UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A SOCIAL MEDIA SENTIMENT ANALYSIS MODEL TO SUPPORT MARKETING INTELLIGENCE IN KENYA.

KIPTANUI DENNIS TOO

P52/65294/2013

SUPERVISOR: DR. LAWRENCE MUCHEMI

A research project report submitted in partial fulfillment of the requirements of the Degree of Master of Science in Computational Intelligence at the University of Nairobi.

June, 2015

DECLARATION

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

Name: Kiptanui Dennis Too **Reg. No:** P52/65294/2013

Signature: _____ **Date:**_____

This project has been submitted with the approval as University supervisor

Name: Dr. Lawrence Muchemi

Sign: _____**Date:** _____

ABSTRACT

Over a decade ago, what commenced as a collection of individual musings scattered across the internet has since evolved into the de facto voice of the global public. It is called social media. The use of social media has brought about the biggest shift of how we gather and respond to information since the advent of internet itself. The sentiments expressed therein have led to undeniable influence in changing the world around.

What do tweets, blogs and posts about your products and services tell you about what they think and feel? These will definitely influence your sales and other KPIs. Every business needs to be able to meticulously translate these social media data for marketing guidance. This can give competitive advantage in terms of early trend detection, alarming on emerging issues and monitoring on competitors activities among others.

Posting reviews online has become an increasingly popular way for people to express opinions and sentiments toward the products bought or services received. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic value in terms of marketing intelligence.

This study, in the literature, investigates the different available social media platforms and sentiment analysis techniques together with approaches to combine several classifiers. The main aim of this study is building a sentiment classifier to classify twitter opinions as positive, negative or neutral. The classifier is specifically used in a business setting for marketing intelligence. This involves analyzing the business products, services, brand and presence. The information yields very useful information and insights for marketing strategies. Our review of literature indicated that support vector machine (SVM) was generally the most accurate machine learning classifier for sentiment analysis and robust on large feature spaces. The study employs a combination of SVM with Naïve Bayes algorithms using the ensemble approach to enhance the overall performance. Model validation is investigated and a prototype is constructed for output presentation.

ACKNOWLEDGEMENT

I express my uttermost gratitude to Jehovah God Almighty for His guidance, help and provision throughout my study. He is my Rock and fortress.

To my supervisor Dr Lawrence Muchemi, I am deeply indebted to you for all the help, suggestions and encouragement I received during the research and when writing this report.

To Mr Isaac Kirwa my father, Mrs Jane Cynthia Kirwa my mother, my brother Emmanuel and my sister Chebet for all their support and prayers during my study.

To Rev Calisto Odede for his encouragement and prayers.

To all my friends including the Njeris for their insights, encouragement and prayers.

Table of Contents

LIST OF FIGURES

LIST OF EQUATIONS

List of Table

ACRONYMS

W-O-M ï Word Of Mouth.

KPI ï Key Performance Indicator.

NLP ï Natural language processing.

VOC ï Voice Of Customer.

ML ï Machine Learning.

SVM ï Support Vector Machines.

MaxEnt ï Maximum Entropy.

NB ï Naïve Bayes.

KNN ï K-Nearest Neighbor.

TF-IDF ï Term Frequency Inverse Document Frequency.

API ï Application Programming Interface.

POS ï Part Of Speech.

IDC ï International Data Corporation.

PMI ï Pointwise Mutual Information.

IMC ï Integrated Market Communication.

LIWC ï Linguistic Inquiry and Word Count.

MPQA ï Multi-perspective Question Answering.

CSV ï Comma Separated Values.

CHAPTER ONE

INTRODUCTION

1.1 Background

Humans are innately social beings and we continually seek ways to keep in touch with each other. Besides a myriad of reasons, one that is very obvious is that we actually value and rely on each otherôs opinions. This is on various aspects of life. Alexander Graham Bell invention of the telephone in the 19th century was just a steppingstone in giving people a means to communicate even when they are geographically dispersed. Technology has further endeavored to brace this natural human act of providing even improved interaction. The discovery of the internet in the late 20th century which is a global system of interconnected computer networks that link billion of devices worldwide via the internet protocol suite brought about new dimensions and capabilities. Web 2.0 which is a furtherance of the World Wide Web (WWW) allows users to not only read content but also input user generated content.

Social media is hot and an integral part of todayôs web ecosystem providing numerous opportunities for applications with significant social and economic impact (Valkanas et al, 2013). User generated content is its fount creating virtual communities in line with the societal nature of humankind. Social networks such as Facebook, Twitter, LinkedIn and YouTube are now the trend. According to digital insights release on social media 2014, Facebook has more than 1.28 billion active users every month and twitter about 1 billion users. These users share and discuss any and all kinds of information: political, academic, economic and social. For businesses it represents a marketing opportunity that transcends the traditional middleman and connects companies directly with customers (Sisira et al, 2011). Information is passed from one social network user to others and thus electronic word-of-mouth (E-w-o-m). W-o-m communications is a pervasive and intriguing phenomenon even in business. It has been generally found that both satisfied and dissatisfied customers tend to spread positive and negative w-o-m respectively regarding products and services which they purchase and use (Anderson, cited in Muller et al 2001).

As social media gives people the liberty to air their opinions, social sites are an endless stream of experiences, opinions and sentiments. Sentiment analysis is the computational study of how opinions, attitudes, emotions, and perspectives are expressed in language. Sentiment analysis strives to make use of the words, how they write them and emoticons to automatically deduce the authorsô feelings swiftly and more accurately.

Marketing intelligence discusses about changes of marketing environment to help preparation and adoption of marketing plans (Ghasemi et al, 2013). It is an area in business where sentiment analysis can have grand significance. ñWhat other people thinkò has always been an important piece of information during the decision making process (Sachin et al., 2013). Marketing Intelligence is used to develop decision-oriented and functional solutions for market management. In essence, it can be used to assess market entry opportunities and to formulate market development plans and penetration strategies. Bruce Temkin (2008) believes that in the next few years, companies will be using sentiment analysis as it is a critical tool.Kenya is experiencing digital evolution. The current trend in marketing intelligence though depends largely on surveys and other business data.

Business has been an integral part of day to day activities since time immemorial. Most businesses today strive to employ some form of business intelligence applications in their operations. This is to aid in decision making process by keeping track of internal processes and company data. How do they keep tabs on the completion or analyze what is happening on their industry or markets where they do business? Business intelligence is not very effective when it lacks context. Marketing intelligence provides this context because it demonstrates what is going on in the market as a whole.

Marketing is a unique function of business whose concern and responsibility must permeate all areas of the enterprise. One of the shortest deýnitions of marketing is ñmeeting needs proýtablyò (Philip Kotler, 2002). In business, change is occurring at an accelerating rate; today is not like yesterday, and tomorrow will be different from today. Marketing has the prime task of keeping enterprises effectively in business. To realize these changes in time or well in advance, enterprises need to always stay in touch with its stakeholders. The ubiquitous nature of social media has provided a salient platform to monitor company relations at various facets of the market. The enormous volume of opinions and comments by the public towards a company, their

events, products and services create unprecedented opportunities to mine text data in real time and to analyze general sentiments. Customers voices on ongoing issues, complaints, feedback, suggestions to improve and even compliments for a service may have great value and importance (Goutam et al., 2014).

1.2 Problem Statement

Most of the data in social media is unstructured. According to IDC survey, unstructured data takes a lionôs share in digital space by occupying approximately 80 percent by volume compared to 20 percent for structured data (Goutam et al, 2014). The opportunity cost of any business to ignore unstructured data is paramount in todayôs fierce competitive world. But it is a challenge analyzing this kind of data and also quite impossible to manually classify the plethora of sentiments expressed by in social media precisely or without being biased. Data mining, Natural Language Processing and machine learning approaches have been applied to try to analyze sentiments from unstructured data.

Marketing intelligence involves primary data collected and analyzed by a business about markets that it anticipates participating in with the intention of using it in making decisions (marketing intelligence, 2013). There is no set plan for how companies gather marketing intelligence but by collecting data about the markets, they can get valuable insights into how to grow their business. Marketing intelligence can be defined as an ongoing, holistic knowledge of all aspects of the aspect place (Quirkôs marketing research, 2004).

In East Africa and specifically Kenya, the airline industry has faced a drastic increase in its dynamicity. Considering its costly status, it requires weighty attention to analyze its market. This however has not been extensively explored. The East African region is vastly populated with a myriad of ethnic backgrounds which use many different languages both local and European to express their opinions electronically. Moreover, a corpus encompassing this diversity does not exist. There is need to correctly harvest and store data specific to this problem area. A study should therefore be carried out in this airline marketspace to effectively identify challenges and hence come up with appropriate solutions.

1.3 Objectives

The general objective of this research is to build a sentiment analysis model that implements three machine learning algorithms to create a model that can be used for marketing intelligence. The algorithms are support vector machines, multinomial naïve Bayes and Bernoulli Naïve Bayes. The trained model should be able to receive extracted opinions and prevalent talking points and effectively classify them as positive, negative or neutral. These sentiment scores are precious actionable knowledge about products, services, brands and the company can provide unprecedented dimensions in decision making.

The model would complement the marketing functionality of a business by providing useful insights from social media data. In particular, it will identify positive, negative and neutral sentiments about the market.

This involves the following specific objectives:

I. Developing a technique to harvest data from social media platforms and creating appropriate repository.
II. Create and utilize a sentiment corpus to include commonly used phrases and words within the Kenyan blogosphere.
III. Apply a combination of support vector machine learning(SVM), Multinomial Naïve Bayes and Bernoulli Naïve Bayes algorithms to develop a sentiment analysis classification model to adeptly classifier opinions as positive or negative if subjective or neutral otherwise.
IV. Create a prototype that presents the outputs of the analysis and summarizes the opinions both through the computer and on handheld devices.

1.4 Research Questions

Based on the above research objectives, the research questions include:-

I. What is needed in mining data from social media?

II. How can we effectively perform feature extraction and categorization for sentiment analysis to identify the marketing requirements of a company?
III. What are the various techniques at hand used in creating a sentiment classifier and what are their constraints?
IV. How can the classifier accuracy be increased?
V. For output presentation, using which graphical representation formats strongly relays the analysis outcome? How can summarization of the sentiments be included in the output?

1.5 Rationale of the study

Marketing to a big extent dictates how well a business will perform. Social media has popped in the business scene to become the fastest growing marketing tool (Goutam Chakraborty and Krishna Pagolu, 2014). To make sense of the continuously growing data in social media for business purposes, we need to device approaches more effectively and accurately analyze the data for insights. If we cannot scrutinize to find what is in this large proportion of opinions voiced, we fail to obtain revolutionary value concealed therein.

1.6 Scope

The scope of this research is limited to activities that lead to collection of business orienteddata, preprocessing,creating a machine learning classifier,model building, evaluation and assessment of the model for a local Kenyan company.

Employing machine learning approach in sentiment analysis, a combination of Support Vector Machine and Naïve Bayes which are both supervised learning algorithms areselected. However, the performance of these supervised learning methods relies on manually labeled training data making this approach domain dependent. The flexibility of the proposed classifier and model might not perform with relative high degree of success unless maybe in an identical domain and setting. The developed model will strongly consider direct opinions.

The study will undertake and recommend a model deployment method for any typical productive environment within the same domain.

1.7 Assumptions and limitation of the research

The assumptions and limitations in this study are outlined here-under:

I. With reference to sentiment classification dimension, assumption is made that the positive-negative and extreme-average imbalances do not exceptionally displace the harmony of opinions.
II. The study assumes that even though geographical elements are not included in this research, there is a degree of uniformity.
III. The study assumes a given set of feature vectors used in support vector machine classification algorithm. The list of vectors is by no means authoritative and complete but sound enough for analysis to meet the marketing requirements.
IV. Due to time constraints, a small corpus is built for this study.
V. Kenyans use several languages in social media to express their emotions. The study is limited to expressions made in English, Swahili and semi-standardized Shengô languages.

CHAPTER TWO

LITERATURE REVIEW

In this section, a review of existing literature on sentiment analysis is done. A description is made employment of machine learning and statistical techniques to determine sentiments in social media opinions. A review of the application of various techniques for sentiment analysis is also done.

2.1 Sentiment Analysis

Opinions are central to almost every aspect related to humans and remain one of the key influences of our behavior. Sentiment analysis is the computational study of opinions, sentiments and emotions expressed in text. Origin of sentiment analysis is rooted in the disciplines of psychology, sociology and anthropology and flow from the theory of affective stance and appraisal theory which focus on emotions in shaping cognitions (Rambocas et al, 2013). It has many names, opinion mining, appraisal extraction and even subjectivity analysis (Kadam et al, 2013). Often itôs not important to know what users are saying, but how they are saying it. Sentiment analysis seeks to automatically associate a piece of text with a sentiment score (Kumar et al, 2013).

Pang et al (2002) classified documents by positive or negative polarities. They conducted their study on movie reviews. They compared three different machine learning techniques: Naïve Bayes, Support Vector Machine and Maximum Entropy. Later they added the dimensionality of analysis by rating review scores. Discovering that the best classifier among the three was the support vector machine, they also realized that machine learning approach is greatly domain dependent. This is because words can depict different emotions in different domains.

Hu and Liu (2004)studied sentiment analysis of customer reviews on products. They proposed a framework to analyze and compare consumersô opinions of competing products; Opinion

Observer prototype used supervised rule discovery to extract features and their corresponding pros and cons. They perform the study in three stages. They identified the feature first, then for each feature, they computed the polarities before finally summarizing those reviews.

Go et al. (2009) used distant learning to acquire sentiment data. Considering tweets ending with positive emoticons like ñ☺ò ñ:-*ò and negative emoticons like ñ☹òò:-/ò as positive and negative respectively. Comparing SVM, NB and MaxEnt models, SVM outperformed the rest. They also considered that feature space and unigram performed best. Pak and Paroubek (2010) used emoticons to identify tweets and train a classifier. They collected data using a similar distant learning paradigm but classified in terms of subjective or objective. They reported that POS and bigrams both help. Both the two studies are based on n-gram models.

Qi, Guo and Hinrich (2013) used support vector machine (SVM) with a large range of features, POS features, stylistic features, emoticons, domain name, readability scores and other statistics to classify tweets. They did feature selection of all features described using mutual information and 10-fold cross validation. They discovered that simple statistics of tweets such as word count or readability scores can assist in twitter sentiment analysis.

Turney et al (2002) while again studying customer reviews introduced an unsupervised algorithm as recommended or not recommended. The classified items were based on fixed, syntactic phrases used for expressing opinions. They utilized phrases rather than words and implement the algorithm based mainly on Pointwise Mutual Information and Information Retrieval to compute semantic orientation of the given review. The study was carried out the in three steps. 1) Began by first extracting phrasal lexicon from reviews. 2) Then each phraseôs polarity was identified. 3) With regard to the average polarity of a phrase, a review was polarized.

Mullen, Tony and Nigel Collier (2004) introduced the concept analysis using support vector machines (SVM) to bring together diverse sources of information. They introduced models using features and combine them unigram models. They used the method of semantic orientation with PMI. After experimenting with the movie review domain, they discovered that hybrid SVMs

which combine unigram-style SVM feature SVMs with the ones based on real-valued favorability measures obtained superior performance.

Mohammed et al (2013) studied on how to build-state-of-the-art in sentiment analysis of tweets. They incorporate support vector machine which had an F-score of 69.02 in the message level and 88.93 in the term level. They implemented a variety of features based on surface form and lexical categories. Most gain in performance was led by sentiment lexicon features along with n-gram features.

