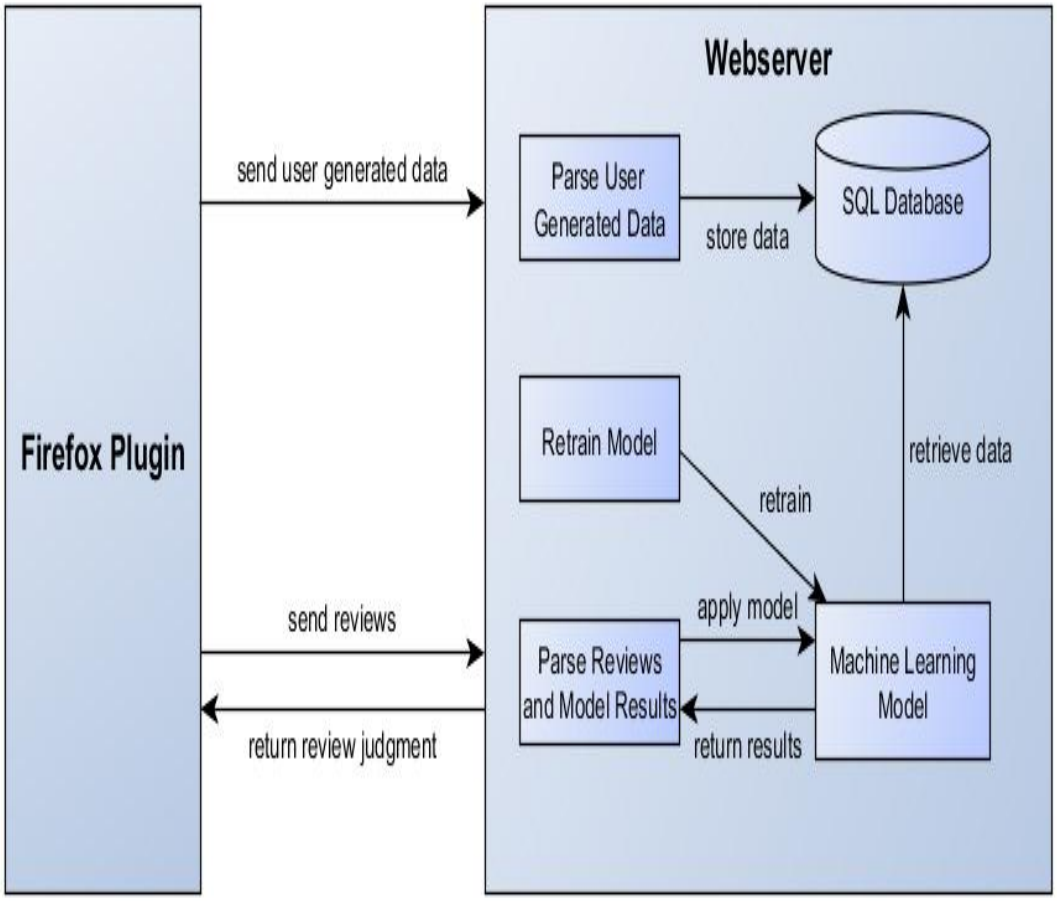
Pattern based approach where we use the unigram model of features

Simple Bag of words model with information gain heuristic

Machine Learning, where we use the Naïve Bayes technique.

These three approaches inform the creation of a conceptual model shown in Figure that is used in this research work.

Figure 2.3 Conceptual Model



CHAPTER THREE: METHODOLOGY

3.0 Preamble:

This chapter is about the research methods that were followed to achieve the overall objectives outlined in section 1.3 above. We explain the sources of training data that were used, data cleaning, the way features were selected and categorized, how classification was achieved and finally how the classifier will be evaluated.

3.1 Research Design

These were summarized as shown in the table below:

Table 3.1 Research Design Activities

No	Objective	Activities	How objective is achieved
1	design	Collect, clean and categorize classifier training data from tweeter	The graph and tweeter APIs were customized and used to collect data from tweeter. The data collected was stored in SQL database then cleaned and stored in files labeled as positive, negative and neutral
2	Develop	Design and build an opinion classifier which will categorize data as positive ,negative and neutral	We used machine learning technique of Naïve Bayes and Natural Language Toolkit of Python programming in

			implementing the classifier
3	Evaluation for efficacy	Customize a classification tool to focus on the Kenyan context	Build a web base application that has an interface that helps in collecting, analyzing and categorizing opinions in Kenyan context.

3.2 Research design

This research design was preferred because of the tinny and informal nature of Social Media data, a scenario that necessitates use of machine learning and computational linguistics techniques. Traditional approaches of natural language processing are challenged by the lack of language formality usage on Social Media.

3.3 Data Collection

3.5.1 Data Collection

We used keywords for current topical issues as search criteria for collecting relevant data Twitter. This exercise was repeated a number of times with the result being a dataset of over 100 tweets fetched on various topical issues.

The data fetched from site was then decoded from JSON (JavaScript Object notation) to text and stored in a MySQL database . All tweets fetched were stored since they don't exceed 140 characters. For each tweet fetched we stored the following attributes:

i. The unique post identifier

ii. Post message

iii. The time the tweet was created.

iv. The item searched for

v. The source of the message i.e Twitter or any other social media.

vi. Polarity which is null as at the point of fetching the data. This field is later updated

3.1.2 Data Preparation for Classifier training

The raw data harnessed from Twitter is directly unsuitable for use in training a polarity classifier since it often has unwanted characters such as links, exclamation marks question marks and other irrelevant characters. These characters are not of essence to this study except for emoticons that are used as means of expression.

The entire corpus stored in the MySQL database was then picked and transferred it to three text files. The three files correspond to the three categories of features on which the classifier is to be trained namely positive, negative and neutral. Time was taken to manually read and clean the contents of each file line by line while removing irrelevant characters, words and sentiments. The result of this exercise was three files each of which is a collection of sentiments that express one category of polarity. We currently have 981 negative sentiments in the file for negatives, 367 neutral sentiments in the file for neutrals and 559 positive sentiments in the file for positives.

During training of the classifier, 75% of the cleaned data combined is used as training data while 25% of the data is used as the testing set

3.1.3 Naïve Bayes Classifier

The Naïve Bayes Classifier is based around the Bayes rule which is a way of looking at conditional probabilities that allows you to flip the condition around in a convenient way. A conditional probability is a probability that event X will occur, given the evidence Y. That is normally written P(X | Y). The Bayes rule allows us to determine this probability when all we have is the probability of the opposite result, and of the two components individually:-P(X | Y) = P(X)P(Y | X) / P(Y).

3.1.4 Feature Extraction and Classifier Development

The NLTK toolkit was used to realize a feature extractor and a Naïve Bayes classifier. The Naïve Bayes classifier requires that the training features be rendered in ‘feature label’ pair for it to recognize that a particular feature belongs to a particular label. In our case the file category (negative, positive or neutral) from which a sentiment comes from is the label for that sentiment while the feature is the sentiment. Since our sentiment

classification is restricted to the sentence level, the data had to be split into a file per sentiment per category.

We use scripts to achieve this splitting. We then use an algorithm in the classifier program that returns a dictionary of feature sets that are rendered as 'feature label' pair as explained above. The returned dictionary of feature sets enables the classifier to learn to associate a feature with a particular label that is positive, negative or neutral. The algorithm is also structured to discard stop words from the dictionary of feature sets used for training.

We used the Naive Bayes classifier that ships with the Natural Language toolkit and customized it to suit our research needs. Its customization is currently complete and functional. It does the following:-

1. Makes the necessary imports of resources such as the categorized corpus reader that we use to read training data stored in files and a set of stop words.
2. Reads in training data from files.
3. Obtains the categories of labels to be used.
4. Generates a feature set for training the classifier using a bag of words model.
5. Splits the feature set into a training set and testing set.
6. Trains a Naïve Bayes classifier imported from NLTK toolkit using the training set.
7. Tests the classifier accuracy using the testing set.
8. Prints out the accuracy and informative features.
9. Retrains the classifier with the entire (combined Training and testing set) set.
10. Connects to the database and loops through each message or tweet as it classifies and updates the polarity of each record.

