Predicting Sales in E-Commerce using Bayesian Network Model

Everlyne Nasambu Wamukekhe

A Thesis submitted in partial fulfillment for the degree of Master of Science in Software Engineering in Jomo Kenyatta University of Agriculture and Technology

2015

DECLARATION

This thesis is my original work and has not been submitted for a degree in any other university

---------- ----------

Everlyne Nasambu Wamukekhe **Date**

This thesis has been submitted for examination with our approval as University Supervisors

---------- ----------

Prof.Waweru Mwangi **Date**

---------- ----------

Dr. Michael Kimwele **Date**

JKUAT, Kenya

ACKNOWLEDGEMENT

I am thankful to the Almighty God for the gift of life and the strength to carry out this research. I wish to express my deepest appreciation to my supervisors, Prof. Waweru Mwangi and Dr. Micheal Kimwele for all their assistance, valuable guidance and encouragement throughout the study. There is so much that they have taught me that I will carry with me for the rest of my life. I also appreciate Judy Gateri for moral support that she has provided to me .Special thanks go to my family members and friends for their untiring support and prayers during this thesis work. This journey would not have been possible without your constant inspiration, encouragement and unconditional support

DEDICATION

This thesis is dedicated to my husband Prof. Masinde and my children Paul, Natalia and Victor I thank you all for your numerous sacrifices and patience during the entire course.

TABLE OF CONTENTS

Content **Page**

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMNS

E- Commerce	**-**	Electronic Commerce
NPL	**-**	Natural Language processing
ROC	**-**	Receiver Operating Characteristics curve

DEFINITION OF TERMS

A tweet is an update message posted on ones profile on Twitter. A tweet is limited to 140 characters (Wanyama, 2012).

E-commerce refers to the use of the Internet and the Web to conduct digitally enabled commercial transactions between and among organizations and individuals (Pearson, 2010).

Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Stock at el., 2010).

Social media is an umbrella term that describes websites that connect individuals somehow. A hallmark of social media is the user generated content. This model contrasts with the editorially controlled style of old media. Social media is sometimes called Web 2.0. (Wanyama, 2012).

Social Network is a social structure made up of persons called "nodes", which are tied or connected by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige (Wanyama, 2012).

ABSTRACT

Social media has become an increasingly important part of our daily lives in the last few years. With the convenience built into smart devices, many new ways of monitoring and predicting products sales have been made possible via social-media applications. An area of substantial research is that of predicting product sales, such as books, video games and movie tickets. There are a number of prediction models that have been used to predict future sales however these models attempt to solve the problem by making assumptions. These models assume that independent variables are truly independent. In theory, there should be zero correlation between any of the independent variables. In practice, however, many variables are related, sometimes quite highly. Therefore, different prediction techniques/methods have been and are being researched on and proposed to address this drawback. The aim of this study was to identify ways of improving prediction of product sales in mobile phones. Consequently, the study realized a predictive model that could classify sentiments from social media by combining natural language processing and the predictive model to compute the probability and present an improved predictive model. The process involved analyses of sentiments from Facebook and Twitter. A predictive model was created that performed classification on 300 annotated Facebook and Twitter sentiments. We compared the result of our model against open source model such as Markov model. The naïve bayes-model recorded a total precision of 93.33% while the receiver operating characteristic curve was 97%. The model predicted 150 of the sentiments belong to preference class No with precision of 96.43%. This means that the model correctly predicted the sentiments to be in class No with 96.43% accuracy. We therefore conclude from the receiver curve that the performance of the model used in this study to analyze data is acceptable and hence the posterior probabilities generated are informative. Markov model recorded a total precision of 91.67% while the receiver operating characteristics curve was 97.86%. The model predicted 161 of the sentiments belong to preference class No with precision of 98.57%. This means that the model correctly predicted the sentiments to be in class No with 98.57% accuracy. When we compare the two models naïve bayes is better because it has a high precision of 93.33% while Markov had a precision of

91.67%. The results obtained from experiments with the model indicate that it is capable of performing classification with an accuracy of 93.33% for sentiments obtained from Social Media. This is near human accuracy, as apparently people agree on sentiment only around 80% of the time. Most of the sentiments in this data are expressed partly in informal language. It can therefore be concluded that the model of classification has proved to be very accurate and efficient in predicting sales in e-commerce. This will assist the phone manufacturing companies in predicting the future levels of sales of their products.

CHAPTER ONE
INTRODUCTION

1.1 Background to the study

Introduction of the World Wide Web, electronic commerce has revolutionized traditional commerce and boosted sales and exchange of merchandise and information. E-commerce refers to the use of the Internet and the Web to conduct business transactions. A more technical definition is that e-commerce involves digitally enabled commercial transactions between and among organizations and individuals (Neck & Manz, 2010).