Gitau and miriti (2011) researched on sentiment analysis using unigrams, emoticons and bigrams mined from twitter. Using Naïve Bayes classification they explored Kenyan issues.

2.1.1 Subjectivity and objectivity

Not all text pieces usually contain useful opinions. Subjectivity/Objectivity classification comes in to separate subjective sentences from objective sentences. Pang et al (2002) studied sentiment analysis by determining whether a sentence is subjective or objective. Subjective sentences usually express opinions whereas objective sentences just states facts.

2.1.2 Levels of sentiment analysis

When performing sentiment analysis, the overall sentiment analysis score can be computed from different levels:

a. Document level sentiment analysis.

It essentially operates on opinionated text in the unit of a document. In this document level classification, a single review about a single topic is considered (Varghese et al, 2013). This becomes a challenge when dealing with comparative statements. All sentences in the document

may not be relevant or express any opinion hence objective/subjective classification is quite imperative.

b. Sentence level sentiment analysis.

Sentence level sentiment analysis evaluates polarity of each sentence. Sentence-level subjectivity classiýcation is useful because most documents contain a mix of subjective and objective sentences (Wiebe et al, 2005). In case of simple sentences, a single sentence bears a single opinion about an entity. Complex sentences are not desirable to sentence level sentiment analysis (Varghese et al, 2013).

c. Phrase level sentiment analysis.

Phrase level sentiment classification is a more meticulous approach to opinion mining. It is sometimes referred to as aspect level sentiment analysis. The context of opinion depends on the wording. The words that appear very near to each other are considered to be in a phrase (Varghese et al, 2013). Turney et al (2004) used phrases rather than just words and applied piontwise mutual information (PMI) between words to label the polarities.

2.1.3 Review of Sentiment Analysis Techniques

There basically exist four main categories to perform sentiment analysis:

I. Keyword spotting.
II. Lexical affinity.
III. Statistical methods.
IV. Concept-level techniques.

I. Keyword spotting

It is the most naïve approach. It categorizes text based on the presence of affect words. This words are usually given some sentimental values in a given linguistic annotation schemes. Hence, the presence of the keywords most probably indicated the orientation of sentiment polarity. It is not robust to negation and relies on surface features (Cambria, Erik et al, 2013).

II. Lexical affinity

This approach is much more powerful than keyword spotting. In addition to detecting affect words it assigns arbitrary words a probable ñaffinityò to particular emotions (Cambria, Erik et al, 2013). This approach is not robust to negation and sentences with other meaning.

III. Statistical methods

This method utilizes machine learning approach. Pang et al (2002) observed machine learning techniques (Naïve Bayes, Support Vector Machines and Maximum Entropy) outperformed human-produced baselines. Some earlier study with machine learning algorithms on movie reviews. Basically, feeding a machine learning algorithm a large training corpus of affectively annotated texts, the system might not only learn the affective valence of affect keywords but also other arbitrary keywords (Cambria, Erik et al, 2013).

IV. Concept-based techniques

These methods employ web ontologies or semantic networks as it heavily relies on knowledge bases. Superior to purely syntactical techniques, concept-based technique can detect multi-word expressions (Cambria, Erik et al, 2013). They have the ability to analyze multi-word expressions related to concepts that explicitly convey emotion.

2.1.4 Machine Learning in Sentiment Analysis

Machine learning (ML) approach applicable to sentiment analysis mostly belongs to supervised classification. It basically includes two sets of data: a training and a test set. The training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to validate the performance of the automatic classifier. Machine learning techniques like Naïve Bayes (NB), maximum entropy (MaxEnt), and support vector machines (SVM) have achieved great success in text categorization. The other most well- known machine learning methods in the natural language processing area are K-Nearest neighborhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model (G.Vinodhini and Chandrasekaran, 2012).

The support vector machine (SVM) is a state-of-the-art learning algorithm proved to be effective on text categorization tasks and robust on large feature space (Saif Mohammad, 2013). This approach was proposed by Vapnik (1995). Pang et al (2004) applied SVM, MaxEnt and NB to classify movie reviews as positive or negative. They found out that SVM performed better than Naïve Bayes and Maximum Entropy classifiers. Ye et al (2009) experiment of applying support vector machine (SVM), Naïve Bayes and N-gram model to the destination reviews depicted that the SVM outperforms the others. Yun-qing xia et al (2007) when proposing a unified collocation framework (UCF) realized that SVM classifier assigns correct labels to most of the true opinions. Situation for accuracy of sentiment analysis is similar to precision in opinion extraction. It dropped by 0.172 when SVM classifier was not used. This justified the importance of SVM classifier in their opinion mining method. Songho Tan (2008) presents an empirical study of sentiment categorization on Chinese documents. Besides investigating the various feature selection methods, he also compared several learning methods including centroid classifier, KNN, NB, winnow classifier and SVM on Chinese sentiment corpus. He found out that SVM performed better sentiment classification than the rest. Riu Xia et al (2011) made a comparative study of effectiveness of ensemble technique for sentiment classification. They employed three well-known text classification algorithms, namely naïve Bayes, MaxEnt and SVM as base classifiers for each of feature sets and too concluded that SVM performed even better. Multiple variants of SVM have been developed in which multi-class SVM is used for Sentiment classification (Kaiquan Xu, 2011). In most of the comparative studies it is found that SVM outperforms other machine learning methods in sentiment classification (Vinodhini and Chandrasekaran, 2012).

Naïve Bayes is a simple but effective classification algorithm (Vinodhini and Chandrasekaran, 2012). It is a family of probabilistic classifiers based on applying Bayes theorem with substantial assumptions between features. Melville et al (2009) used NB to present a unified framework in which they used background lexical information in terms of word-class associations. Rui Xia (2011) applied NB algorithm on feature sets when making a comparative study of the effectiveness of ensemble technique for sentiment classification. This algorithm is widely used for document classification (Songbo Tan, 2008). Its assumption of word independence makes it efficient.

Maximum entropy (MaxEnt) classification is a machine learning classification method that generalizes logistic regression to multiclass problems. MaxEnt models are feature-based models (A Go et al, 2009). The idea behind MaxEnt models is that one should prefer the most uniform models that satisfy a given constraint (Nigam et al, 2009). Given an independent variable, the model predicts possible outcomes for a categorically distributed dependent variable. Pang et al (2002) reports that from previous studies, MaxEnt outperformed NB sometimes, but not always. This is because unlike NB, MaxEnt makes no assumptions about the relationships between features therefore it might potentially perform better when conditional independence assumptions are not met. McDonald, Ryan (2009) experiment and analysis gave significant support for the mixture weight method for training very large-scale conditional maximum entropy models with L2 regularization.

The idea behind centroid classification algorithm is simple and straightforward (Sangbo Tan, 2008). Rocchio classifier is the centroid classifier approach that is applied to text classification using term frequency-inverse document frequency (tf-idf). Erik et al (2013) applied centroid algorithm to perform open domain sentiment analysis. They blended the largest existing taxonomy of common knowledge with natural-language-based semantic network of common-sense knowledge then applied multi-dimensional scaling on the resulting knowledgebase.

Other machine learning algorithms have also been used but not widely. K-Nearest Neighbor (KNN) is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document (Vinodhini and Chandrasekaran, 2012). Winnow is a mistaken-driven method using multiplicative scheme. Glance, Natalie et al (2005) while deriving

marketing intelligence from online discussion empirically found winnow to be a very effective document classification algorithm. Rudy Prabowo (2009) combines rule-based classification (RBC), supervised learning and machine learning into a new combined method and test the method on movie reviews. The results portrayed improved classification effectiveness in terms of micro- and macro- averaged F1.

2.2 Social Media

Social media is the combination of three elements: content, user communities and Web 2.0 technologies (Ahlqvist et al, 2008). It employs mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user generated content (Kietzmann, Jan H., et al, 2011). Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics, it can be construed as a form of collective wisdom(Asur et al, 2010). In business, the impact of consumer-to-consumer communications has been greatly magnified in the market place (mangol et al, 2009). It presents an interesting opportunity for harnessing the data into a form that allows for specific predictions about particular outcomes, without having to institute market mechanism. This can provide great insights and principles organizations can follow to interact with the market environment through integrated marketing communications (IMC). Jansen et al (2009) investigated twitter microblogging as a form of electronic word of mouth (w-o-m) for sharing consumer opinions concerning brands. W-o-m is considered a powerful marketing medium for companies to influence customers.

2.2.1 Examples of social media

There are several social media sites available today:

i. Twitter

It is an online social networking and microblogging service. It was created in 2006 and has gradually gained popularity with 1 million users (socialbaker 2014). It allows users to send and read short messages (140 characters) called ñtweetsò. These tweets are opinions that

contain diverse sentiments towards various topics. Sentiment analysis can be performed on the tweets and the character length constraint and the word level granularity aptly suits its setting (Kumar et al, 2012). Jansen et al (2009) study investigated the effects of services in commercial sector, namely, the impact on the relationship between company and consumer. Contrary to techniques used for harvesting data from online resources, twitter provides application programming interface (API). Despite its unprecedented availability of data it usually comes at a premium price and only a fraction is provided free. Off the shelve approaches are usually unable to operate due to real-time computational demands and therefore some computational costs. Valkanas et al (2013) evaluated the extent to which analysis processes are affected by the limitations. After using a regular API and Gardenhouse API, they observed that samples differ by an order of magnitude but with identical temporal patterns and periodicity.

ii. Facebook

It is a social networking site founded in 2004. Users create personal profiles, add friends exchange messages and status updates. Unlike Twitter, posts can be much lengthier hence quite laborious when doing any form of text mining and analytics. It is the most widely used social networking site with about 1.28 billion registered users (socialbakers 2014). Its robust nature is quite evident. It was blocked in China in July 2009 after riots in the western province of Xinjiang. This was because it could intensify the political unrest as the communication links would reach even greater audience in real time.

iii. Google+

It is a social networking site that Google describe as a social layer enhancing many of its properties. It was referred to as Google Circles before. In addition to having Gmail email services, it also has +1 button to signal appreciation and the ability to comment on YouTube comments. It is the second-largest social networking site in the world with over 600 million active users as of October 2013 (socialbakers 2013).

iv. LinkedIn

LinkedIn is a business-oriented social networking service founded in 2002. It allows users to create profiles and connections to each other mostly for professional networking. It is the largest professional networking site with more than 300 million users as of June 2014 (socialbakers 2014).

v. YouTube

YouTube is special social media platform based on video-sharing. It was created in 2005. According to grits report (2014) YouTube uploads 24hrs video every minute.

2.3 Ensemble Learning

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of the predictions (Dietterich,2000). It is quite obvious that groups of people can often make better decisions than individuals. The same is believed in machine learning. The idea behind ensemble mechanisms is to exploit the characteristics of several independent learners by combining them in order to achieve better performance than the best baseline classifier (Fersini et al, 2014).The main objective of ensamble is to maximize individual accuracy and diversity. In essence, it strives to achieve performance by combining the opinions of multiple learners.

Two necessary conditions should be satisfied to achieve a good ensemble: accuracy and prediction diversity (Fersini et al, 2014).

Below is a figure depicting ensemble method architecture.

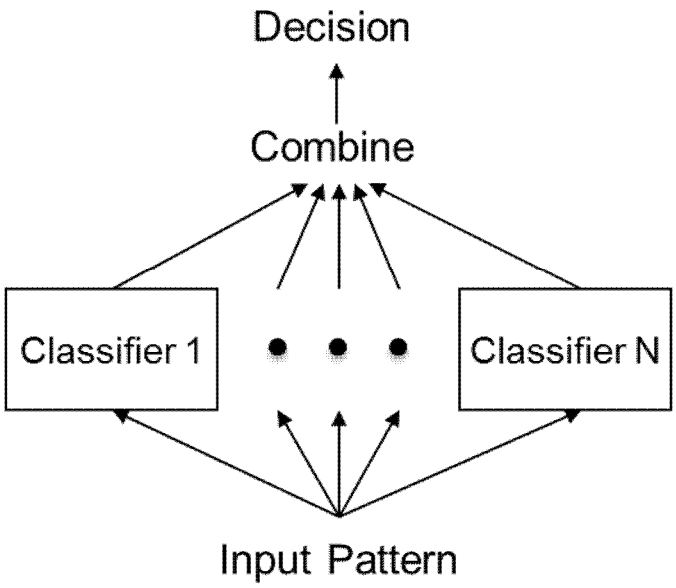

Fig 2.1 Ensemble Learning overview

2.3.1 Reasons for ensemble method.

The three fundamental reasons to consider an ensemble method approach according to Dietterich (2000) are:

1. Statistical - By constructing an ensemble the algorithm can ñaverage" their votes and reduce the risk of choosing the wrong classier.
2. Computational - An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual classifiers.
3. Representational ï When the true function f cannot be represented by any of the hypothesis in H (weighted sums of the hypotheses drawn from H might expand the space).

2.3.2 Popular ensemble methods.

There are several techniques used to construct ensembles but the most popular ones are:

1. Bagging.

It comes from the phrase Bootstrap AGGregation. It is the way decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multi-sets the same cardinality/size as your original data.

In bagging, generating complementary base-learners is left to chance and to the instability of the learning method. An unstable algorithm is one that when there is a small change in the training set, it causes a large difference in the base learners.

2. Boosting

It is an approach to calculate the output using several different models and then average the result using a weighted average approach. Here, we actively try to generate complementary base-learners by training the next learner on the mistakes of the previous learner.

The disadvantage of the original boosting method is that it requires a very large training sample.

3. Stacking

It is a similar to boosting: you also apply several models to you original data. The difference here is, however, that you don't have just an empirical formula for your weight function, rather you introduce a meta-level and use another model/approach to estimate the input together with outputs of every model to estimate the weights or, in other words, to determine what models perform well and what badly given these input data.

The combiner function f() is another learner and not restricted to being a linearcombination as in voting.

2.4 Self-report imbalances

There have been social psychological studies done on social media usersô behavior. This is to try to better grasp the framework around their emotions. Lin, et al. (2013) studied on how to control influence of individual existing biases. Communicative nature of social media causes a reactive characteristic of social data do problematically model or mirror the world events (Rost, Mattias, et al. 2013). Many different aspects of emotion can be evaluated, including: individualsô self-reports of feelings, neurological changes, autonomic system reactions, and bodily actions (Mauss & Robinson, 2009). Self-report is any method which involves asking a participant about their feelings, attitudes, beliefs and so on (wikipedia). Guerra et al (2014) raise awareness over the fact that opinions voiced over social media platforms are impacted by many social and psychological factors which need to be considered so as to build an effective and reliable sentiment analysis system. He mentions two kinds of imbalances:

i) Positive-negative sentiment report imbalance.
ii) Extreme-average sentiment report imbalance.

2.4.1 Positive-negative sentiment report imbalance:

Based on psychological research, people tend to express positive feelings more than negative feelings. This could be rather complicated to classify as one event may be positive to one group but negative to another. A game for example, the supporters of winning team will be happy as oppose to the loosing team. Also the credibility of the expressions comes to question as some will voice positive emotions for public display.

2.4.2 Extreme-average sentiment report imbalance:

This one explores the human tendency to report extreme experiences more than average experiences. This can be an inhibitor of getting the overall outlook of the situation in the marketplace.