The classifier has been tested and found to be working at an accuracy of 81.1% according to its inbuilt NLTK accuracy module. A look at the classified records shows that the

classifier is able to determine the polarity of most of the sentiments and the more training of the data increase the more the level of accuracy goes up.

3.1.5 Evaluation of Classifier

In experimenting with the Naïve Bayes Classifier, we relied on the NLTK metrics module which provides functions for calculating accuracy, precision and recall for the Classifier.

3.1.6 Implementation of the system

The Sentiment Analyzer System was implemented using Java language for the application Logic whereas the interface was realized using Java Server Faces (JSF). MySQL database was used as the back end for storing data fetched from Facebook and Twitter and other application data.

3.4 System Design Specification

In this section we specify the functionalities of the system, associated inputs and outputs together with associated data sources. We employed the use of agile software development Methodology to guide in the object oriented development of the application software.

The Sentiment Analyzer System was designed to carry out Sentiment Analysis for Kenyan issues based on data obtained from Facebook and Twitter. This is in line with our earlier stated objectives in section 1.3 above. To ensure realization of the objectives, the research designed the application with four main modules outlined below:-

i. Twitter API interface modules

ii. Classification Module

iii. Reports Module

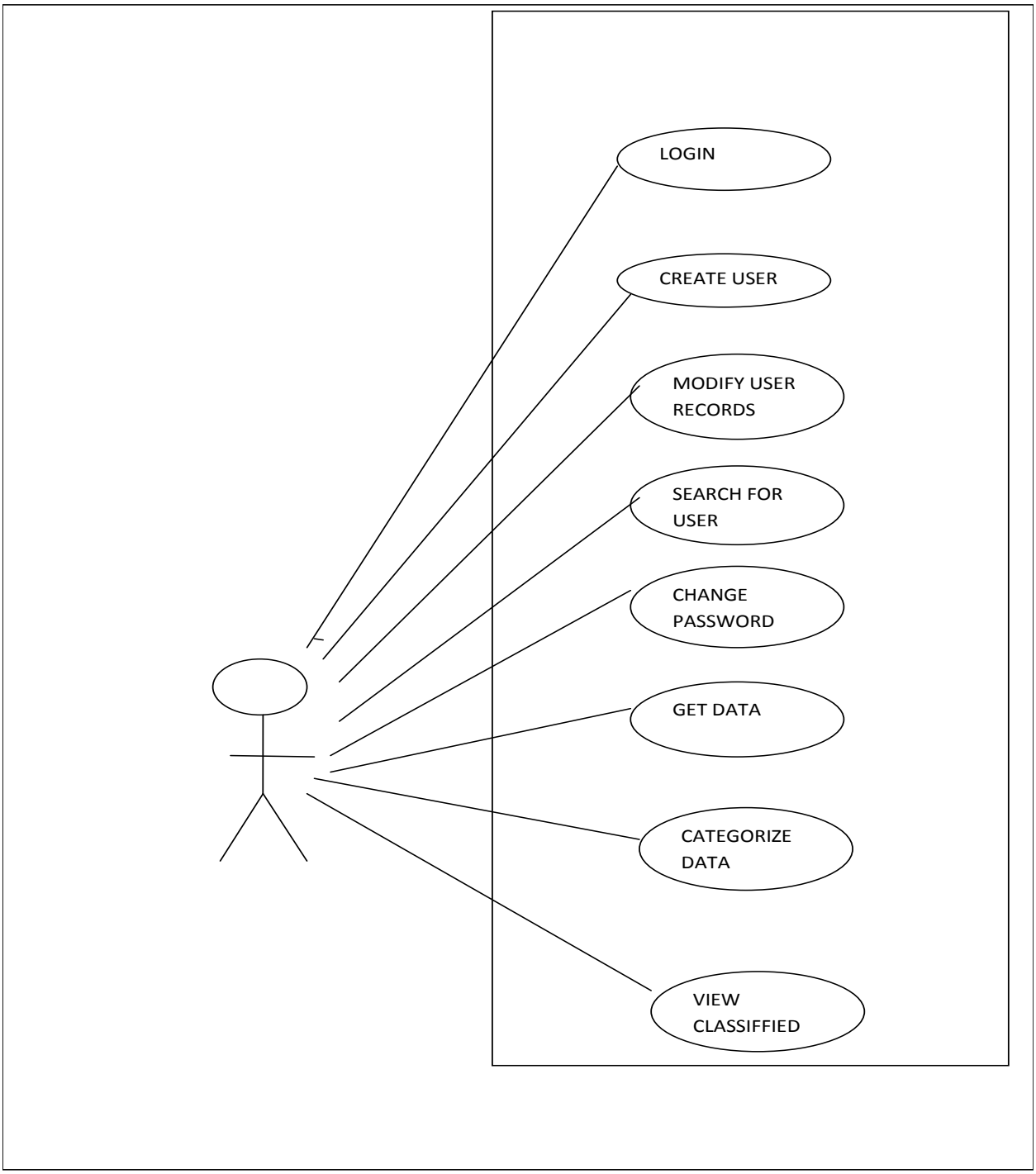
iv. User administration

The research employed Use Cases to describe system – actor interactions.

A use case is a sequence of actions that provide a measurable value to an actor. Another way to look at it is a use case describes a way in which a real-world actor interacts with the system. The overall goal of the system is to classify Kenyan sentiments obtained from Twitter into three categories; positive, negative or neutral. It therefore has to provide interface for

searching and fetching relevant Twitter data. Diagrammatically, the system Use Case Model is as shown in figure below:

Figure 3.1 . Use Case Model



3.5 ACTORS

Two actors are involved namely:-

i. The System

This refers to the Sentiment Analyzer System which constitutes various modules

Interconnected to provide various functionalities that fulfill the objectives of this research.

ii. User of the system

This is the human actor interested in benefiting from the utilities offered by the System functions.

3.5.1 USE CASES:

Table 3.2 Login authenticate user

Use case name	Authenticate user
Actors	User and sentiment analyzer
Description	Authenticate user functionality in order for them to access the system's functionality
Precondition	1.The account of the user has been created 2.The user has the right to access the system
Triggers	The user launches the application by launching the browser by typing in the address of the application.
Basic course	The following shall be the procedure to authenticate users of

	the system: 1. The actor enters address and the browser launches the application. 2. The system responds by availing the log in form. 3. The user responds by providing user name and password. 4. The system responds by displaying the interface of the system. 5. End of the system.
Exceptions paths	1.If user has not entered user id • System informs the user invalid user name • System functionality remain unavailable to the user • System fails the log in attempts 2.If user has not entered password • System notifies the user wrong password or to reenter password. 3.If the user name or password is invalid • System notifies the user wrong password or user name. • The functionality remain unavailable for the user

Table 3.3 Extract data from tweeter

Use case name	Extract data
Actors	User, sentiment analyzer and tweeter
Description	This functionality allows users to fetch data tweeter
preconditions	1.the user must have an account 2.the actor has been authenticated 3.the functionality is available through the interface
Triggers	The actor select fetch data from tweeter
Basic course	The following shall be the procedure for extracting data from tweeter. 1.The actor select from the interface extract data 2.The system provide button for extracting data 3.Th actor types the word to be searched 4.The system sends search request to tweeter through tweeter API 5.Tweeter responds by providing the results. 6. The actor views the results 7.the system stores the results 8.end of the use case
Exception path	1.If no result found the system alert the actor

Table 3.4 Classify data

Use case name	Classify data
Actors	User, sentiment analyzer, python classifier
Description	This functionality enable classification of data into positive, negative and neutral
Precondition	1. The actor has been authenticated and has access to the system. 2.The functionality is available on the user

	interface
Triggers	The actor select the functionality on the interface
Basic course	• The following shall be the procedure for classifying data • Navigate through the interface and select classify • System display the function for classification • User to specify the key word to be classified • System calls python classifier • End of the use case
Post condition	Polarity of each message is stored

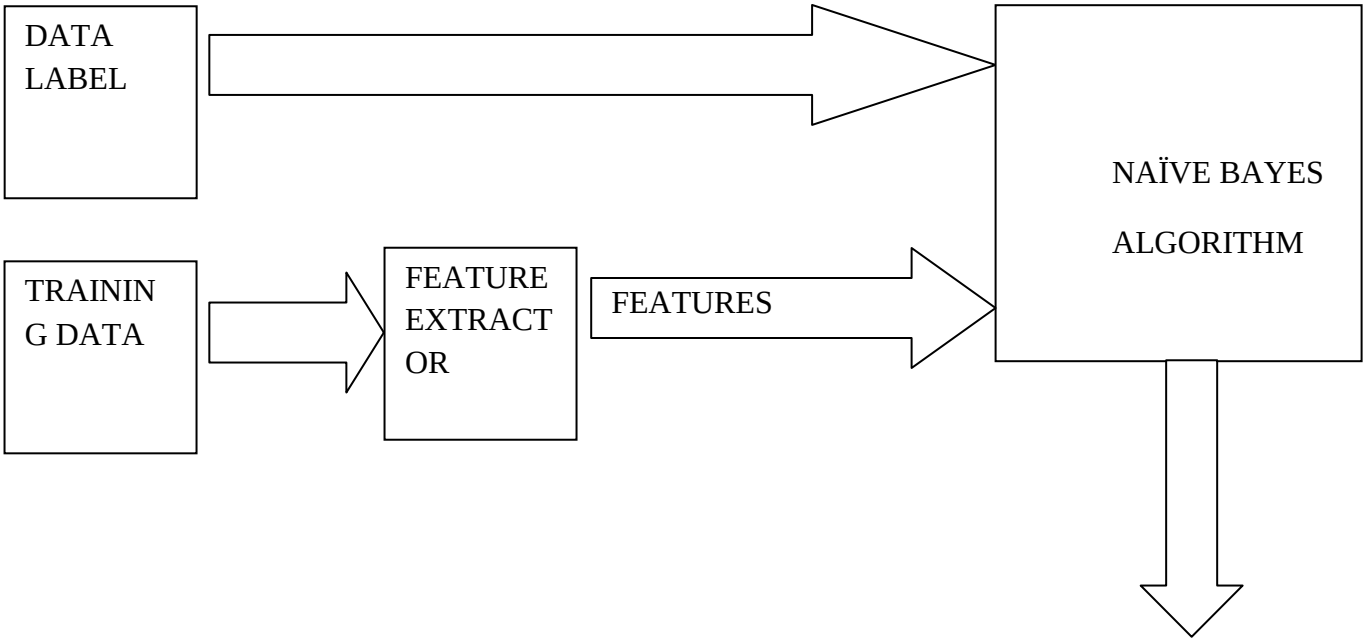
Table 3.5 View classification results or report

Use case name	View results or report
Actors	User ,sentiment analyzer
description	Allows the user to view the results
Precondition	• The actor has access to view results • The system has the functionality to view results
Trigger	The actor select view report
Basic course	The following shall be the procedure to view report • The actor navigate view report menu • The system provides the functionality through the interface • The actor select the keyword then view the report on the selected keyword

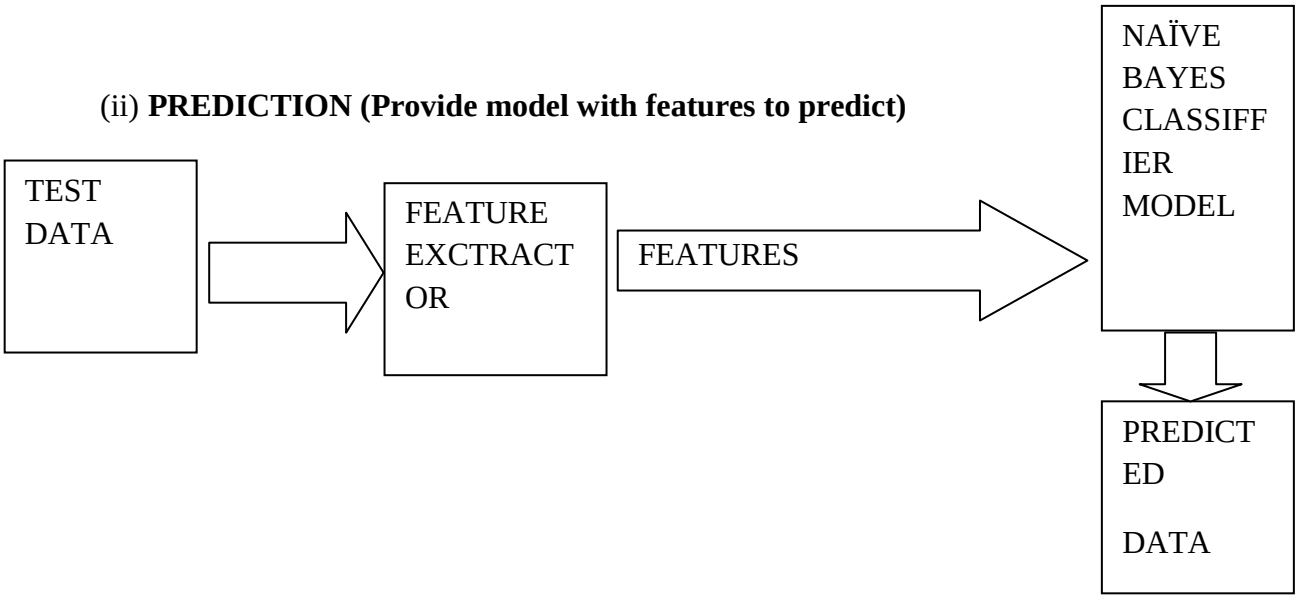
	• The system generate the results in a tabular form and a bar chart • End of the use case

(i) Feed the algorithm with label and feature pairs to build the classifier model

Figure 3.2 Classifier model



(ii) **PREDICTION (Provide model with features to predict)**



The source of data for this study is Twitter. This data includes text, emoticons and acronyms. Twitters provide millions of users a chance to express their private opinions about different issues in the society. Thus the data available on this site and other social network sites has immediate applicability in business environments such as gaining information from summarized views of users on topics (Yessenov andMisailovic, 2009).

Data for this study was collected from Twitter using APIs that are readily available and free for use in accessing data from the two social networks. It is a fact that the style and nature of writing and Twitter is not strictly formal. The data collected therefore had to go through some cleaning exercise before it was used in training the classifiers. We therefore had to remove irrelevant characters such as URL links and repeated characters. This was meant to ensure that the classifiers are trained on appropriate data for better performance.

Feature selection/extraction

In order to perform machine learning, it is necessary to extract clues from the text that may lead to correct classification (Yessenov and Misailovic, 2009). Clues about the original data are usually stored in the form of a feature vector, F = (f1; f2; : : : fn). Each coordinate of a feature vector represents one clue, also called a feature, fi of the original text. The value of the coordinate may be a binary value, indicating the presence or absence of the feature, an integer or decimal value, which may further express the intensity of the feature in the original text. In most machine learning approaches, features in a vector are considered statistically independent from each other.

The selection of features strongly influences the subsequent learning. The goal of selecting good features is to capture the desired properties of the original text in the numerical form. The study endeavored to select the properties of the original text that are relevant for the sentiment analysis task. Unfortunately, the exact algorithm for finding best features does not exist. We therefore relied on intuition, domain knowledge, and experimentation for choosing a good set of features.

Bag of words

We used a simple bag of words model to perform feature extraction. Usually opinion detection is based on the examination of adjectives in sentences though this can be misleading in cases that include sarcastic sentiments and negations. As a result we considered polarities based on sentences more than they appear based on adjectives.

Classification

This research used a classification algorithm to predict the label for a given input sentence. There are two main approaches for classification: supervised and unsupervised. In supervised classification, the classifier is trained on labeled examples that are similar to the test examples, whereas unsupervised learning techniques assign labels based only on internal differences (distances) between the data points. In this classification approach each sentence is considered independent from other sentences (Yessenov and Misailovic, 2009). The label we were interested in this project is the polarity of the sentence. We built a classifier based on the Naïve Bayes method which is a supervised classification technique, and trained it to perform classification.