2.5 Sentiment Lexicon

Sentiment lexicons are lists of words with associations to positive and negative sentiments (Saif Mohammad, 2013). The list contains words that are used in training the sentiment classifiers. Languages that have been studied mostly are English and in Chinese (Vinodhini and Chandrasekaran, 2012). Examples of sentiment lexicons include:

1. Sentiment140 lexicon a Stanford University project (Go et al, 2009) is a set of about 1.6 million tweets with positive and negative emoticons. By applying machine learning approach, tweets are classified positive or negative based on emoticons.
2. Multi-Perspective Question Answering (MPQA) Wiebe et al (2005) is a lexicon with subjective expressions.
3. SentiWordNet Lexicon Esuli et al (2006). It provides an extension for WordNet associating synsets with an emotional value.
4. Bing et al (2004) while studying mining and summarizing customer reviews constructed Bing Liu lexicon. It is useful as it includes misspellings, morphological variants, slang, and social-media mark up.
5. LWIC (Linguistic Inquiry and Word Count) evaluates emotions in a text by employing pre-classified words in a dictionary.

Kouloumpis et al (2011) found out that while analyzing the microblogging domain, though part-of-speech (POS) features may not be particularly useful for sentiment analysis, sentiment lexicon was to some extend useful.

2.6 Feature Extraction

Feature extraction involves identifying the aspects being commented on e.g. Flight delay (Henrique and Flavia, 2010). For sentiment analysis, feature extraction is one of the most complex tasks as it requires use of Natural Language Processing to automatically identify the features in the opinions under analysis. It also largely reflects on the efficiency of the overall sentiment analysis task.

Wu et al (2009) use phrase dependency parsing. In dependencygrammar, structure is directly determined by the relation between a head and its dependents. The dependent is a modifier or complement and the head plays a more important role in determining the behaviors of the pair. The authors want to compromise between the information loss of the word level dependency in dependency parsing as it does not explicitly provide local structures and syntactic categories of phrases and the information gain in extracting long distance relations. Hence they extend the dependency tree node with phrases.ò

Hu et. al (2009) used frequency principle. The frequent item set was used to extract the most relevant features from a domain and pruned it to obtain a subset of all the relevant features. They extract the nearby adjectives to a feature as an opinion word regarding that feature. Using a seed set of labeled adjectives, which they manually develop for each domain; they further expand it using WordNet and use them to classify the extracted opinion words as positive or negative.

Mukherjee et al (2012) presented a novel approach to identify feature specific expressions of product reviews with different features and mixed emotions. They did this by exploiting the opinion expressions association to those features.

Most of the work has been done for product review. Siqueira et al, (2010) studied feature extraction for sentiment analysis of opinions on services. It is particularly difficult to identify the features being commented on in this case. They present a domain-free process.

2.7 Challenges in sentiment analysis

Sentiment analysis is a pretty new research field still gaining ground. Some challenges observed are:

2.7.1 Named Entity Extraction

Named entities are definite noun phrases that refer to specific types of individuals, such as brands, products, organizations and so on. The goal of named entity extraction is to identify all textual mentions of the named entities in a text piece (Pak et al, 2010).

2.7.2 Misspelling and creative spelling

Social media users usually have no standardized format of expressing their opinions. Users often wrongly spelling words or use some artistic dimensions to write their words. These words usually create trouble when trying to analyze them.

2.7.3 Slang and informal language

Many generation and group of people use different informal languages with regard to various contexts. In Kenya, the difference in diversity and for a concept of recognition, so many people especially the youth use informal language. These languages differ from region to region.

2.7.4 Information Extraction

As information from the data sources come in many shapes and sizes, complexity of natural language can make it very difficult to access the information in the opinion text (Pak et al., 2010). Unlike humans, machines cannot identify the relevant information as easily as humans do. NLP tools are still trying to build a general purpose representation of meaning from unrestricted text.

2.7.5 Co-reference Resolution

It usually occurs at the aspect of sentiment analysis and entity level. When co-referring words are not found out, effective sentiment analysis cannot be carried out (Pak et al, 2010). The two different types of opinions are direct opinion where emotions are explicitly expressed on a target while comparative opinion compares several targets. The comparative texts may contain co-references which must be effectively resolved to correct results.

2.7.6 Relation Extraction

This is a major research area in NLP. Relation extraction is the task of finding the syntactic relation between words in a sentence (Pak et al, 2010). Semantic of a sentence is realized by knowing the word dependencies.

2.7.7 Domain Dependency

Many classifiers trained to classify opinion polarities in a specific domain and perform relatively well may produce miserable results if applied in a different domain. Woller, Martin et al (2013) studied how to perform domain-independence sentiment analysis for movie reviews. Essentially this is still an unresolved challenge in sentiment analysis.

2.8 Marketing Intelligence

At the beginning in the 1970ôs, together with administration, production, finance and personnel, sales unit was one of the important areas in business. Progressively this area was enhanced by market studies and communication. It grew with the realization of the crucial aspect customers played. It has so far bloomed though not basically more important than others, it has evolved to become an essential part of any organization. Experts believe marketing agenda is to get into the customerôs head so as to realize what he wants and then how best the organization can approach it. Essentially, it is an analytical tool.

Fig 2.2 marketing evolution

2.8.1 Economic environments

Marketing is used to measure satisfaction, evaluate potential, discover new openings, verify the segmentation method adopted, reinforce sales arguments etc. (Philippe M and Christophe B, 2002). As a general rule, marketing develops when offer outstrips demand, in other words, when the products or services that companies produce exceed what customers can buy (Malaval et al, 2013). Economists refer to four types of economic environments:

a) Production economy ï It is characterized by a situation where customersô needs are not satisfied by the companies. In this phase, there is no real need for marketing. Companies need to keep producing while still maintaining economic efficiency. This economy is greatly characterized by very low competitive pressure this is because companies are certain to place what they produce. In our current generation, finding this type of environment is an herculean task as most enterprises have ventured pretty much in all industries. However, in situations where there are conflicts or natural disasters, this environment can be observed temporarily.

b) Distribution economy ï Here, the supply and demand forces are at equilibrium. This is a very challenging task due to the dynamic state of the marketplace. It is next to impossible to supply to customersôdemands with meticulous accuracy.

c) Market economy ï This corresponds to an economy where the supply every so often greatly surpasses the demand. In this scenario, marketing is very imperative as the customers cannot purchase all the goods or services offered.

d) Environment economy ï It closely relates to the market economy and follows on form it by taking into account those non-economic organizations.

2.8.2 Growth of marketing

Marketing is a constantly changing discipline. Erragcha et al (2014) study highlights mutations of marketing techniques based on social and technological developments. Finding earlier approaches involved passive users engaging in unidirectional communication, in the latest marketing paradigm, consumers have changed and become more sensitive to the society.Since the advent of the Internet, social nature of humanity has been enhanced. Marketing has also experienced very rapid development. Combining web 2.0, web-marketing and social media considers consumers as active players. Making companies listen to voice of customer (VOC). Indeed,we can identify consumersô attitudes towards brandsthrough their interaction the Web: blogs, forums, social networks (Facebook, Twitter), online citizen media (Erragcha et al, 2014).

2.8.3 Marketing problem solving modes

There have been observed approaches to making marketing decisions. Basically, decision making is dependent on three factors:

- Marketing problem.
- Decision maker.
- Decision environment.

The result is four different marketing problem-solving modes: Optimizing, Reasoning, Analogizing, and Creating (ORAC) (Wierenga and van Bruggen 1997;2000). Clocks of mind depicts hard calculations while clouds of mind a situation having free flow of thought without a clear goal.

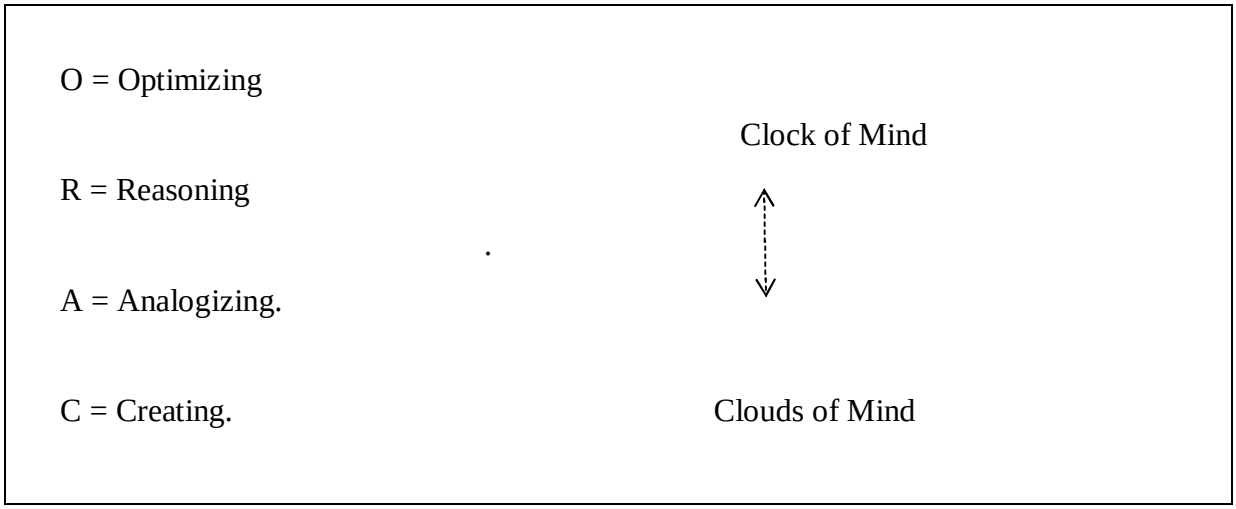

Fig. 2.3 The ORAC model of marketing problem-solving modes.

1. Optimizing: Usually possible with clear and precise insights in the mechanism behind thevariable that we want to optimize. This mode actually implies the presence of an objectively best solution achievable by proper use of the marketing instrument. An example can be choosing the most desirable advertising campaign once the budget and relevant reach and costs are known.

2. Reasoning: Here, a marketer has a mental model of certain marketing phenomena and uses this as the foundation for drawing conclusions. An example is a situation where the

marketing managers have a representation of the factors the companyôs brand market share. This could be observed to vary with respect to geographical location. The managers could then reason giving possible causes, (i) customers deviant preferences; (ii) ineffective sales taskforce; (iii) strong competition (Goldstein 2001). Marketing research comes into play here to assist in scrutinizing these causes more explicitly.

3. Analogizing: A simple illustration is the use of solar system to study an atom. In essence, it utilizes the idea of relating relatively similar problems in making marketing decisions. Experiences with earlier product introductions in the market act as points of reference

4. Creating: The building blocks are novel and effective ideas and solutions. This is a very important mode that automatically employs divergent thinking. Marketing managers are always probing better and more effect approaches to marketing.

2.8.4 Marketing Intelligence

Marketing intelligence involves primary data collected and analyzed by a business about markets that it anticipates participating in with the intention of using it in making decisions (marketing intelligence, 2013). There is no set plan for how companies gather marketing intelligence but by collecting data about the markets, they can get valuable insights into how to grow their business. Glance, Natalie et al (2005) studied on approaches of deriving marketing intelligence from online discussions. They presented a system that gathers and annotates online discussion relating to consumer products using variety of state-of-the-art techniques.

Marketing intelligence can be defined as an ongoing, holistic knowledge of all aspects of the aspect place (Quirkôs marketing research, 2004).

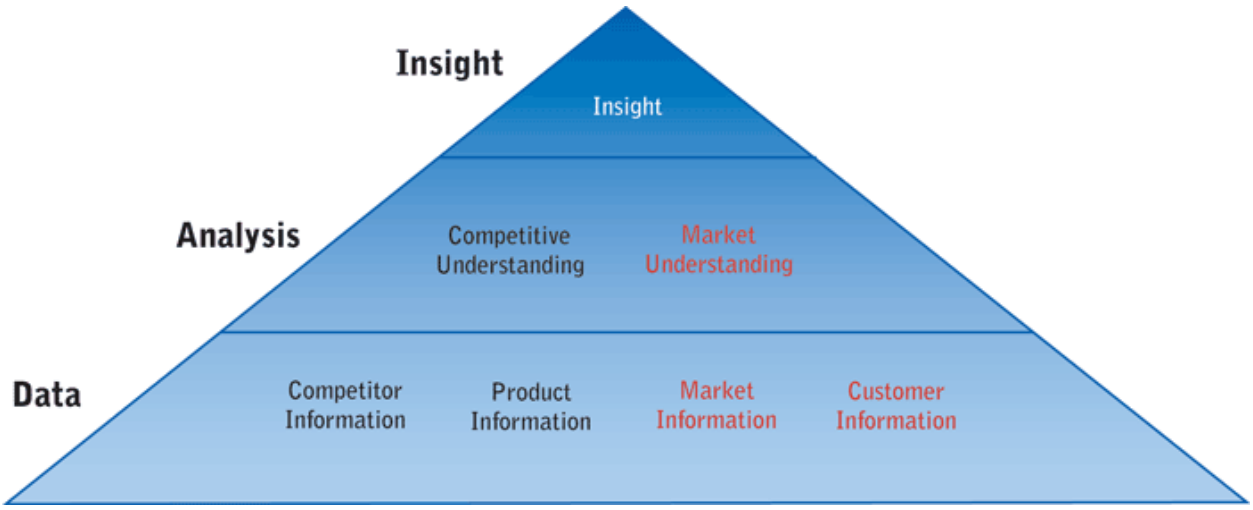

Fig. 2.4 marketing intelligence pyramid

The information face of the market intelligence pyramid is built upon a foundation consisting of four basic areas: competitor information, product information, market information and consumer information. Each of these areas of knowledge can be a unique discipline in and on itself.

2.9 Ethical issues

Social media created revolutionary opportunities to connect people. Humphrey et al (2010) studied content analysis of twitter in which the amount of personally identifiable information in twitter messages was coded. There may be risks associated with information sharing on social media. Considering safety implication is very important due to social media broadcastability.

Sentiment analysis involves accessing personal data of some kind which can lead to the disruption of some important normative values. Violating peoplesô personal information is one of the most obvious ethical objections. Information obtained from anonymous data does not link individuals to the messages hence no direct sense of privacy violation.

This study intends to utilize public data from twitter. The data is made anonymous to shield individualôs privacy. Authorôs identity is excluded during harvesting of the data.

2.10 Conceptual model

From the previous work done in this field and literature studied and cited above, this study identifies the most appropriate approach. These approaches form the building blocks of a conceptual model used for this research.