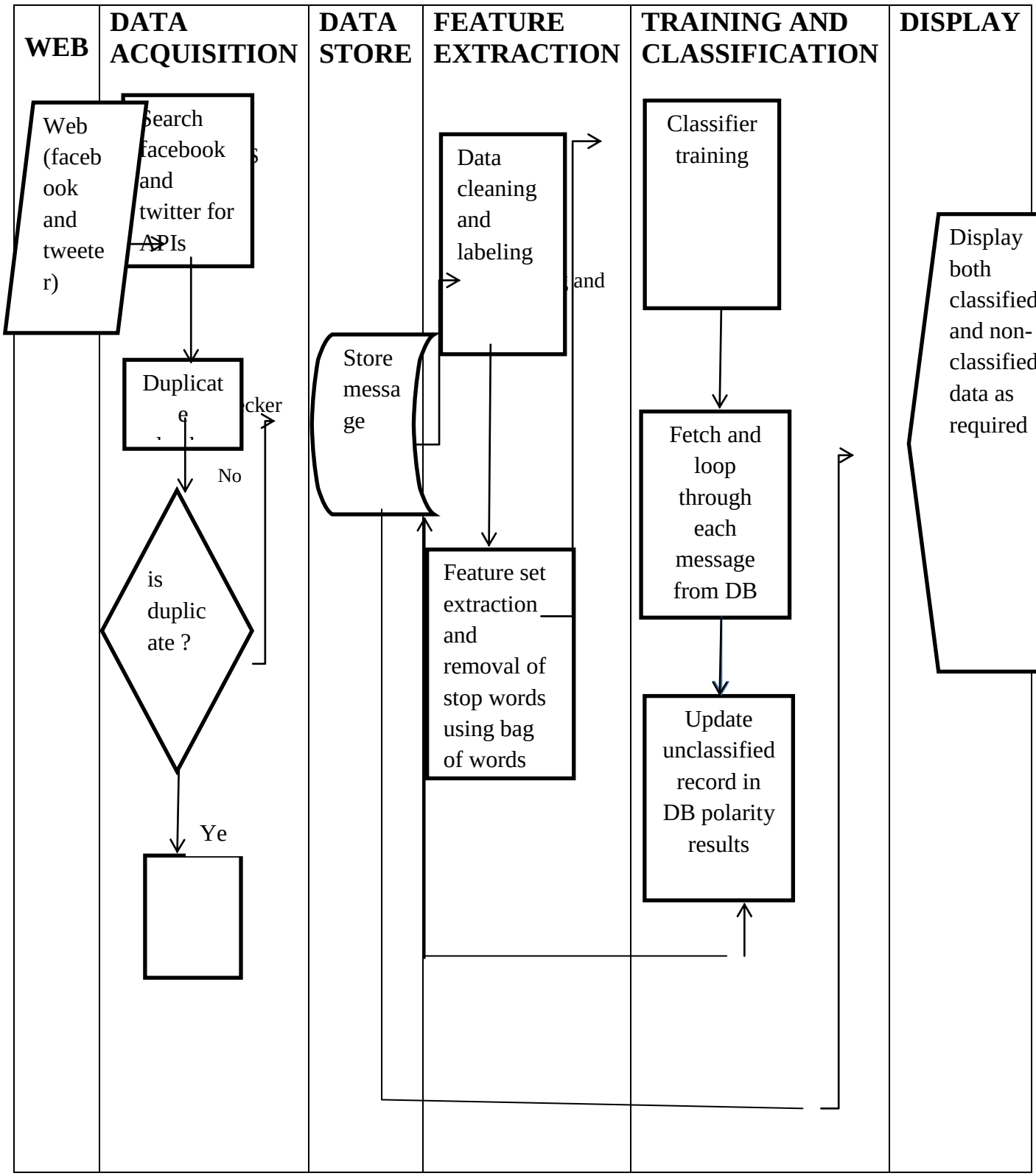
3.6 Architectural Design

3.6.1 Architectural Overview

Figure 3.3 shows an overview of the system components and the interconnections between them that enable it to perform Sentiment analysis.

Figure 3.3 Sentiment System Overview

Web: - The system interacts with the Web in order to fetch data from Twitter



Twitter APIs: - These APIs provide an interface for interaction between

Sentiment Analyzer System and the social network.

Duplicate Checker: - The purpose of this function in the Data acquisition module is to ensure there is no duplication of messages fetched from and Twitter. It uses the message Ids to enforce this check.

Data Store: - The purpose of this module is to store data fetched from Twitter

into a MySQL database.

Feature Extraction Module: - This module does extraction of feature sets which are

rendered to the classifier in 'feature - label' pairs where the feature is the word and the label is the polarity. It also removes stop words which have been found to be insignificant in the classification task. It is called into action throughout the classification process.

Training and Classification Module: - This module is charged with training a Naïve Bayes

Classifier based on the features returned by the feature extractor. The trained classifier is then used to determine the polarity of sentiments in the database.

Display Module: - This module provides visualizations for the data classification results in form of percentages and Charts.

3.7 System Implementation

The Sentiment Analyzer System was implemented as a three tier application using Java Enterprise Edition tools as given below:-

3.7.1 Front End

Java Server Faces (JSF) was used in development of forms that constitute the user

Interfaces/forms.

AJAX – AJAX was used in input validations on user interfaces

Richfaces – we also used rich face components such as text Cascading Style Sheets (CSS):- this was used in managing styles for the interfaces. An example of CSS usage is in the coloring scheme used to display the three polarities. We used red color for negative sentiments, blue for neutral while green denotes positive sentiments.

3.7.2 Application Logic/Middle tier

The business logic was implemented using Java programming Language. This involved definition of entity properties into Plain Old Java Objects (Pojos) whereas the logic was done in Java Bean classes.

3.7.3 Backend

MySQL database was used to implement backend objects which include tables, relationships, constraints and sequences. Hibernate was used to provide the requisite mapping between the database and the application logic.

3.7.4 Classifier Module

This module was implemented using Python programming language and the Natural Language Toolkit (NLTK) tools. We customized the Naïve Bayes Classifier that is available in the NLTK toolkit. Part of NLTK data was also used in providing a set of stop words. As earlier explained the feature set function checks and removes any appearance of stop words from the sentiments being classified.

CHAPTER FOUR: PROTOTYPE EVALUATION

4.0 Preamble

The main purpose of evaluation is to make conclusions about the developed prototype and to improve its effectiveness. The following criteria were used to evaluate the overall system upon completion of development. A positive answer to each of the questions below confirmed the success of this research.

Is the tool able to collect training data from both English and Swahili?

Is the chosen method of classification ideal for the kind of data that has been collected?

Once the data is collected, is the classifier able to be trained?

Are we using the correct features for classification?

Are the results of the tests on the data accurate in terms of quality, or are the results unfavorably skewed towards one sentiment?

Is the Web application developed able to provide sentiment analysis on data of topical issues extracted from Twitter?

In chapter 3 various tests were run to test the working of individual component to ensure that each functions properly as a unit, these tests showed that the various units were functioning as desired.

Integration testing was also done after the unit testing to verify that all elements of the prototype mesh properly and that the overall system function/performance had been achieved. From the results obtained it was noted that the design objectives of the project had been achieved since for every input data was being classified correctly.

When all these tests had been done, it was now time to do the prototype evaluation. It was used for overall testing of the prototype after the end of the development process, to verify that the developed prototype meets requirements or to identify differences between expected and actual results. Black-box testing approaches in which test data are derived from the specified functional requirements without regard to the final program structure were used exclusively during evaluation.

In order to carry out reliable evaluation of the prototype, several rules were followed. One was to

separate the training and test data sets. The training set (seen data) was used to build the prototype and the test set (unseen data) was used to measure the prototype's performance, at this point no changes were done to the knowledge base or any modules of the prototype. This made it possible to obtain predictive accuracy figures for the prototype, and helped to avoid over fitting on a particular test data set.