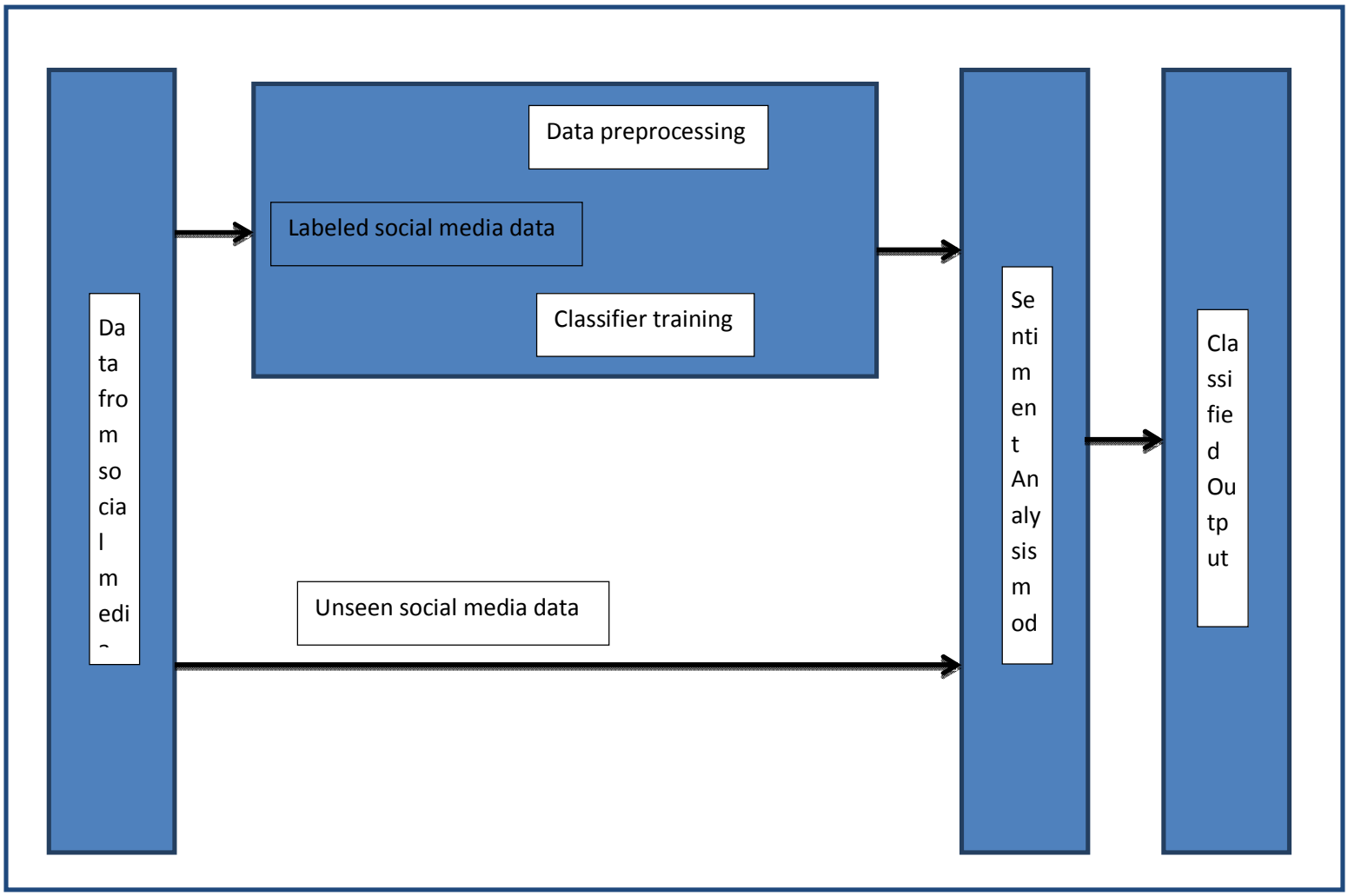

Fig 2.5 Conceptual Framework

CHAPTER THREE

RESEARCH METHODOLOGY

In this chapter, the research methodology is outlined. Data collection, Text preprocessing, sentiment detection, sentiment analysis and model development are done.

3.1 Research Design

There are two fundamental research approaches: qualitative and quantitative approaches. Despite the ongoing debate, recent development in research methodologies suggest that the two approaches should be integrated in comprehensive research designs in order to improve research rigor and address several of the epistemological and methodological criticisms (Kelle, 2006; Olsen, 2004). Sentiment analysis has the potential of employing pluralism. It can be used as a complementary research technique.

This research design outlines: the data collection from source; data preprocessing and cleaning; building classifier using support vector machine; system design and implementation; evaluation. The chosen design is formidable as social media data in Kenya is sparse more so with regard to the ones that are business oriented.

3.2 Sentiment analysis system architecture design

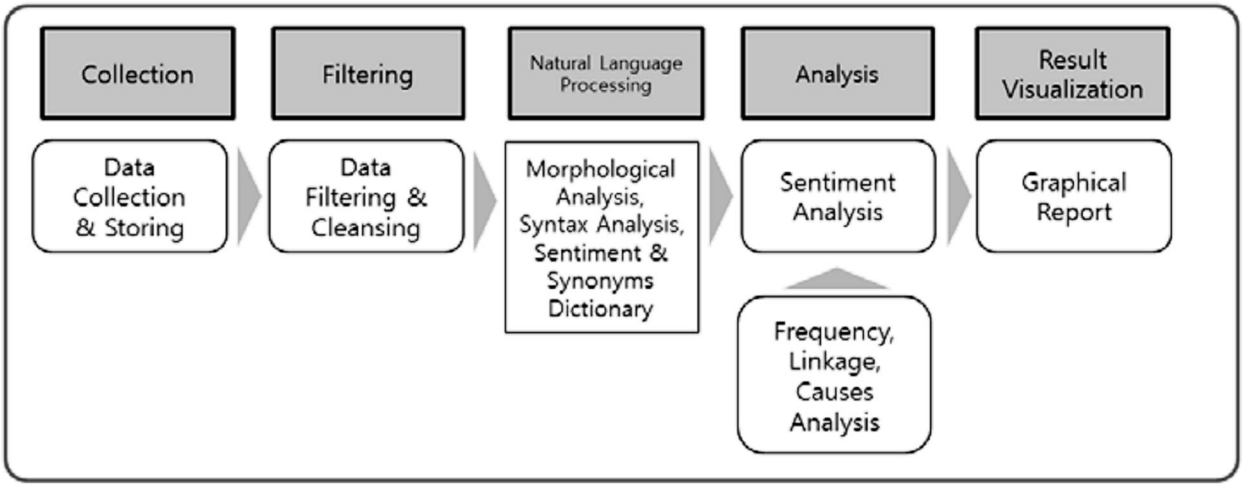

Fig. 3.1 Sentiment analysis system architecture

Choi, Chi-Hwan, et al (2013) proposes a sentiment analysis architecture design which is adopted for this study. This design is used in coming up with a prototype. Prototyping involves building a working model of the candidate system for evaluation. The two main approaches are: Throwaway and evolutionary. Evolutionary approach is used for this study so as to constantly refine it.

3.3 Data Collection

This study is based on data from social media and particularly twitter and Facebook. Harvesting data from social media is provided for by several APIs. Valkanas et al (2013) demonstrated that though we have the free API which provides for a relatively low tweet count compared to the commercial ones, they main difference is the magnitude, otherwise they both observe identical periodicity and temporal patterns. For this research, the free twitter API and Facebook Graph API are customized to fetch data from the twitter and Facebook sites. Due to the customized web crawler instead of the open source crawler, web crawling speed has been accelerated significantly (Choi et al, 2013). The tweets and status collected are stored before text preprocessing is applied. Labels of positive, negative and neutral are appended to the training set of social media data. The collected data includes text, emoticons and common acronyms.

Twitter API provided by Twitter Inc. is a free and readily available tool for use in accessing the twitter network platform. Though the content is limited for free version, Valkanas et al (2013) report showed the difference is only in magnitude. Otherwise they have relative identical temporal patterns and periodicity. Facebook on the other hand applies a Graph API. It connects the various elements of Facebook components.

For this exercise, keywords are used particularly the ones oriented to the business sphere (product, service, brand and company). JSON (JavaScript Object Notation) is implemented to text and stored in the MongoDB. For each post fetched we stored the following attributes:

i. Post message
ii. Post timestamp
iii. Unique post identifier

iv. Polarity

To check the distribution of word frequency, Zipfôs law is applied. It implements the principle that given a corpus, the frequency of any word will be inversely proportional to its rank in the frequency table.

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}.$$ é . (1)

Sample of the opinion data collected which form the corpus are shown below:

Love these 2 super stars @woowoopops @laurahind. Thanks for the sweet ride in bizclass @KenyaAirways @LandRoverUK

KenyaAirways

RT @alykhansatchu: crude oil bear market represents a strong financial tailwind for airlines @BW @KenyaAirways

Ma3Route @KenyaAirways @chriskirwa more like matatu helicopters..

jaredogeda Shouldn't this be a major concern if JKIA isnt accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways #KOT

RavS82 @Ma3Route @jaredogeda @Kenya_Airports @KenyaAirways #KOT not really a concern its us going backwards in life in past years

RT @RavS82: @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isnt accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways

RT @RavS82: @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways

alykhansatchu @BW @KenyaAirways Watching KQ share

RT @RavS82 @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways

Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways #KOT

The more reason why you should visit Kenya..#WhyILoveKenya @MagicalKenya @KenyaAirways @kwskenya @DBK017 @mcta_kenya http://t.co/RsLTMP2EFg

alykhansatchu @BW @KenyaAirways , except those who trade in futures.

KenyaAirways Luggage tag no. 0706KQ395726 Ticket No. 706-5836648984

crude oil bear market represents a strong financial tailwind for airlines http://t.co/CEM5zjfFzU @BW @KenyaAirways http://t.co/JGJlFIPdA3

KenyaAirways to expand codeshare with @EtihadAirways from #AbuDhabi @AUH to #Jeddah,#Dammam,#Riyadh on 31OCT #avgeek

Love these 2 super stars.@woowoopops @laurahind. Thanks for the sweet ride in bizclass @KenyaAirways @LandRover_UK http://t.co/Hp2SvGkMsX

KenyaAirways 0721598765

RT @alykhansatchu: crude oil bear market represents a strong financial tailwind for airlines http://t.co/CEM5zjfFzU @BW @KenyaAirways

Check this advert out via @SusanLUCKYWong Brilliant! @KenyaAirways over to you Get @Lupita_Nyongo and other Kenyans!

dennisomondi Bring me the @KenyaAirways nuts when you get to Mombasa Sir. I will be waiting. Pole for the delay.

RT @mombasasafariss: The more reason why you should visit Kenya..#WhyILoveKenya @MagicalKenya @KenyaAirways @kwskenya @DBK017 @mcta_kenya

ConradMasheti thank you for the good feedback sir. We are glad you enjoyed your #T1A experience @DrOfweneke @KenyaAirways

Thank you @KenyaAirways for today's forum at The Stanley #Nairobi. Looking forward to many more interactive sessions #Partnerships

RT @DrOfweneke: Loving the new terminals at JKIA looking very classy cc @KenyaAirways

3.4 Text Preprocessing

The mined data is usually not only unstructured, It is also contains irrelevant and non-textual characters. Text preparation involves cleaning before analysis is performed (Rambocas et al, 2013). The preprocessing is broken down into 3 steps:

1. Tokenization.

   Tokenization involves splitting a string into its desired constituents seeking to isolate as much sentiment information as possible. Tokenization helps in keeping the vocabulary as small as possible. Emoticons and abbreviations (e.g., OMG, ICB) are identified as part of the tokenization process and treated as individual tokens.

2. Normalization.

   Abbreviations are noted and replaced by the meaning they represent (e.g., ICB ->I canô t believe). Informal intensifiers are also determined such as character repetition and all-caps. All-caps words (e.g I LOVE Kenya Airways) are made into lower case while character repetitions are reduced to three characters.

3. Part-of-speech (POS) tagging.

   Part-of-speech tagging is done by a POS tagger. This basically identifies the various words in the tweet. Any special twitter tokens are noted (e.g., RT, #hashtag,usertag, URLs) and substituted by relevant token.

3.5 Sentiment Detection

Sentiment detection requires appraising and extracting reviews and opinions from textual dataset through the use of computational tasks (Rambocas et al, 2013). In order to perform machine learning, it is necessary to extract clues from the text that may lead to correct classification (Yessenov and Misaivovic, 2009). Objective sentences are discarded as subjective sentences are

further scrutinized for polarity orientation. The detection can be done with various techniques: Unigrams, N-Grams, Negation and Lemmas.

The study endeavored to select the properties of opinions that are relevant for sentiment analysis and particularly targeting commercial orientation. To select the most rewarding features, domain knowledge and experimentation was considered.

To observe a more robust approach for marketing intelligence, both service and product feature extractions are implemented.

3.6 Sentiment Classification

This project employs the use of supervised machine learning approach and specifically a combination of support vector machine (SVM) and Naïve Bayes. Support vector machine is a kind of large-margin classifier which from previous studies has been reported to perform particularly well for sentiment analysis (pang et al, 2002; Dave et al, 2003; Wu et al, 2009). It aims at finding a decision boundary between classes that is maximally far from any point in the training data.

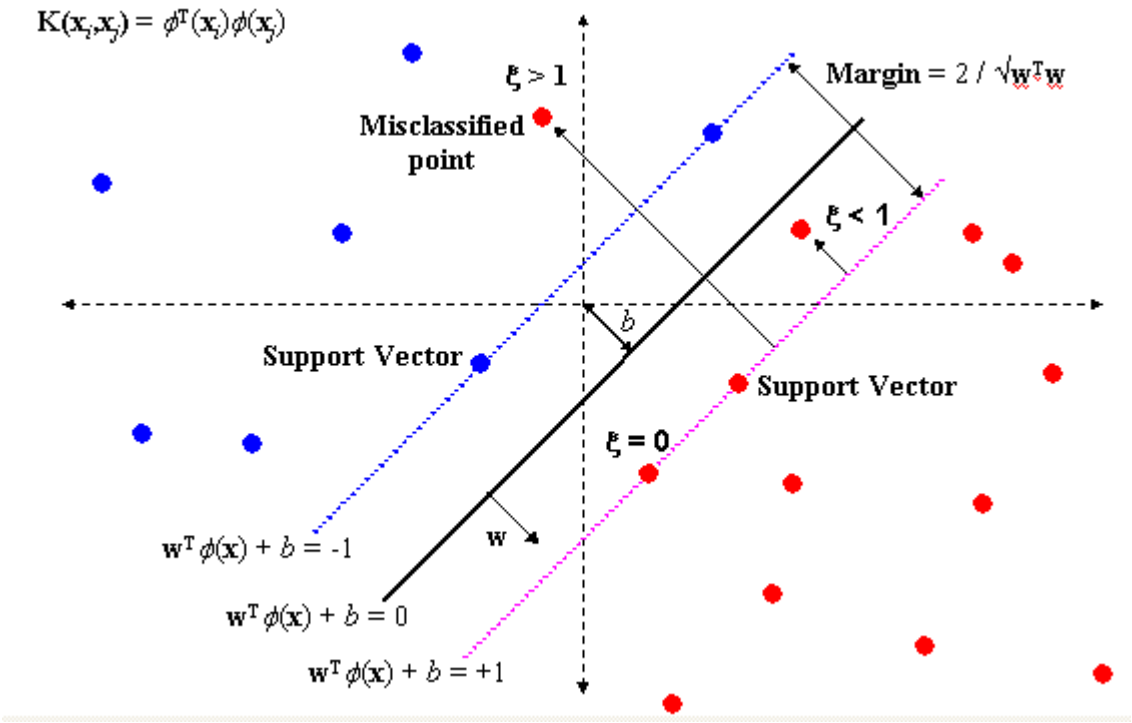

Fig 3.2 Support Vector Machine

Decision hyperplane can be defined by intercept b and normal vector $\vec{w}$ (weight vector).

All points $\vec{x}$ on the hyperplane satisfy:

$\vec{w}^T\vec{x} = -b$ é é é é é é é é é é é é é é é é .é é é é é (2)

Linear classifier

$$f(\vec{x}) = sign\ (\vec{w}\vec{x} + b)$$

é é é é é é é é é é é é é é é é ... (3)

For linearly inseparable data, we utilize the kernel approach. These maps the original feature space to some higher dimensional feature space with separable training set. LibSVM package (Chang and Lin, 2011) and a linear kernel are used in several implementations of similar studies.

$\Phi:\vec{x} \rightarrow \emptyset(\vec{x})$

Common kernel functions in NLP are:

a) String Kernel

   Bag-of-words approach is a string kernel where substrings must be separated by whitespaces.

b) Tree Kernel

   A tree is represented by a vector of integer counts of each sub tree type

To increase the accuracy of the classification, common n-grams are discarded.

Naive Bayes (NB) classifiers belong to the family of probabilistic classifiers based on Bayes theorem. It employs the principle of conditional probability.

$$P(c_j|d_i)=\frac{P(c_j)\prod_{k=1}^{|d_i|}P(w_{d_{ik}}|c_j)}{\sum_{r=1}^{|C|}P(c_r)\prod_{k=1}^{|d_i|}P(w_{d_{ik}}|c_r)}$$ é (4)

NB classifiers have the advantage that it requires only a small amount of training data to estimate the parameters needed for classification. Only the variances of the variables for each class needs to be determined since independent variables are assumed.

To estimate the parameters for a featureôs distribution, one must assume a distribution or generate nonparametric models for the features from the training set (John et al. 1995). These assumptions on distributions of features are called the event model. For this project, we implement Multinomial and Bernoulli distributions as they are considered more popular in document classification.

Multinomial Naïve Bayes is built on the principle of multinomial distribution which is a generalization of a binomial distribution. Applying the model, the likelihood of observing a feature vector F is given by the following:

$$p(F|C)=\frac{(\sum_i F_i)!}{\prod_i F_i!}\prod_i p_i^{F_i}$$

é . (5)

Bernoulli Naïve Bayes on the other hand is built on the foundation of the probability distribution of a random variable. Features are independent Booleans describing inputs. Given F_i as a Boolean expressing the occurrence or absence of the i^{th} term from the vocabulary, then a document given class C has the likelihood computed as follows:

$$p(F_1, \ldots, F_n|C) = \prod_{i=1}^{n} [F_i\, p(w_i|C) + (1 - F_i)(1 - p(w_i|C))]$$ é é é é é é (6)

For this project, the combination of the three classification algorithms was selected so as to try and improve the overall model performance. So the logic follows that, we pick the scores, following the two best out of the three and marge them.

3.7 Classification Assessment

Scikit-Learn metrics module provided functions for calculating accuracy, precision and recall for classifier. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

$$\text{Precision} = \frac{tp}{tp + fp}$$ é é é é é é é é é é é é é é é é é ..é é é é é é é . (7)

$$\text{Recall} = \frac{tp}{tp + fn}$$ é .. é é . (8)

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$ é é é é é é é é é é é é é é é é é é é .. (9)

While building the classifier, training is done so as to predict the label for a given input opinion. This was realized by sklearn (scikit-learn suite for python programming). Python is a very robust

language when it comes to sentiment analysis with libraries used for fitting. SVC and NuSVC implement the ñone-against-oneò approach (Knerr et al., 1990) for multi-class classification.

3.8 Presentation

The general purpose of the analysis is to transfigure the harvested data into meaningful and useful information. Once the analysis is completed, a number of conventional options are used to displaying the result of text analysis (Rambocas et al, 2013). This study utilizes graphical presentation in a web based tool that is mobile device enabled. Marketing intelligence is derived through the interactive analysis of sentiments identified from the input feed.

3.9 Evaluation

To evaluate the overall model, there should be metrics that reflect the success. F1-score is a statistical measure of testôs accuracy combining both precision and recall.The inbuilt NLTK metrics are used to measure accuracy, precision and recall.

The overall model should have the capacity to:

I. Collect opinions from social media sites andpipeline them appropriately into a repository.
II. Identify the relevant features and train the classifier.
III. Accurately test the results especially in terms of quality.
IV. Correctly reflect the analysis in both the web interphase and mobile device interphase.

3.10 Implementation

For this study, the prototype developed is implemented using python for both logic application and interfacing. MongoDB is used as the back end for storing application data and the data fetched. Scikit-Learn, a python module integrating classic machine learning algorithms in the tightly-knit scientific python is used to handle the model development and evaluation.

The Airline industry is selected as the industry of study. This is because while striving to improve the pertinence in Kenyan, it has brought about extensive transpositions to the Kenyan economy. Three airlines are selected for the study: Kenya Airways, Ethiopian Airways and Fly Jambo jet. Kenya Airways is the Kenya national flag carrier, it was founded in 1977 and it is a private-public company. Ethiopian Airways is Ethiopiaôs flag carrier. It was founded in 1945 and has remained the paramount airline in East Africa. Fly Jambo Jet is one of the low cost airlines in Kenya. Though it is owned by Kenya airways, it runs on its solus platform. It majorly focuses on flights destinations within the country.

This project identified the three airlines by literature review for the main reasons of market multidimensional perspective. Ethiopian Airlines brings about the continental and even the global dominance into focus. Fly Jambo jet on the other hand brings the aspect of native market into airlines marketing intelligence perspective.

3.10.1 Prototype BackEnd

It the part of the prototype that handles data scraped from social media platforms. MongoDB is utilized. It is a NoSQL document database. This makes it handle unstructured data from the social platforms more appropriately. It stores JSON-style document which is basically how python handles the data. In case this prototype is to be used in a distributed environment, then MongoDB provides a method for storing data across multiple machines. It is called sharding and it is used to support deployments with very large data sets and high throughput operations. Map-Reduce is also implemented in MongoDB. Essentially, it is a data processing paradigm for condensing large volumes of data into useful aggregated results.

Scripts that scrape the social media are scheduled so that they periodically run and dump the collection in the MongoDB.

3.10.2 Prototype FrontEnd

This is the side that provides for user interface. It is built on django framework. Django is a high-level python web framework that encourages rapid development and clean pragmatic design. It is created inside a python virtual environment which should contain all the required dependency which include NLTK and scikit-learn.

3.10.3 Model

The model is created then passed to a python pickle. It is then loaded from the system disc to the django environment and applied to new sets of data while scraping.

3.11 Algorithm Combination

Advances have been geared towards improving the performance of machine learning classifiers in several dimensions. One of them is combining several machine learning classifiers by primarily combining the set of each individual classifierôs predictions ï typically by voting - and using the result to classify new examples. This is referred to as an ensemble.

Dietterich (2000) defines ensamble methods as learning algorithm that construct a set of classifiers and then classify new data points by taking a (weighted) vote of the predictions.

The most common algorithms include Bayesian averaging, error-correcting output coding, bagging, boosting and decorate. This study utilizes the bagging approach.

BAGGING is basically **B**ootstrap **AGG**regat**ING**. This approach employs the simplest way of combining predictions that belong to the same type. The combination can be realized with voting or averaging where each model receives equal weight. It is a more appropriate approach if the learning scheme is unstable as it improves performances in almost all cases (a learner is unstable if its output classifier undergoes major changes in response to small changes in training data).

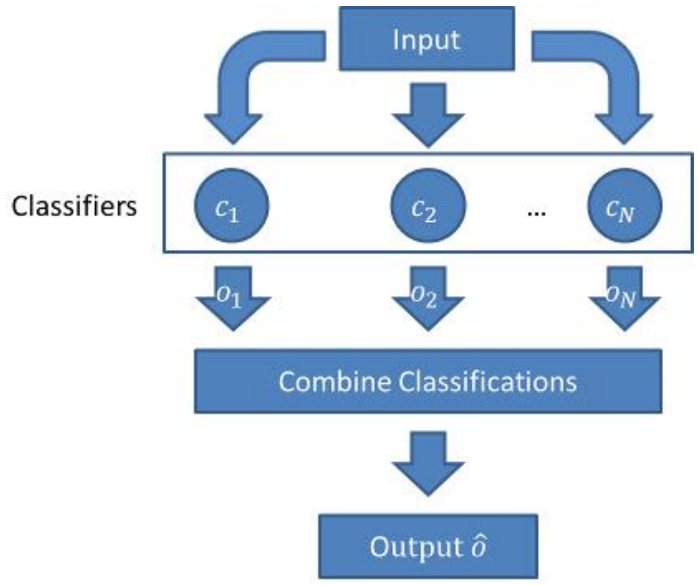

Fig 3.3 Combining Classifiers

The illustration in figure 7 above exhibits a general overview of combining classifiers.

CHAPTER FOUR

ANALYSIS AND DESIGN

4.0 Introduction

The main purpose of collecting data for this research is to build a model. The model is trained and expected to correctly classify unseen data according to the three categories: positive, neutral, negative. All the sentiments are very useful as they give a company the insights from the vocal public.

This chapter outlines the analysis and design as well as the interpretation process for the prototype that is to be built for the purpose of this study.

4.1 Basic Dataset

4.1.1 Prototype Description

Having the model, the project implemented a prototype in the airline industry to instantiate its workability. A use case is employed to describe the prototype. A use case provides a measureable value to an actor or way in which actors interact with the system. Below is a graphical representation of the use case.

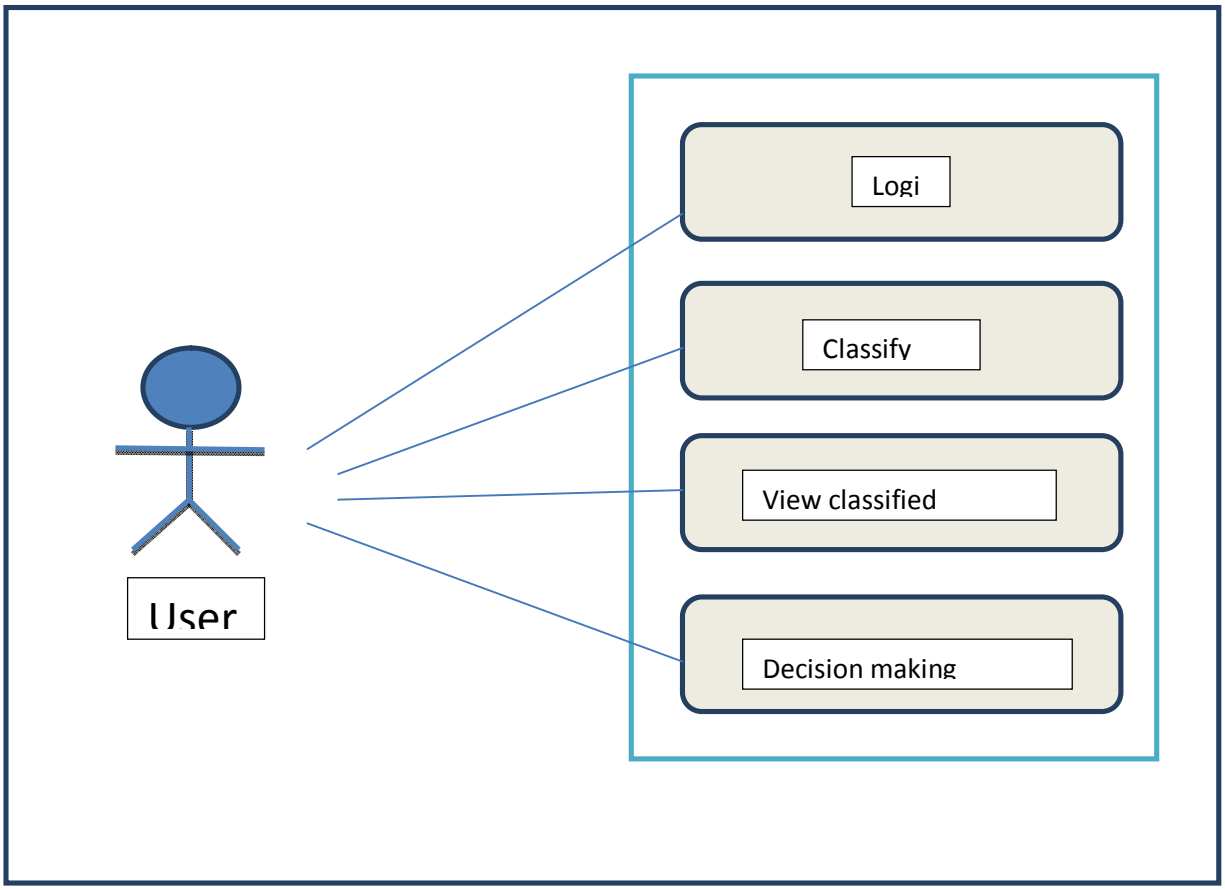

Fig 4.1 Use Case Model

4.1.2 Data Source

Data for this study is scraped from the various social media platforms. The various respective APIs are applied to harvest publicly available opinions. Twitter provides three sources to get their data. Twitter streaming API is the ñdata-pushingò service of the tweets. These tweets are received near real time. The major drawback of the Streaming API is that Twitterôs Steaming API provides only a sample of tweets that are occurring. The actual percentage of total tweets users receive with Twitterôs Streaming API varies heavily based on the criteria users request and the current traffic. Studies have estimated that using Twitterôs Streaming API users can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time.

Twitterôs Search API, involves polling Twitterôs data through a search or username. Through the Search API users request tweets that match some sort of ñsearchò criteria. Twitterôs Search API gives you access to a data set that already exists from tweets that have occurred between 9-11 days earlier. The criteria can be keywords, usernames, locations, named places, etc. Firehose is in fact very similar to the Twitterôs Streaming API as it pushes data to end users in near real-time, but the Twitter Firehose guarantees delivery of 100% of the tweets that match your criteria but at a cost. Firehose is handled by two data providers, GNIP and DataSift.

Facebook is provided for by the graph API. Facebook Graph is a graphic representation of the Facebook community, showing Facebook users and connections. This API presents a simple, consistent approach for developers to interact with the Facebook social graph. For this project, it is used to scrape public Facebook pages belonging to respective companies of interest.

4.1.3 Data preprocessing

Scraped data comes will all sorts of foreign characters such as html tags, elongated words, urls and acronyms. These can be referred to as noise. It is very important to consider them as they hugely reflect on the efficiency of the whole process. Preprocessing consists of three steps:

1. Tokenization.
2. Normalization.
3. Part-of-speech tagging.

Tokenization delimited the data to single sensible units. These units which maybe words, phrases, symbols, or other meaningful elements are called tokens. Text normalization process strives to transform the text to make it consistent. Basically there are two major types of normalization: Stemming and Lemmatization. Both of them are provided for by NLTK. Grammatical tagging is another term given to POS tagging. It is generally the process of appending a tag to the words in a sentence.

4.2 Model Development

4.2.1 Modeling

The model is basically the backbone of the classification. We developed a trained model to observe labeled opinions and classify the unobserved ones. A corpus of 773 opinions is used to train the model with the labels Positive, Negative and Neutral. Both the SVM and Naïve Bayes algorithms are used to complement each other. SVM and NB are often use as baseline for text classification and sentiment analysis. NBSVM is a robust performer that combines both classifiers.