Later on, an open discussion was held to let the human experts justify their solutions.

Standards for evaluation metrics were used to evaluate the performance of the sent analyzer prototype. One of the standards, precision is a measure of the correctness of the prototype's output to that of the manual system to classify the same questions,that is how much of the information that the system returned is actually correct, this was defined as shown in equation 4.1.

$$Precision = \frac{True\ positives\ (t_p)}{True\ positives\ (t_p) + False\ positive\ (f_p)} \ldots\ldots\ldots\ldots equation\ (4.1)$$

The other standard used to evaluate the prototype is recall, which measures how much relevant information is extracted; this was defined as shown in equation 4.2

$$Recall = \frac{True\ positives\ (t_p)}{True\ positives\ (t_p) + false\ negatives\ (f_n)} \ldots\ldots\ldots\ldots equation\ (4.2)$$

Accuracy was also another performance measure used to evaluate the prototype, accuracy refers to the proximity of measurement results to the true value, and this was defined as shown in equation 4.3

$$Accuracy = \frac{True\ positives\ (t_p) + True\ negatives\ (t_n)}{True\ positives\ (t_p) + True\ negatives\ (t_n) + False\ positives\ (f_p) + false\ negatives\ (f_n)}$$

$$\ldots\ldots\ldots\ldots equation\ (4.3)$$

Where

True positives **(t_p)** refer to the number of questions correctly labeled as belonging to a particular subtopic,that is correctly identified.

false positives**(f_p)** refers to the number of questions incorrectly labeled as belonging to a particular subtopic, that is incorrectly identified.

false negatives**(f_n),** refers to the number of questions that were not labeled as belonging to a

particular subtopic but should have been, that is incorrectly rejected.

True negatives**(t_n)** refers to the number of questions that were not labeled as belonging to a particular subtopic and should not have been labeled, that is correctly rejected.

4.1 Data collection for the test

Test data was data that had not been used during the process of developing the prototype. It was obtained from three main sources; some data was collected from the social media in specific tweeter using the tweeter API The most import criteria of these test cases are that they cover normal cases, as well as the most standard, and rare cases.

4.2 Cleaning the data

Data cleaning was done to make a data set consistent with other similar data sets in the system. Once the data for testing was collected the data that was used in purpose sampling was cleaned. Cleaning involved elimination through deletion of incomplete words of typographical errors, ambiguous words were also eliminated.

Data cleaning also involved removal of stop words and also harmonizing words that seems to have multiple meanings.

4.3 Data sampling

Data from tweeter was generated divided into two, one part of the sentence was tested in its raw form while the second part had purpose sampling employed.

For the words obtained from Internet blogs on topics of interest, the test data was classified into two

(i) Some words were tested in their raw form without eliminating any element

(ii) Purpose sampling was employed. Ambiguous words were also eliminated, those with multiple meanings.

4.4 Mode of analysis

The mode of analysis is deductive based on descriptive statistics method. The performance standards will be considered for the various sets and comparisons then made between the accuracy of placement based on data from the various sources.

4.5 Results

The following results were obtained using the sent analyzer. The results were computed in line with the performance standards to be used, we relied on the NLTK metrics module which provides functions for calculating accuracy, precision and recall for the Classifier.

4.5.1 Summary of the results of the data from tweeter

The data from the internet had been divided into two sets, the results that compared the results of the performance of the prototype on data obtained from the Internet were as shown in table 4.1.

Test runs and Presentation of Results

We performed a number of tests whose results are as tabulated below:- Test runs with Naïve Bayes Classifier.

The levels of tests are defined below.

- Level one: The test done in five trials
- Level two: The test done in ten trials
- Level three: The test done in fifteen trials

Table 4.1: Effects of increasing the no of training on unbalanced keywords (Positive 840, Negative 620, Neutral 430)

NO	EXPERIMENT	CLASSIFIER ACCURACY %	POS PRECISION %	POS PRECISION RECALL %	NEG PRECISION RECALL %	NEG PRECISION RECALL %	NEU PRECISION RECALL %	NEU PRECISION RECALL %
1	Classification from level 1 training	54.2	64.6	75.4	63.6	95.1	59	61.5

2	Classification from level 2 training	61.2	76	79.1	69.5	95.2	63.2	71.2
3	Classification from level 3 training	77.3	77.9	80.3	77.3	95.5	68.4	79.3

The levels of tests are defined below.

- Level one: The test done in five trials
- Level two: The test done in ten trials
- Level three: The test done in fifteen trials

Table 4.2: Effects of increasing the no of training on balanced corpus (Positive 360,Negative 360,Neutral 360)

NO	Experiment	Classifier accuracy %	Pos precision %	Pos precision recall %	Neg precision recall %	Neg precision recall %	Neg precision recall %	Neg precision recall %
1	Classification from level 1 training	64.2	65	64.2	57	62	73.2	63
2	Classification from level 2 training	66.6	67	65.7	61.9	65.8	76.2	74.2
3	Classification from level 3 training	79.8	68.2	71.6	73.3	75.2	77.6	78.3

4.6 Sentiment Analyzer System results

The sentiment analyzer system was tested using test cases (see above). All the system features were observed to work as stipulated in the design.

The above test were done to check the effects of running both balanced and unbalanced corpus on the sent analyzer and from the results we are able to see an improvement when the corpus is done in level one, level two and finally level three we can deduce from the results of the test that as more training is done in the extracted data the level of accuracy also goes up as we can see in the unbalanced from level one is 64.2,level two 66.6 and level three it is 79.8 in all this there is an improvement

4.7 Discussion of results

The above test were done to check the effects of running both balanced and unbalanced corpus on the sent analyzer and from the results we are able to see an improvement when the corpus is done in level one, level two and finally level three we can deduce from the results of the test that as more training is done in the extracted data the level of accuracy also goes up as we can see in the unbalanced from level one is 64.2,level two 66.6 and level three it is 79.8 in all this there is an improvement,

We also interpret the following from an accuracy of 78.9%, Positive precision of 68.2%, Positive recall of 73.6%, Negative precision of 85.1%, Negative recall of 95.5%, Neutral Precision of 76.6% and Neutral recall of 40.9 %:-1. A sentiment given a positive classification is only 68.2% likely to be correct. This precision leads to 31.8% false positives for the positive label.73.6% positive recall means that there is a likelihood of 26.4% of getting false negatives and false neutrals in the positive class. A sentiment identified as negative is 85.1% likely to be correct. This means a 14.9% likelihood of having false positives and neutrals in the negative class. A negative recall of 95.5% means that many sentiments that are negative are correctly classified. This implies very few false negatives for the negative category. A sentiment given a neutral classification is only 76.6% likely to be correct. This precision leads to 23.4% false neutrals for the neutral label. A neutral recall of 40.9% means that there is a high likelihood of 59.1% of getting false negatives and false positives in the neutral class.

One possible explanation for the above results is that people use normally positive words

in negative reviews, but the words are preceded by a negating word such as "not" (or some other negative word), such as "not great". And since the classifier uses the bag of words model, which assumes every word is independent, it cannot learn that "not great" is a negative.

CHAPTER FIVE: CONCLUSION AND FUTURE WORD

5.0 Conclusion

In this research work we were able to build an application that has the ability to fetch and store data searched on various topics from the most popular social media site Twitter. A number of classification models were considered as specified in the literature review out of which we chose to use the Naïve Bayes classifier model because of its simplicity in adapting it to the data collected. We developed a Naïve Bayes classifier that integrates an information gain heuristic using the Natural Language Tool Kit and trained it on a preprocessed dataset from Social Media.

The results obtained from experiments with the classifier (see Chapter 4 above) show that the classifier is capable of performing classification with an accuracy of 78.9% for sentiments obtained from Social Media. This is near human accuracy, as apparently people agree on sentiment only around 80% of the time. Most of the sentiments in this data are expressed partly in English, Swahili, Slang and Sheng thus formal language is scarcely used. We therefore conclude that the model of classification selected is ideal for the kind of data collected from social media on Kenyan opinions. Finally, we integrate the techniques and methods developed into a web based application for use in providing Sentiment Analysis with respect to opinions from Kenya social media users. The application provides user friendly features and its architecture can also be used in viewing results of other classification approaches.

5.1 Limitations Faced

A number of limitations were faced as listed:-

1. We were not able to fetch more information associated with sentiments fetched from social media due to restrictions that prevent running of queries that have joins.

2. While fetching data, the requests are limited to a time span of 30 seconds beyond which a fetch operation is timed out and disconnected. This limits the amount of data that can be fetched on a search keyword.

3. The use of positive words preceded by negations such as ‘not’ in negative sentiments leads to erroneous classifications since the classifier uses the bag of words model, which assumes every word is independent. It cannot therefore learn that "not great" is a negative.

4. Based on the data obtained from Kenyan opinions we observed that the language used is mostly slang. Kenyan slang is constantly changing coupled with the idea of lack of inadequate literature on it since it is informal. This made it challenging during preparation of training data and also during classification.

5.2 Future Work

The following improvements can be done on the Sentiment Analyzer System:-

1. Functionalities for accommodating other classifiers other than the Naïve Bayes classifier can be developed into the application. These classifiers include Decision trees and Support Vector Machines. Results from the various classifiers can be compared in a report interface for the best classification technique to be selected.

2. Training on multiple words can also be explored to resolve the limitation.

3. The application can also be enhanced to be accessible on handheld devices such as mobile phones.

4. Integration with multi and cross-lingual language dictionaries to cater for the dynamic

natureof language used on social media.

REFERENCES:

Abbasi, A., Chen, H., Salem, A. (2008). Sentiment analysis in multiple languages:

Abbasi, A., Chen, H., Salem, A. (2008). Sentiment analysis in multiple languages:

ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27.

Approach to Parsing, *in proceedings of IWPT.*

Attributes selection for opinion classification in Web forums, *ACM Transactions on*

attributes vectors, and the role of linguistic analysis, *Proceedings of the 20th international*

Blache, P., Balfourier, J.-M. (2001). Property Grammars: a Flexible Constraint-Based

Blei,D., Ng,A., and M. Jordan (2003). Latent dirichlet allocation, *Journal of Machine*

Blei,D., Ng,A., and M. Jordan (2003). Latent dirichlet allocation, *Journal of Machine*

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM : a library for support vector machines,

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying sources of

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying sources of

conference on Computational Linguistics, p.841-es, August 23-27, 2004, Geneva,

Feature selection for opinion classification in Web forums, *ACM Transactions on*

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007). Large-scale sentiment analysis for

Information Systems (TOIS), Volume 26 Issue 3.

Information Systems (TOIS), Volume 26 Issue 3.

Learning Research, 3(5):993–1022.

Learning Research, 3(5):993–1022.

news and blogs*, In CWSM '07.*

opinions with conditional random fields and extraction patterns, *Human Language*

opinions with conditional random fields and mining models, *Human Language*

Processing (HLT-EMNLP 2005).

Processing (HLT-EMNLP 2005).

Switzerland.

Technology Conference/Conference on Empirical Methods in Natural Language

Technology Conference/Conference on Empirical Methods in Natural Language

APPENDIX

Figure 1 Login screen

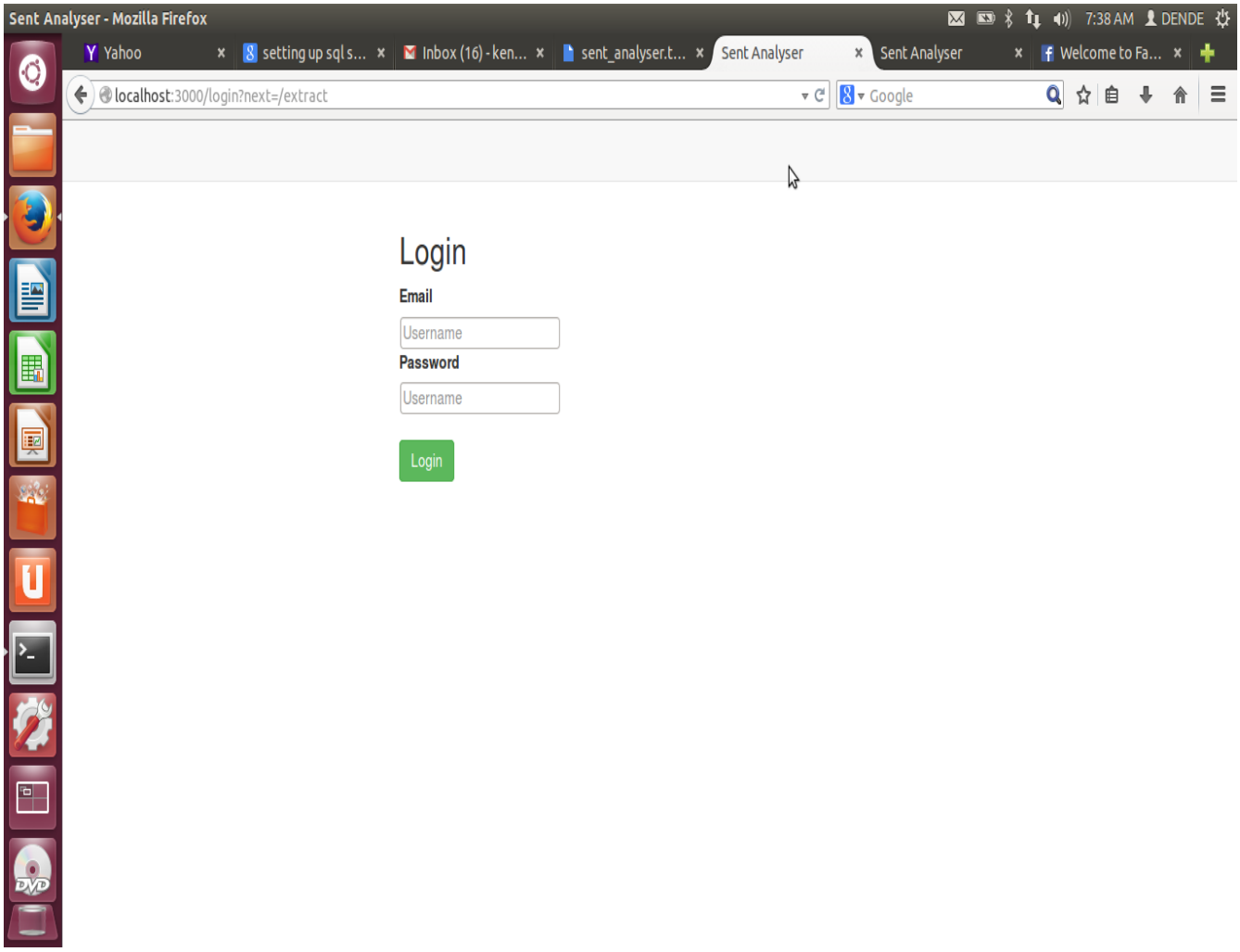
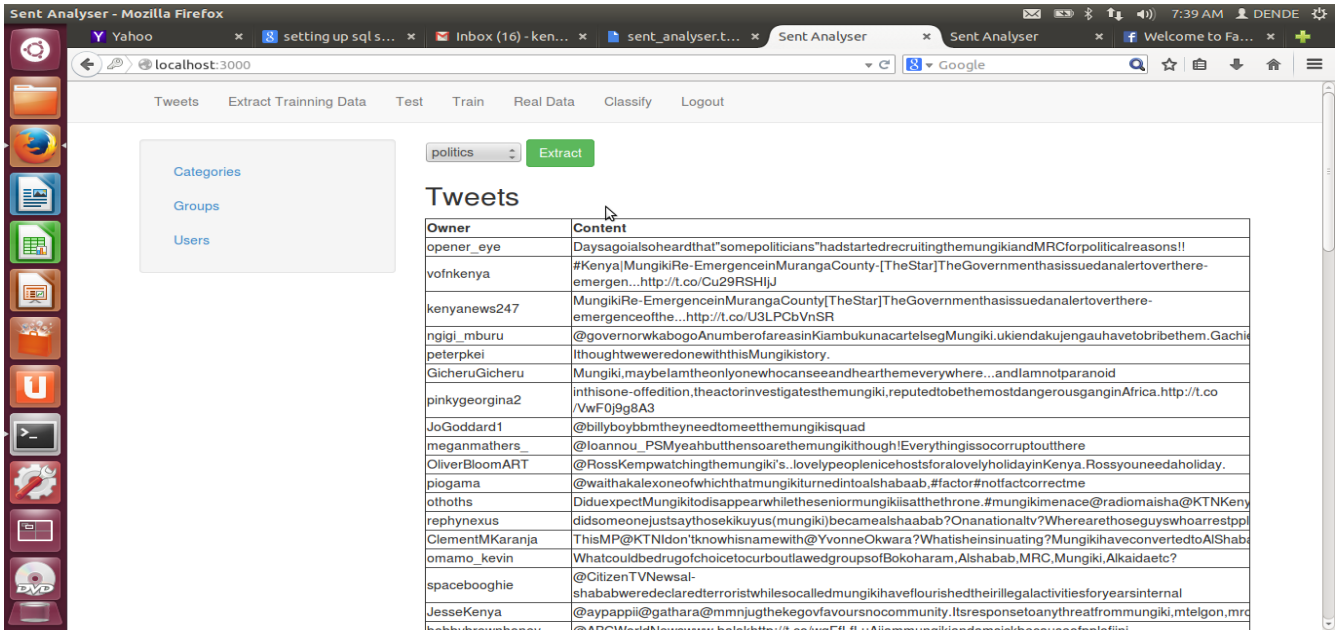