Below is a portion of the manually classified sentiments form the corpus that is used to train the model.

kenya airways	positive	Love these 2 super stars @woowoopops @laurahind. Thanks for the sweet ride in bizclass @KenyaAirways @LandRoverUK
kenya airways	neutral	KenyaAirways
kenya airways	neutral	Ma3Route @KenyaAirways @chriskirwa more like matatu helicopters..
kenya airways	neutral	RavS82 @Ma3Route @jaredogeda @Kenya_Airports @KenyaAirways #KOT not really a concern its us going backwards in life in past years
kenya airways	negative	RT @RavS82: @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways
kenya airways	neutral	alykhansatchu @BW @KenyaAirways Watching KQ share
kenya airways	negative	RT @RavS82 @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways
@FlyJambojet	positive	Thou I missed my flight to Eldoret,the lady at customer service desk at JKIA went her away to explain me as Layman @FlyJambojet thumbs up!
@FlyJambojet	neutral	RT @FlyJambojet: Dont miss your flight. You can check in online from 30 hours to 2 hours prior to departure of your flight on
@FlyJambojet	positive	RT @FlyJambojet: The Early Bird Catches the Lowest Fares http://t.co/l2aLrKUfIy https://t.co/oE4tL8JOSc
@FlyJambojet	negative	@FlyJambojet @qubanltd what flights have the special low rate in the month of November? can't see any via website.
@FlyJambojet	neutral	@EdgarYatich try @FlyJambojet we need you early tomorrow.
@FlyJambojet	positive	RT @FlyJambojet: The Early Bird Catches the Lowest Fares http://t.co/l2aLrKUfIy https://t.co/oE4tL8JOSc
@FlyJambojet	positive	Thou I missed my flight to Eldoret,the lady at customer service desk at JKIA went her away to explain me as Layman @FlyJambojet thumbs up!
@FlyJambojet	neutral	@FlyJambojet @ticketsasa on the 10th of November, i.e
@flyethiopian	positive	flyethiopian Congrats
@flyethiopian	positive	I fly Ethiopian Airlines..... The safest Airliner... Come and visit us.
@flyethiopian	positive	Ethiopian Airlines to Start Doha Service from Dec 2014 http://t.co/Gsupat1pJv
@flyethiopian	negative	RT @AirTravellerorg: #Ethiopian Airlines Delays #Tokyo Launch http://t.co/kIVeCmCtwL
@flyethiopian	positive	SPECIAL DISCOUNTS AND OFFERS ON ETHIOPIAN AIRLINES.
@flyethiopian	neutral	RT @airlineroute: Ethiopian Airlines Nov 2014 Kinshasa/Brazzaville Service Changes http://t.co/4ufdMOLakc
@flyethiopian	positive	Ethiopian Airlines to take passengers from Dublin to Los Angeles next year http://t.co/3kUe2joef4 via @IrishTimes
@flyethiopian	negative	@theboybutler @HBaldwinMP @DouglasCarswell Didn't Ethiopian Airlines recently buy a few Airbuses ?
@flyethiopian	neutral	Contractair Ltd: B777 Captains - Ethiopian Airlines- *New Improved Terms, BONUS* http://t.co/ZugGoGcjko
@flyethiopian	neutral	Ethiopian Airlines flight from the 1950s
@flyethiopian	positive	RT @flyethiopian: Ethiopian won Best African Airline of the Year & Best African Airline to West Africa 2014 Award by AKWAABA .

4.2.2 Evaluation

The simplest metric that can be used to evaluate a classifier is the accuracy. It measures the percentage of inputs in the test set that the classifier correctly labels. Confusion matrices are tables where each cell indicates how often a label was predicted when the correct label was another.

4.3 Cross-validation

Cross validation is the de facto approach used to assess the accuracy and validity of a statistical model. It is used for this study as well. A K-fold cross validation is a technique that performs multiple evaluations on different test sets before combining the scores from those evaluations. The original corpus is subdivided into N subsets called folds. A valid model should show good predictive accuracy

4.3.1 Precision

It is a quantity within cross validation that represents a fraction of retrieved instances that are relevant. It is also called positive predictive value. High precision means that an algorithm returned substantially more relevant results than irrelevant.

4.3.2 Recall

Recall on the other hand is the fraction of relevant instances that are retrieved. A high recall means that an algorithm returned most of the relevant results. It is also called sensitivity.

4.3.3 F1-Score

F1 score can be interpreted as a weighted average of both precision and recall. It reaches its best at 1 and its worst at 0.

4.4 Evaluation

From the results illustrated in figure 7, those computed metrics qualify the model as well above average. It can be said that it has the ability to classify unseen data with accuracy close to a human being. This is because it has a precision average of 0.8, a recall average of 0.79 and an F-score of 0.83.

Table 4.1 : Cross validation results

	Precision	**Recall**	**F-1 Score**
Cross Validation 1	0.55	0.56	0.55
Cross Validation 2	0.82	0.71	0.73
Cross Validation 3	0.90	0.78	0.77
Cross Validation 4	0.76	0.78	0.75
Cross Validation 5	0.95	0.96	0.95
Cross Validation 6	0.82	0.87	0.84
Cross Validation 7	0.77	0.86	0.79
Cross Validation 8	0.93	0.93	0.93
Cross Validation 9	0.78	0.74	0.76
Cross Validation 10	0.77	0.49	0.51

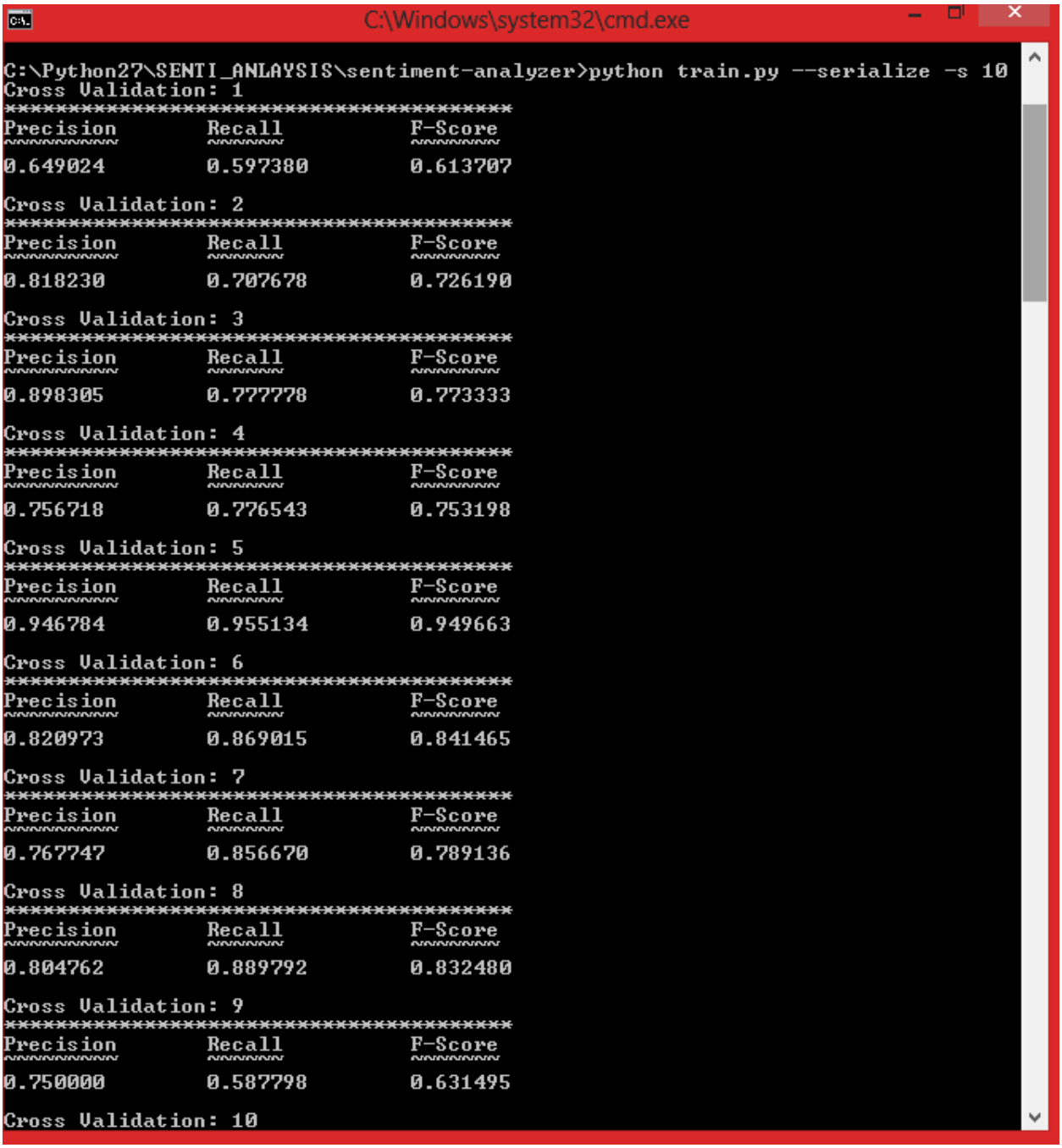

Fig 4.2: Cross validation (screenshot)

Fig 9 below actualizes the metrics above after running an instance of the model.

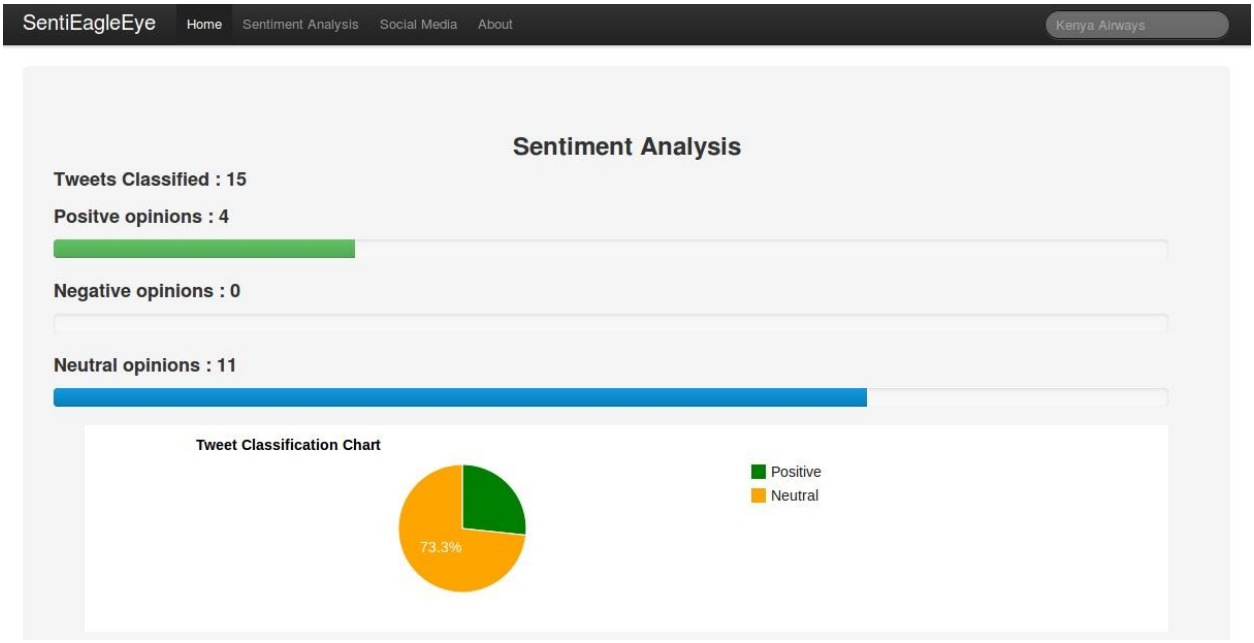

Fig 4.3. Sentiment Analysis instance.

A sample of the classified opinions are displayed after an instance and the performance of the model can be weighed by going inspecting the output. This in essence portrays the staging of the model metrics.

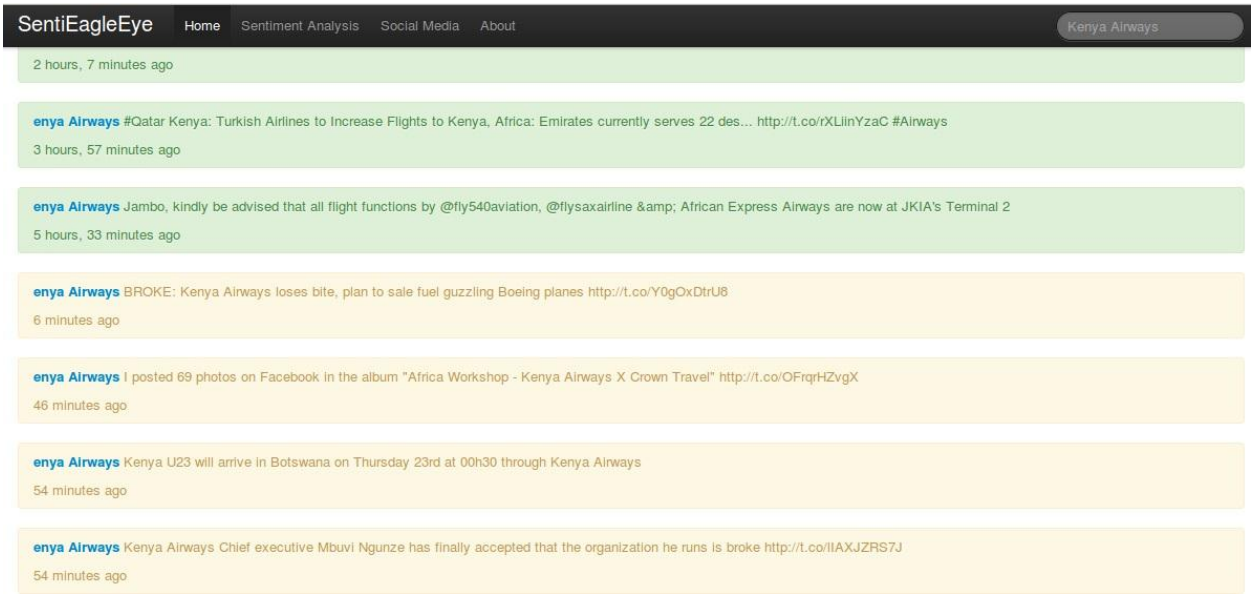

Fig 4.4 Sample Classified Sentiments.

CHAPTER FIVE

CONCLUSION

5.1 Findings and Conclusion

Sentiment analysis is increasingly becoming a vital phenomenon in our digital world. In Business, it can play a major role as it surveys on the various discussions within the environment of operation. In this research work, we strived to build a machine learning model that can accurately classify social media data into three possible classes: positive, negative, or neutral. The model integrates three machine learning classification algorithms. The algorithms operate by logic to try to counterbalance each otherôs performance so as to boost the overall performance. It is then implemented into a prototype that harvests public opinions about an airline company, classify them appropriately and save them with a sentiment score. This sentiment scores are queried from the database and used to generate visualization.

Evaluation of the model proved the classifiers metrics to be statistically good. Two cross validations are done performed: a fivefold and a tenfold. The tenfold cross validation which is considered the standard metric for this particular study consist of average scores of 0.81 for precision, 0.77 for recall and 0.76 for F-1 score. Precision in simple terms means that the classifier returned substantially more relevant results than irrelevant, recall on the other hand means that the classifier returned most of the relevant results. Our results as defined and discussed earlier are close to human accuracy. It can be concluded that the classification model is able to classify data within our domain, correctly with reasonable level of accuracy.

A corpus encompassing the diversity of languages within the East African region does not exist. There is need to correctly harvest and store data specific to this problem area. Therefore, a study was carried out in theEastern African airline marketspace to effectively develop a corpus, identify challenges and hence come up with appropriate solutions. This study aimed at building a sentiment analysis model that implements three machine learning algorithms to create an ensemble model that can be used for marketing intelligence.

The data was collected using twitter Application Programming Interface. This data included phrases and words about Kenya Airways, Fly Jambojet and Ethiopian Airlines which were first

cleaned and passed through the developed ensemble model to classify them as positive, negative or neutral. The opinions collected however contained Amharic alphabet for the Ethiopian Airlines which were unrecognizable to the English alphabet. This challenge led to the elimination of several opinions during cleaning which could be vital for several decisions. There were also a number of opinions aired through pictures and videos which were not captured in this study. These sentiment scores provide precious actionable knowledge about services, competitorsô information and the company at large to provide unprecedented dimensions in decision making.

Django python web framework was used to integrate the model build. It is set in a virtual environment where all its dependencies are made available. The application provides user friendly features. The same architectural approach can be used in other marketing domains of closely similar characteristics with speculated results above average.

From the study, it became evident that there is a great need for a corpus comprising of East African languages to create a pool from which this airlines and other major stakeholders can draw meaning and insights from. The Airlines greatly benefit from the model since a part from creating a trend analysisapproach, it also surfaces several integral issues to their companies and the whole industry at large.

The model would complement the marketing functionality of a business by providing useful insights from social media data. In particular, it will identify positive, negative and neutral sentiments about the market.