Figure 2 Extraction of the data from Tweeter



Owner	Content
opener_eye	Daysagoialsoheardthat"somepoliticians"hadstartedrecruitingthemungikiandMRCforpoliticalreasons!!
vofnkenya	#Kenya|MungikiRe-EmergenceinMurangaCounty-[TheStar]TheGovernmenthasissuedanalertoverthere-emergen...http://t.co/Cu29RSHIjJ
kenyanews247	MungikiRe-EmergenceinMurangaCounty[TheStar]TheGovernmenthasissuedanalertoverthere-emergenceofthe...http://t.co/U3LPCbVnSR
ngigi_mburu	@governorwkabogoAnumberofareasinKiambukunacartelsegMungiki.ukiendakujengauhavetobribethem.Gachi
peterpkei	IthoughtweweredonewiththisMungikistory.
GicheruGicheru	Mungiki,maybeIamtheonlyonewhocanseeandhearthemeverywhere...andIamnotparanoid
pinkygeorgina2	inthisone-offedition,theactorinvestigatesthemungiki,reputedtobethemostdangerousganginAfrica.http://t.co/VwF0j9g8A3
JoGoddard1	@billyboybbmtheyneedtomeetthemungikisquad
meganmathers_	@loannou_PSMyeahbutthensoarethemungikithough!Everythingissocorruptoutthere
OliverBloomART	@RossKempwatchingthemungiki's..lovelypeoplenicehostsforalovelyholidayinKenya.Rossyouneedaholiday.
piogama	@waithakalexoneofwhichthatmungikiturnedintoalshabaab,#factor#notfactcorrectme
othoths	DiduexpectMungikitodisappearwhiletheseniormungikiisatthethrone.#mungikimenace@radiomaisha@KTNKeny
rephynexus	didsomeonejustsaythosekikuyus(mungiki)becamealshaabab?Onanationaltv?Wherearethoseguyswhoarrestppl
ClementMKaranja	ThisMP@KTNIdon'tknowhisnamewith@YvonneOkwara?Whatisheinsinuating?MungikihaveconvertedtoAlShab
omamo_kevin	WhatcouldbedrugofchoicetocurboutlawedgroupsofBokoharam,Alshabab,MRC,Mungiki,Alkaidaetc?
spaceboochie	@CitizenTVNewsal-shababweredeclaredterroristwhilesocalledmungikihaveflourishedtheirillegalactivitiesforyearsinternal
JesseKenya	@aypappii@gathara@mmnjugthekegovfavoursnocommunity.Itsresponsetoanythreatfrommungiki,mtelgon,mrc

SAMPLE CODES FOR SENT ANALYZER

```
"""
Django settings for twitter_app project.

For more information on this file, see
https://docs.djangoproject.com/en/1.6/topics/settings/

For the full list of settings and their values, see
https://docs.djangoproject.com/en/1.6/ref/settings/
"""

# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
import os
#BASE_DIR = os.path.dirname(os.path.dirname(__file__))
BASE_DIR = '/home/dende/apps/sent_analyser/sent_analyser/'

# Quick-start development settings - unsuitable for production
# See https://docs.djangoproject.com/en/1.6/howto/deployment/checklist/

# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = 'kq0n!(plkqffa@+n96rjq*dog@o=b^-82^kh+f4@9c3#px6wii'

# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True

TEMPLATE_DEBUG = True

ALLOWED_HOSTS = []


# Application definition

INSTALLED_APPS = (
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'sent_analyser',
    'app',
)

MIDDLEWARE_CLASSES = (
    'django.contrib.sessions.middleware.SessionMiddleware',
```

```
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
    'django.middleware.clickjacking.XFrameOptionsMiddleware',
)

ROOT_URLCONF = 'sent_analyser.urls'

WSGI_APPLICATION = 'sent_analyser.wsgi.application'


# Database
# https://docs.djangoproject.com/en/1.6/ref/settings/#databases

DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'sent_analyser',
        'USER': 'root',
        'PASSWORD': 'r00t',
        'HOST': '127.0.0.1',
        'PORT': '3306',
    }
}

# Internationalization
# https://docs.djangoproject.com/en/1.6/topics/i18n/

LANGUAGE_CODE = 'en-us'

TIME_ZONE = 'UTC'

USE_I18N = True

USE_L10N = True

USE_TZ = True


# Static files (CSS, JavaScript, Images)
# https://docs.djangoproject.com/en/1.6/howto/static-files/

# Absolute filesystem path to the directory that will hold user-uploaded files.
# Example: "/var/www/example.com/media/"
MEDIA_ROOT = BASE_DIR + 'media/'
```

```
# URL that handles the media served from MEDIA_ROOT. Make sure to use a
# trailing slash.
# Examples: "http://example.com/media/", "http://media.example.com/"
MEDIA_URL = '/media/'

# Absolute path to the directory static files should be collected to.
# Don't put anything in this directory yourself; store your static files
# in apps' "static/" subdirectories and in STATICFILES_DIRS.
# Example: "/var/www/example.com/static/"
STATIC_ROOT = BASE_DIR + 'static/'

# URL prefix for static files.
# Example: "http://example.com/static/", "http://static.example.com/"
STATIC_URL = '/static/'

TEMPLATE_DIRS = (
      BASE_DIR + 'templates/',
   # Put strings here, like "/home/html/django_templates" or
"C:/www/django/templates".
   # Always use forward slashes, even on Windows.
   # Don't forget to use absolute paths, not relative paths.

)
LOGIN_URL = '/login'

STATICFILES_DIRS = (
   '/home/dende/apps/sent_analyser/static/',
)
# Context processors
TEMPLATE_CONTEXT_PROCESSORS = (
   'django.contrib.auth.context_processors.auth',
   'django.core.context_processors.debug',
   'django.core.context_processors.i18n',
   'django.core.context_processors.media',
   'django.core.context_processors.static',
   'django.core.context_processors.request',
   'django.contrib.messages.context_processors.messages',
)
```