5.2 Limitation of study

A number of limitations were faced as listed below:

1. Language diversity. Some of important information could not be harvested as they were in different languages. Ethiopian Airlines for example is made of a public community communicating in Amharic.
2. The prototype like many other natural language approaches cannot identify and learn from sarcasm.

3. Twitter provides for free API to harvest public data. It is however limited to only about 5000 tweets which can be as small a fraction of the whole dataset as 1%. Facebook also sell much of their data compared to the freely available public posts.
4. The sources of legitimate data is quite limited i.e, Fly Jambojet which is one component relevant for our study do not have an official LinkedIn account hence data from this particular platform could not be sufficiently utilized.
5. Environmental discrepancies is another issue, this can be explained as the number of activities involved in one business environment differentiating with the other.
6. Challenge in filtering data from Facebook since Facebook restricts running of queries with joins.

5.3 Recommendation for future work

1. Integration with multi and cross-lingual language dictionaries to carter for the dynamic nature of language user on social media.
2. Accommodate sentiments expressed in Google+ and LinkedIn.
3. Perform trend analysis on sentiments over a given time period.
4. Incorporate a fairly accurate cross-domain sentiment classification.

References

1. AlAhlqvist, Toni; Back, A., Halonen, M., Heinonen S (2008). ñSocial media road maps exploring the features triggerd by social mediaò. VTT Tiedotteita- Valtion Teknillinen T(2454): 13

2. Asur, Sitaram, and Bernardo A. Huberman."Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on.Vol. 1.IEEE, 2010.

3. Blair-Goldensohn, Sasha, et al. "Building a sentiment summarizer for local service reviews." WWW Workshop on NLP in the Information Explosion Era. 2008.

4. Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis."IEEE Intelligent Systems (2013): 1.

5. Cambria, Erik, et al. "Semantic multi-dimensional scaling for open-domain sentiment analysis." (2013): 1-1.

6. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST)2.3 (2011): 27.

7. Choi, Chi-Hwan, et al. "Sentiment Analysis for Customer Review Sites." (2013).

8. Cortes, Corinna, and Vladimir Vapnik."Support-vector networks." Machine learning 20.3 (1995): 273-297.

9. Dave, Kushal, Steve Lawrence, and David M. Pennock."Mining the peanut gallery: Opinion extraction and semantic classification of product reviews."*Proceedings of the 12th international conference on World Wide Web*.ACM, 2003.

10. Dietterich, Thomas G. "Ensemble methods in machine learning." *Multiple classifier systems*.Springer Berlin Heidelberg, 2000.1-15.

11. Dģeroski, Saso, and Bernard Ç enko. "Is combining classifiers with stacking better than selecting the best one?." *Machine learning* 54.3 (2004): 255-273.

12. Erragcha, Nozha, and Rabiaa Romdhane. "New Faces of Marketing In The Era of The Web: From Marketing 1.0 To Marketing 3.0." Journal of Research in Marketing 2.2 (2014): 137-142.

13. Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." Proceedings of LREC.Vol. 6. 2006.

14. Fersini, E., E. Messina, and F. A. Pozzi. "Sentiment analysis: Bayesian Ensemble Learning." *Decision Support Systems* 68 (2014): 26-38.

15. Gitau, E., and Miriti, E. (2011).An approach for Using Twitter to perform Sentiment Analysis in Kenya, University of Nairobi, Kenya.

16. Glance, Natalie, et al. "Deriving marketing intelligence from online discussion.ò Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.ACM, 2005.

17. Glance, Natalie, et al. "Deriving marketing intelligence from online discussion."Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.ACM, 2005.

18. Glance, Natalie, et al. "Deriving marketing intelligence from online discussion."Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.ACM, 2005.

19. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford (2009): 1-12.

20. Goldenberg, Jacob, Barak Libai, and Eitan Muller. "Talk of the network: A complex systems look at the underlying process of word-of-mouth." Marketing letters 12.3 (2001): 211-223.

21. Gonçalves, Pollyanna, et al. "Comparing and combining sentiment analysis methods." Proceedings of the first ACM conference on online social networks.ACM, 2013.

22. Guerra, Pedro Calais, Wagner Meira Jr, and Claire Cardie. "Sentiment analysis on evolving social streams: How self-report imbalances can help." Proceedings of the 7th ACM international conference on Web search and data mining.ACM, 2014.

23. Hu, Minqing, and Bing Liu."Mining and summarizing customer reviews."Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.ACM, 2004.

24. Humphreys, Lee, Phillipa Gill, and Balachander Krishnamurthy. "How much is too much? Privacy issues on Twitter." Conference of International Communication Association, Singapore. 2010.

25. Jansen, Bernard J., et al. "Twitter power: Tweets as electronic word of mouth."Journal of the American society for information science and technology 60.11 (2009): 2169-2188.

26. John, George H., and Pat Langley."Estimating continuous distributions in Bayesian classifiers."*Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.

27. Kadam, Sachin A., and Mrs Shweta T. Joglekar. "Sentiment Analysis, an Overview." International Journal of Research in Engineering & Advanced Technology, 1/4, p1 7 (2013).

28. Kamps, Jaap, et al. "Using wordnet to measure semantic orientations of adjectives." (2004): 1115-1118.

29. Kietzmann, Jan H., et al. "Social media? Get serious! Understanding the functional building blocks of social media." Business horizons 54.3 (2011): 241-251.

30. Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." ICWSM 11 (2011): 538-541.

31. Kumar, Akshi, and Teeja Mary Sebastian."Sentiment analysis on twitter."IJCSI International Journal of Computer Science Issues 9.3 (2012): 372-378.

32. Lin, Yu-Ru, et al. "Voices of victory: A computational focus group framework for tracking opinion shift in real time." Proceedings of the 22nd international conference on World Wide Web.International World Wide Web Conferences Steering Committee, 2013.

33. Malaval, Philippe, and Christophe Bénaroya. Aerospace marketing management.Springer Internat.Publ., 2013.

34. Mangold, W. Glynn, and David J. Faulds. "Social media: The new hybrid element of the promotion mix." Business horizons 52.4 (2009): 357-365.

35. Marketing intelligence, 2013. Available from: http://www.businessdictionary.com/definition/marketing-intelligence.html.[15 July 2014].

36. Mauss, Iris B., and Michael D. Robinson. "Measures of emotion: A review."Cognition and emotion 23.2 (2009): 209-237.

37. Mcdonald, Ryan, et al. "Efficient large-scale distributed training of conditional maximum entropy models." Advances in Neural Information Processing Systems. 2009.

38. Mehdi Ghasemi ,RasoolSajediRaeisi Seyyed Ali NabaviChashemi. The Role of Knowledge Management on Marketing Intelligence of Employees of an Organization

(Case Study: Insurance Companies of Mazandaran Province). International Research Journal of Applied and Basic Sciences (2013).

39. Melville, Prem, Wojciech Gryc, and Richard D. Lawrence."Sentiment analysis of blogs by combining lexical knowledge with text classification." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.ACM, 2009.

40. Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." *arXiv preprint arXiv:1308.6242* (2013).

41. Mukherjee, Subhabrata, and Pushpak Bhattacharyya. "Feature specific sentiment analysis for product reviews." Computational Linguistics and Intelligent Text Processing.Springer Berlin Heidelberg, 2012.475-487.

42. Nigam, Kamal, John Lafferty, and Andrew McCallum."Using maximum entropy for text classification." IJCAI-99 workshop on machine learning for information filtering.Vol. 1. 1999.

43. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. 2010.

44. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.Association for Computational Linguistics, 2002.

45. Philip Kotler. Marketing Management, Millenium Edition.Tenth edition. (2002).

46. Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang,òMining Product Reviews Based on Shallow Dependency Parsingò, SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009

47. Quirkôs marketing research 2004, Marketing intelligence vs marketing research. Available from: < http://www.quirks.com/articles/a2004/20041209.aspx> [15 July 2014]

48. Rambocas, Meena, and João Gama. Marketing Research: The Role of Sentiment Analysis.No. 489. Universidade do Porto, Faculdade de Economia do Porto, 2013.

49. Rost, Mattias, et al. "Representation and communication: challenges in interpreting large social media datasets." Proceedings of the 2013 conference on Computer supported cooperative work. ACM, 2013.

50. Sachin A. Kadam, Shweta T. Joglekar Sentiment: Analysis an overview. International Journal of Research in Engineering & Advanced Technology, (2013).

51. Sangani, Chirag, and Sundaram Ananthanarayanan."Sentiment Analysis of App Store Reviews." Methodology 4: 1.

52. Siqueira, Henrique, and Flavia Barros."A feature extraction process for sentiment analysis of opinions on services." *Proceedings of International Workshop on Web and Text Intelligence*. 2010.

53. Tan, Songbo, and Jin Zhang."An empirical study of sentiment analysis for chinese documents." Expert Systems with Applications 34.4 (2008): 2622-2629.

54. Tan, Songbo, and Jin Zhang."An empirical study of sentiment analysis for chinese documents."*Expert Systems with Applications* 34.4 (2008): 2622-2629.

55. Turney, Peter D. "Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics.Association for Computational Linguistics, 2002.

56. Valkanas, George, et al. "Mining Twitter Data with Resource Constraints."

57. Varghese, Raisa, and M. Jayasree."A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING." 2013.

58. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2.6 (2012).

59. Wiebe, Janyce, and Ellen Riloff."Creating subjective and objective sentence classifiers from unannotated texts." Computational Linguistics and Intelligent Text Processing.Springer Berlin Heidelberg, 2005.486-497.

60. Wilson, Theresa, et al. "OpinionFinder: A system for subjectivity analysis."Proceedings of hlt/emnlp on interactive demonstrations.Association for Computational Linguistics, 2005.

61. Wollmer, Martin, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context." Intelligent Systems, IEEE 28.3 (2013): 46-53.

62. Xia, Rui, Chengqing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences 181.6 (2011): 1138-1152.

63. Xia, Yun-Qing, et al. "The unified collocation framework for opinion mining."Machine Learning and Cybernetics, 2007 International Conference on. Vol. 2.IEEE, 2007.

64. Xu, Kaiquan, et al. "Mining comparative opinions from customer reviews for Competitive Intelligence." *Decision support systems* 50.4 (2011): 743-754.

65. Ye, Qiang, Ziqiong Zhang, and Rob Law. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." *Expert Systems with Applications* 36.3 (2009): 6527-6535.

66. Yessenov, Kuat, and Saġa Misailovic. "Sentiment analysis of movie review comments." *Methodology* (2009): 1-17.

67. Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu,òPhrase Dependency Parsing for Opinion Miningò, EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, Volume 3

Appendices

Appendix 1 – Sample codes

Twitter harvesting code.

```
import sys

import time

from pymongo import Connection

from pymongo.errors import ConnectionFailure

from bson import json_util

import tweepy

from tweepy import API

from tweepy import Stream

from tweepy import OAuthHandler

from tweepy.streaming import StreamListener


ckey = 'enter your customer key'

csecret = ǿenter your customer secretô

atoken = 'enter your access token'

asecret = 'enter your access secret'


auth = OAuthHandler(ckey, csecret)

auth.set_access_token(atoken, asecret)
```

```
con = Connection(host="localhost", port=27017)

db = con.sentiEagleEye

twitterStreaming = db.twitterStreamingflyethiopian


track=('Ethiopian Airlines') # Identifies the phrase to track

api = API(auth)


search = []

page = 1

maxPage = 1000

while(page<=maxPage):

tweets = api.search(track)

for tweet in tweets:

search.append(tweet)

print tweet

for tweet in search:


        # Empty dictionary for storing tweet related data

data ={}

data['created_at'] = tweet.created_at

data['geo'] = tweet.geo

data['id'] = tweet.id

data['source'] = tweet.source

data['text'] = tweet.text
```

```
data['retweeted'] = tweet.retweeted

data['lang'] = tweet.lang

time.sleep(7)


    # Insert to MongoDB process

twitterStreaming.insert(data)

page = page+1
```

Facebook posts harvesting code.

```
import urllib2

import json

import sys

from pymongo import Connection

from pymongo.errors import ConnectionFailure


con = Connection(host="localhost", port=27017)

db = con.sentiEagleEye # Declares a MongoDB database

facebook = db.facebookPageStatus #Declares a MongoDB database collection


def create_post_url(graph_url, APP_ID, APP_SECRET):

        #create authenticated post URL
```

```
        post_args = "/posts/?key=value&access_token=" + APP_ID + "|" + APP_SECRET

        post_url = graph_url + post_args


        return post_url


def render_to_json(graph_url):
        #render graph url call to JSON

        web_response = urllib2.urlopen(graph_url)

        readable_page = web_response.read()

        json_data = json.loads(readable_page)


        return json_data



def main():
        #simple data pull App Secret and App ID

        APP_SECRET = "enter your Facebook APP_SECRET"

        APP_ID = "enter your Facebook APP_ID"


        #to find go to page's FB page, at the end of URL find username

        list_companies = ["PrideofAfrica", "FlyJambojet", "Ethiopianairlines"]

        graph_url = "https://graph.facebook.com/"


        for company in list_companies:
```

```
#make graph api url with company username

current_page = graph_url + company


#open public page in facebook graph api

json_fbpage = render_to_json(current_page)


#gather our page level JSON Data

page_data = (json_fbpage["id"], json_fbpage["likes"],

                        json_fbpage["talking_about_count"],

                        json_fbpage["username"])

print page_data


#extract post data

post_url = create_post_url(current_page, APP_ID, APP_SECRET)

json_postdata = render_to_json(post_url)

json_fbposts = json_postdata['data']


#facebook.insert(json_fbposts)


#print post messages and ids

for post in json_fbposts:


        print post

        facebook.insert(json_fbposts)
```

```
if __name__ == "__main__":
        main()
```

Train model code.

```
import argparse
import cPickle
import cProfile
import datetime
import numpy
import os
import scipy
import string

from sklearn import cross_validation
from sklearn import metrics
from sklearn import svm
from sklearn import naive_bayes
from sklearn.utils import check_arrays

import datasettings
```

```
from analyzer.parser import parse_imdb_corpus

from analyzer.parser import parse_training_corpus

from analyzer.vectorizer import SENTIMENT_MAP

from analyzer.vectorizer import Vectorizer




class Trainer(object):

    """Trains the classifier with training data and does the cross validation.

    """



def __init__(self):

        """Initializes the datastructures required.

        """

        # The actual text extraction object (does text to vector mapping).

        self.vectorizer = Vectorizer()



        # A list of already hand classified tweets to train our classifier.

        self.data = None



        # A list containing the classification to each individual tweet

        # in the tweets list.

        self.classification = None
```

```
        self.classifier = None
        self.scores = None


def initialize_training_data(self):
        """Initializes all types of training data we have.
        """
        corpus_file = open(os.path.join(datasettings.DATA_DIRECTORY,
                                        'full-corpus.csv'))

classification, tweets = parse_training_corpus(corpus_file)

        reviews_positive = parse_imdb_corpus(
os.path.join(datasettings.DATA_DIRECTORY, 'positive'))

        num_postive_reviews = len(reviews_positive)
        class_positive = ['positive'] * num_postive_reviews

        reviews_negative = parse_imdb_corpus(
os.path.join(datasettings.DATA_DIRECTORY, 'negative'))
        num_negative_reviews = len(reviews_negative)
        class_negative = ['negative'] * num_negative_reviews

        self.data = tweets
        self.classification = classification
```

```
        #self.date_time = date_time
        #self.retweet = retweets
        #self.favorited = favorited


def initial_fit(self):
"""Initializes the vectorizer by doing a fit and then a transform.
        """
        # We map the sentiments to the values specified in the SENTIMENT_MAP.
        # For any sentiment that is not part of the map we give a value 0.
        classification_vector = numpy.array(map(
lambda s: SENTIMENT_MAP.get(s.lower(), 0),
                                self.classification))

        feature_vector = self.vectorizer.fit_transform(self.data)

return (classification_vector, feature_vector)

def build_word_dict(self):
        """ Build sentiment dictionary and build vector of
weights for tweets.
        """
fileIn = open(os.path.join(datasettings.DATA_DIRECTORY,
                        'AFINN-96.txt'))
```

```
wordDict = {}
line = fileIn.readline()
while line != '':
temp = string.split(line, '\t')
wordDict[temp[0]] = int(temp[1])
line = fileIn.readline()
fileIn.close()

fileIn = open(os.path.join(datasettings.DATA_DIRECTORY,
                           'AFINN-111.txt'))
line = fileIn.readline()
while line != '':
temp = string.split(line, '\t')
wordDict[temp[0]] = int(temp[1])
line = fileIn.readline()
fileIn.close()

    word_dict_vector = []
for tweet in self.data:
        word_list = tweet.split()
sum = 0
for word in word_list:
if word in wordDict.keys():
sum += wordDict[word]
```

```
            word_dict_vector.append(sum)


    return word_dict_vector


def transform(self, test_data):
"""Performs the transform using the already initialized vectorizer.
        """
        feature_vector = self.vectorizer.transform(test_data)


def score_func(self, true, predicted):
        """Score function for the validation.
        """
    return metrics.precision_recall_fscore_support(
    true, predicted,
            pos_label=[
                SENTIMENT_MAP['positive'],
                SENTIMENT_MAP['negative'],
                SENTIMENT_MAP['neutral'],
                ],
    average='macro')


def cross_validate(self, k=10):
        """Performs a k-fold cross validation of our training data.
```

```
        Args:
k: The number of folds for cross validation.
        """
        self.scores = []

        X, y = check_arrays(self.feature_vector,
                            self.classification_vector,
                            sparse_format='csr')
cv = cross_validation.check_cv(
k, self.feature_vector, self.classification_vector,
classifier=True)

for train, test in cv:
self.classifier1.fit(self.feature_vector[train],
                     self.classification_vector[train])
self.classifier2.fit(self.feature_vector[train],
                     self.classification_vector[train])
self.classifier3.fit(self.feature_vector[train],
                     self.classification_vector[train])
            classification1 = self.classifier1.predict(
                self.feature_vector[test])
            classification2 = self.classifier2.predict(
                self.feature_vector[test])
            classification3 = self.classifier3.predict(
```

```
            self.feature_vector[test])


classification = []
for predictions in zip(classification1, classification2,
                       classification3):
            neutral_count = predictions.count(0)
            positive_count = predictions.count(1)
            negative_count = predictions.count(-1)
if (neutral_count == negative_count and
                negative_count == positive_count):
classification.append(predictions[0])
elif (neutral_count > positive_count and
                neutral_count > negative_count):
classification.append(0)
elif (positive_count > neutral_count and
                positive_count > negative_count):
classification.append(1)
elif (negative_count > neutral_count and
                negative_count > positive_count):
classification.append(-1)
classification = numpy.array(classification)


self.scores.append(self.score_func(y[test], classification))
```

```
def train_and_validate(self, cross_validate=False, mean=False,
serialize=False):
    """Trains the SVC with the training data and validates with the test data.

    We do a K-Fold cross validation with K = 10.
    """
    self.classification_vector, self.feature_vector = self.initial_fit()

    self.classifier1 = naive_bayes.MultinomialNB()
    self.classifier2 = naive_bayes.BernoulliNB()
    self.classifier3 = svm.LinearSVC(loss='l2', penalty='l1',
                                     C=1000,dual=False, tol=1e-3)

if cross_validate:
        self.cross_validate(k=cross_validate)
else:
self.classifier1.fit(self.feature_vector,
                             self.classification_vector)
self.classifier2.fit(self.feature_vector,
                             self.classification_vector)
self.classifier3.fit(self.feature_vector,
                             self.classification_vector)

if serialize:
```

```
        classifiers_file = open(os.path.join(
            datasettings.DATA_DIRECTORY, 'classifiers.pickle'), 'wb')
cPickle.dump([self.classifier1,
                    self.classifier2,
                    self.classifier3], classifiers_file)
        vectorizer_file = open(os.path.join(
            datasettings.DATA_DIRECTORY, 'vectorizer.pickle'), 'wb')
cPickle.dump(self.vectorizer, vectorizer_file)


return self.scores


def build_ui(self, mean=False):
        """Prints out all the scores calculated.
        """
for i, score in enumerate(self.scores):
print "Cross Validation: %d" % (i + 1)
print "*" * 40
if mean:
print "Mean Accuracy: %f" % (score)
else:
print "Precision\tRecall\t\tF-Score"
print "~~~~~~~~~~\t~~~~~~~\t\t~~~~~~~~"
precision = score[0]
recall = score[1]
```

```
            f_score = score[2]
print "%f\t%f\t%f" % (precision, recall, f_score)



print



def bootstrap():
    """Bootstrap the entire training process.
    """
parser = argparse.ArgumentParser(description='Trainer arguments.')
    parser.add_argument('-c', '--corpus-file', dest='corpus_file',
metavar='Corpus', type=file, nargs='?',
help='name of the input corpus file.')
    parser.add_argument('-p', '--profile', metavar='Profile', type=str,
nargs='?', help='Run the profiler.')
    parser.add_argument(
        '-s', '--scores', metavar = 'Scores', type=int, nargs='?',
help='Prints the scores by doing the cross validation with the '
        'argument passed as the number of folds. Cannot be run with -p '
        'turned on.')
    parser.add_argument(
        '-m', '--mean', action='store_true',
help='Prints the mean accuracies. Cannot be run with -p/-s turned on.')
```

```
    parser.add_argument(
        '--serialize', action='store_true',
help='Serializes the classifier, feature vector and the '
             'classification vector into the data directory with the same '
             'names.')
args = parser.parse_args()


trainer = Trainer()
    trainer.initialize_training_data()


if args.profile:
if isinstance(args.profile, str):
cProfile.runctx(
                'trainer.train_and_validate()',
                {'trainer': trainer, 'serialize': args.serialize},
                {}, args.profile)
print 'Profile stored in %s' % args.profile
else:
cProfile.runctx(
                'trainer.train_and_validate()',
                {'trainer': trainer, 'serialize': args.serialize},
                {}, args.profile)
else:
scores = trainer.train_and_validate(cross_validate=args.scores,
```

```
mean=args.mean,

serialize=args.serialize)

if args.mean:

        trainer.build_ui(mean=True)

if args.scores:

          trainer.build_ui()


return scores


if __name__ == '__main__':

scores = bootstrap()
```

Appendix 2 – Corpus subset.

kenya airways	positive	5.27E+17	2014-10-29T13:07:23.000Z	Love these 2 super stars @woowoopops @laurahind. Thanks for the sweet ride in bizclass @KenyaAirways @LandRoverUK
kenya airways	neutral	5.27E+17	2014-10-29T12:45:43.000Z	KenyaAirways
kenya airways	positive	5.27E+17	2014-10-29T12:39:09.000Z	RT @alykhansatchu: crude oil bear market represents a strong financial tailwind for airlines @BW @KenyaAirways
kenya airways	neutral	5.27E+17	2014-10-29T12:31:53.000Z	Ma3Route @KenyaAirways @chriskirwa more like matatu helicopters..
kenya airways	negative	5.27E+17	2014-10-29T12:24:31.000Z	jaredogeda Shouldn't this be a major concern if JKIA isnt accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAirways #KOT
kenya airways	neutral	5.27E+17	2014-10-29T12:18:59.000Z	RavS82 @Ma3Route @jaredogeda @Kenya_Airports @KenyaAirways #KOT not really a concern its us going backwards in life in past years
kenya airways	negative	5.27E+17	2014-10-29T12:18:21.000Z	RT @RavS82: @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isnt accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAi
kenya airways	negative	5.27E+17	2014-10-29T12:18:09.000Z	RT @RavS82: @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAi
kenya airways	neutral	5.27E+17	2014-10-29T12:18:07.000Z	alykhansatchu @BW @KenyaAirways Watching KQ share
kenya airways	negative	5.27E+17	2014-10-29T12:15:01.000Z	RT @RavS82 @Ma3Route @jaredogeda Shouldn't this be a major concern if JKIA isn't accessible via Mombasa Road?!?! @Kenya_Airports @KenyaAi
@FlyJambojet	neutral	5.28E+17	2014-10-31T16:31:08.000Z	RT @FlyJambojet: Dont miss your flight. You can check in online from 30 hours to 2 hours prior to departure of your flight on
@FlyJambojet	positive	5.28E+17	2014-10-31T16:18:52.000Z	RT @FlyJambojet: The Early Bird Catches the Lowest Fares http://t.co/l2aLrKUfIy https://t.co/oE4tL8JOSc
@FlyJambojet	positive	5.28E+17	2014-10-31T16:14:31.000Z	Thou I missed my flight to Eldoret,the lady at customer service desk at JKIA went her away to explain me as Layman @FlyJambojet thumbs up!
@FlyJambojet	neutral	5.28E+17	2014-10-31T15:53:27.000Z	Hope they operate here "@OliverCheruiyot: @EdgarYatich try @FlyJambojet we need you early tomorrow."
@FlyJambojet	negative	5.28E+17	2014-10-31T15:34:29.000Z	@FlyJambojet @qubanltd what flights have the special low rate in the month of November? can't see any via website.
@FlyJambojet	positive	5.28E+17	2014-10-31T16:18:52.000Z	RT @FlyJambojet: The Early Bird Catches the Lowest Fares http://t.co/l2aLrKUfIy https://t.co/oE4tL8JOSc
@FlyJambojet	neutral	5.28E+17	2014-10-31T15:23:14.000Z	@EdgarYatich try @FlyJambojet we need you early tomorrow.
@FlyJambojet	positive	5.28E+17	2014-10-31T14:53:11.000Z	RT @FlyJambojet: The Early Bird Catches the Lowest Fares http://t.co/l2aLrKUfIy https://t.co/oE4tL8JOSc
@FlyJambojet	positive	5.28E+17	2014-10-31T16:14:31.000Z	Thou I missed my flight to Eldoret,the lady at customer service desk at JKIA went her away to explain me as Layman @FlyJambojet thumbs up!
@FlyJambojet	neutral	5.28E+17	2014-10-31T13:03:17.000Z	@FlyJambojet @ticketsasa on the 10th of November, i.e
@flyethiopian	positive	5.30E+17	2014-11-04T05:05:11.000Z	Ethiopian Airlines to Start Doha Service from Dec 2014 http://t.co/Gsupat1pJv
@flyethiopian	neutral	5.29E+17	2014-11-04T03:01:06.000Z	RT @airlineroute: Ethiopian Airlines Nov 2014 Kinshasa/Brazzaville Service Changes http://t.co/4ufdMOLakc
@flyethiopian	positive	5.29E+17	2014-11-03T23:04:25.000Z	#Ethiopian Airlines goes to Dublin and Los Angeles - East African Business Week http://t.co/j4y25kVhHQ
@flyethiopian	positive	5.29E+17	2014-11-03T22:18:50.000Z	Ethiopian Airlines to take passengers from Dublin to Los Angeles next year http://t.co/3kUe2joef4 via @IrishTimes
@flyethiopian	negative	5.29E+17	2014-11-03T21:02:23.000Z	@theboybutler @HBaldwinMP @DouglasCarswell Didn't Ethiopian Airlines recently buy a few Airbuses ?
@flyethiopian	negative	5.29E+17	2014-11-03T10:49:55.000Z	RT @AirTravellerorg: #Ethiopian Airlines Delays #Tokyo Launch http://t.co/kIVeCmCtwL
@flyethiopian	positive	5.29E+17	2014-11-03T10:45:14.000Z	I fly Ethiopian Airlines..... The safest Airliner... Come and visit us.
@flyethiopian	neutral	5.29E+17	2014-11-03T10:29:26.000Z	Contractair Ltd: B777 Captains - Ethiopian Airlines- *New Improved Terms, BONUS* http://t.co/ZugGoGcjko

@flyethiopian	positive	5.29E+17	2014-11-03T07:38:02.000Z	SPECIAL DISCOUNTS AND OFFERS ON ETHIOPIAN AIRLINES.
@flyethiopian	neutral	5.28E+17	2014-10-31T02:47:39.000Z	Ethiopian Airlines flight from the 1950s
@flyethiopian	positive	5.28E+17	2014-10-31T09:44:08.000Z	flyethiopian Congrats
@flyethiopian	positive	5.28E+17	2014-10-31T09:32:56.000Z	RT @flyethiopian: Ethiopian won Best African Airline of the Year & Best African Airline to West Africa 2014 Award by AKWAABA .
@flyethiopian	positive	5.28E+17	2014-10-31T09:32:27.000Z	RT @flyethiopian: While Bole International Airport won Most Passenger Friendly International Airport (East Africa) Award by AKWAABA

Appendix 3 – Installation instructions.

a) Python and Virtual environment installation.

1. Python comes preinstalled in linux. If you are using windows, search, download and install python 2.7.
2. Virtual environment a tool to keep the dependencies required by different projects in separate places. It is installed by either using the setup tools or git cloning using the virtual-clone script https://pypi.python.org/pypi/virtualenv-clone .
3. After python and the virtual environment are set up, create an instance of an environment by entering the command: virtualenv virtual_environment_name. Activating the virual environment will be by the command: source virualenvname/bin/activate or C:\windows\path\to\created\env\activate for linux or windows respectively.

b) MongoDB Installation.

1. Download and install the latest version of MongoDB http://www.mongodb.org/ . N/B it is an open-source NOSQL document database.
2. You can create your own database and give it a name. This can be done by using the command: use database_name. Or you can just create the database by inserting documents into the named database.
3. A MongoDB database contains a collection of documents. A collection is created by the command: db.createCollection().

c) Facebook Scraper installation.

1. Facebook scraper used in this project sources its data from public Facebook webpages. First select the pages you want to scrape and enter them in the script search parameter area.
2. Copy the Facebook status harvesting code found in appendix one above and paste it in your development environment.
3. Using your Facebook details, login to your Facebook account and create an APP.
4. Using the credentials of your new app, enter the APP_SECRET and APP_ID in the code copied.
5. Ensure MongoDB server is running, run the Facebook scraping script either on your local machine or in the server.

d) Twitter Scraper installation.

1. Log in to your twitter account using your credentials. Redirect to the Twitter developers site and create an app. You will be issued with a set of the Appôs keys which are useful for authentication to the Twitter platform.
2. Copy the keys and fill them appropriately in the twitter scraping script. The keys should remain private to you.
3. Select the phrase you would like to consider and enter it in the track field within the script.
4. Ensuring the database server is running, run the script either in your local machine or in a server similar to the Facebook script.

e) Model Training.

1. Having collected sufficient number of opinions, you can check this by counting the documents in the collection by using *db.collectionName.count()*, export the harvested opinions into a comma separated value(CSV) format.
2. In the transfer the .csv file into the folder named data.
3. Run the training script to create a classifier and a vectorizerpickle which are used to introduce the classified model into the django web framework.

f) sentiEagleEye installation.

1. Using the python setup tools, install django 1.5.8 and all other required packages listed in the requirements text.
2. Scrape other opinions and after appending them to the trained model, they will be classified as positive, negative or neutral.
3. You can also do a search which will return a set of classified tweets as positive, negative or positive.