creating user-uploaded

```
from django.conf.urls import patterns, include, url
from django.contrib.staticfiles.urls import staticfiles_urlpatterns
```

```
from django.contrib import admin
from app.views import *
admin.autodiscover()

urlpatterns = patterns('',
    # Examples:
    # url(r'^$', 'twitter_app.views.home', name='home'),
    # url(r'^blog/', include('blog.urls')),
    url(r'^$', Tweets.as_view()),
    url(r'^login$', Login.as_view()),
    url(r'^logout$', Logout.as_view()),

    url(r'^user$', CreateUser.as_view()),
    url(r'^tweets$', Tweets.as_view()),
    url(r'^extract$', ExtractTweets.as_view()),
    url(r'^categories$', Categories.as_view()),
    url(r'^groups$', UserGroups.as_view()),
    url(r'^categorized$', Categorize.as_view()),
    url(r'^learn$', Train.as_view()),
    url(r'^admin/', include(admin.site.urls)),
    url(r'^report$', Report.as_view()),

)

urlpatterns += staticfiles_urlpatterns()
```

configuration of tweeter

```
"""
WSGI config for twitter_app project.

It exposes the WSGI callable as a module-level variable named ``application``.

For more information on this file, see
https://docs.djangoproject.com/en/1.6/howto/deployment/wsgi/
"""

import os
os.environ.setdefault("DJANGO_SETTINGS_MODULE",
"sent_analyser.settings")

from django.core.wsgi import get_wsgi_application
application = get_wsgi_application()
```

forms

```
@import url('widgets.css');

/* FORM ROWS */

.form-row {
   overflow: hidden;
   padding: 8px 12px;
   font-size: 11px;
   border-bottom: 1px solid #eee;
}

.form-row img, .form-row input {
   vertical-align: middle;
}

form .form-row p {
   padding-left: 0;
   font-size: 11px;
}

/* FORM LABELS */

form h4 {
   margin: 0 !important;
   padding: 0 !important;
   border: none !important;
}

label {
   font-weight: normal !important;
   color: #666;
   font-size: 12px;
}

.required label, label.required {
   font-weight: bold !important;
   color: #333 !important;
}

/* RADIO BUTTONS */

form ul.radiolist li {
   list-style-type: none;
}
```

```
form ul.radiolist label {
   float: none;
   display: inline;
}

form ul.inline {
   margin-left: 0;
   padding: 0;
}

form ul.inline li {
   float: left;
   padding-right: 7px;
}

/* ALIGNED FIELDSETS */

.aligned label {
   display: block;
   padding: 3px 10px 0 0;
   float: left;
   width: 8em;
   word-wrap: break-word;
}

.aligned ul label {
   display: inline;
   float: none;
   width: auto;
}

.colMS .aligned .vLargeTextField, .colMS .aligned .vXMLLargeTextField {
   width: 350px;
}

form .aligned p, form .aligned ul {
   margin-left: 7em;
   padding-left: 30px;
}

form .aligned table p {
   margin-left: 0;
   padding-left: 0;
}
```

```
form .aligned p.help {
   padding-left: 38px;
}

.aligned .vCheckboxLabel {
   float: none !important;
   display: inline;
   padding-left: 4px;
}

.colM .aligned .vLargeTextField, .colM .aligned .vXMLLargeTextField {
   width: 610px;
}

.checkbox-row p.help {
   margin-left: 0;
   padding-left: 0 !important;
}

fieldset .field-box {
   float: left;
   margin-right: 20px;
}

/* WIDE FIELDSETS */

.wide label {
   width: 15em !important;
}

form .wide p {
   margin-left: 15em;
}

form .wide p.help {
   padding-left: 38px;
}

.colM fieldset.wide .vLargeTextField, .colM fieldset.wide .vXMLLargeTextField
{
   width: 450px;
}

/* COLLAPSED FIELDSETS */

fieldset.collapsed * {
```

```
    display: none;
}

fieldset.collapsed h2, fieldset.collapsed {
    display: block !important;
}

fieldset.collapsed h2 {
    background-image: url(../img/nav-bg.gif);
    background-position: bottom left;
    color: #999;
}

fieldset.collapsed .collapse-toggle {
    background: transparent;
    display: inline !important;
}

/* MONOSPACE TEXTAREAS */

fieldset.monospace textarea {
    font-family: "Bitstream Vera Sans Mono",Monaco,"Courier
New",Courier,monospace;
}

/* SUBMIT ROW */

.submit-row {
    padding: 5px 7px;
    text-align: right;
    background: white url(../img/nav-bg.gif) 0 100% repeat-x;
    border: 1px solid #ccc;
    margin: 5px 0;
    overflow: hidden;
}

body.popup .submit-row {
    overflow: auto;
}

.submit-row input {
    margin: 0 0 0 5px;
}

.submit-row p {
    margin: 0.3em;
```

```
}

.submit-row p.deletelink-box {
    float: left;
}

.submit-row .deletelink {
    background: url(../img/icon_deletelink.gif) 0 50% no-repeat;
    padding-left: 14px;
}

/* CUSTOM FORM FIELDS */

.vSelectMultipleField {
    vertical-align: top !important;
}

.vCheckboxField {
    border: none;
}

.vDateField, .vTimeField {
    margin-right: 2px;
}

.vURLField {
    width: 30em;
}

.vLargeTextField, .vXMLLargeTextField {
    width: 48em;
}

.flatpages-flatpage #id_content {
    height: 40.2em;
}

.module table .vPositiveSmallIntegerField {
    width: 2.2em;
}

.vTextField {
    width: 20em;
}

.vIntegerField {
```

```
    width: 5em;
}

.vBigIntegerField {
    width: 10em;
}

.vForeignKeyRawIdAdminField {
    width: 5em;
}

/* INLINES */

.inline-group {
    padding: 0;
    border: 1px solid #ccc;
    margin: 10px 0;
}

.inline-group .aligned label {
    width: 8em;
}

.inline-related {
    position: relative;
}

.inline-related h3 {
    margin: 0;
    color: #666;
    padding: 3px 5px;
    font-size: 11px;
    background: #e1e1e1 url(../img/nav-bg.gif) top left repeat-x;
    border-bottom: 1px solid #ddd;
}

.inline-related h3 span.delete {
    float: right;
}

.inline-related h3 span.delete label {
    margin-left: 2px;
    font-size: 11px;
}

.inline-related fieldset {
```

```
    margin: 0;
    background: #fff;
    border: none;
    width: 100%;
}

.inline-related fieldset.module h3 {
    margin: 0;
    padding: 2px 5px 3px 5px;
    font-size: 11px;
    text-align: left;
    font-weight: bold;
    background: #bcd;
    color: #fff;
}

.inline-group .tabular fieldset.module {
    border: none;
    border-bottom: 1px solid #ddd;
}

.inline-related.tabular fieldset.module table {
    width: 100%;
}

.last-related fieldset {
    border: none;
}

.inline-group .tabular tr.has_original td {
    padding-top: 2em;
}

.inline-group .tabular tr td.original {
    padding: 2px 0 0 0;
    width: 0;
    _position: relative;
}

.inline-group .tabular th.original {
    width: 0px;
    padding: 0;
}

.inline-group .tabular td.original p {
    position: absolute;
```

```
    left: 0;
    height: 1.1em;
    padding: 2px 7px;
    overflow: hidden;
    font-size: 9px;
    font-weight: bold;
    color: #666;
    _width: 700px;
}

.inline-group ul.tools {
    padding: 0;
    margin: 0;
    list-style: none;
}

.inline-group ul.tools li {
    display: inline;
    padding: 0 5px;
}

.inline-group div.add-row,
.inline-group .tabular tr.add-row td {
    color: #666;
    padding: 3px 5px;
    border-bottom: 1px solid #ddd;
    background: #e1e1e1 url(../img/nav-bg.gif) top left repeat-x;
}

.inline-group .tabular tr.add-row td {
    padding: 4px 5px 3px;
    border-bottom: none;
}

.inline-group ul.tools a.add,
.inline-group div.add-row a,
.inline-group .tabular tr.add-row td a {
    background: url(../img/icon_addlink.gif) 0 50% no-repeat;
    padding-left: 14px;
    font-size: 11px;
    outline: 0; /* Remove dotted border around link */
}

.empty-form {
    display: none;
}
```