The Internet and other digital networks are the driving forces behind a dramatic change in the way business dealings are conducted. Increasingly, networks and information technology are being used to electronically design, market, buy, sell, and deliver products and services worldwide. This Internet Economy has seen tremendous growth and with latest fastest internet connectivity technologies has made internet service to be available to almost all consumers; this has presented all marketers to trade via e-commerce websites (Maury & Kleiner, 2002).

The expansion of the Internet in the past decade has given researchers new avenues to explore the art of prediction. First, online forums and blogs have allowed individuals to share thoughts, opinions and information with one another on any imaginable subject. Dhar and Chang (2009) examined the predictive ability of Internet chatter from user-generated content on blogs and forums and its impact on music album sales. It was concluded that online chatter predicted album sales during the first two weeks following the album release and the week preceding the album release. In a similar research (Mishne & Glance 2006) predicted movies and found correlation between references to movies in weblog posts both before and after their release and the movies' financial success.

The growth in size and popularity of social media sites like Facebook and Twitter has given researchers investigating the science of prediction another source of data (Dhar & Chang, 2009). Twitter in particular has become a popular website whereby stock market investors have increasingly turned to as an investment tool. Twitter's

appeal as an investment tool lies in the user's ability to relay company information, investment ideas and market sentiment in a short, concise manner. Analyzing the impact of sentiments analysis in prediction (Assur & Huberman, 2010) found out that it had some impact especially after the movie had been released. Sentiments have become a pointer to stimulus in social media and a combination of sentiments and keywords increases the prediction accuracy.

Sentiment analysis and topic detection are two growing areas in Natural Language Processing (NLP), and there are increasing trends of using them in social media analytics. Many companies use sentiment analysis to mine information about what people think and feel about their products, while political organizations use it to gather information about parties the people support. Topic detection is another emerging trend in social media analytics, and marketing companies use it to find out the current subjects people are talking about and the emerging topics in which people are interested (Phua, 2013). Sentiment analysis has often been used to identify attitudes of people towards certain products or political views. Elaborating on a comprehensive literature about the various methods used in opinion mining and sentiment analysis, the most basic approach considers whether a document or a word or phrase within the document contains positive or negative sentiment. Other more complex approaches perform ranking of attitudes into more than two classes (i.e., "star" ratings) and tries to find the sources and targets of these attitudes (Pang & Lee, 2011)

According to (Zhang et al, 2011), Bayesian prediction has been used in predicting sales in e-commerce. The model has been used in predicting music sales and proved to be meaningful. A hierarchical Bayesian model was developed based on a logistic diffusion process that allowed for the generalization of various adoption patterns out of discrete data and was applied in a situation where the eventual number of adopters was unknown. It was based on prelaunch data such as the success of previous records and updated sequentially when the first sales data of the participant record were available (Lee, 2003). Naïve Bayesian learning has been adversely used to study adoption behaviours in social media by predicting individuals' adoption probabilities from observed adoption data.

1.2 Statement of the problem

The growth in size and popularity of social media sites like Facebook and Twitter has enabled researchers to use it as another source of data for prediction (Asur and Huberman, 2010). Research on predicting product sales, such as books, video games and movie tickets has shown that, the evolution of blog posts over time has exhibited positive correlation with book sales and in some cases is able to predict spikes in sales (Gruhl, et al, 2005).

Despite the existence of a number of prediction models that have been used to predict future sales, these models attempt to solve the problem by making the assumptions. They assume that independent variables are truly independent; they cannot elegantly handle missing values and hence assumption is made about the missing data to give a value, they assume fixed increments/decrements in the score values for variables on an increment scale (Allison, 2012).

A large body of research has focused on prediction models for future sales in e-commerce. Using Twitter data for predicting film box office revenues during the opening and second weeks of each movie (Asur and Huberman 2010) found a tweet rate time series exhibited strong correlation with movie earnings. Examining the predictive ability of Internet chatter from user-generated content on blogs and forums and its impact on music album sales, (Dhar & Chang, 2009) concluded that online chatter is predictive of album sales during the first two weeks following the album release and the week preceding the album release. Taking a similar approach with movies (Mishne & Glance, 2006) found correlation between references to movies in weblog posts both before and after their release and the movies' financial success.

This research aims at using the the naïve bayes for predicting sales in e-commerce using social media platform such as Facebook and Twitter as ways of predicting product sales in mobile phones. Consequently, the study will realize a predictive model that will classify sentiments from social media by combining semantic web and the predictive model to compute the probability and present an improved predictive model. The process will include analyzing sentiments from Facebook and

Twitter. This will assist the companies in future to predict product sales and improve prediction levels of sales in e-commerce.

1.3 General objective

The general objective of this study was to develop a predictive model and implement the proposed model by studying and comparing existing predictive models.

1.4 Specific objectives

The specific objectives of the study are to

i. Identify models used in prediction in e-commerce.
ii. Identify the techniques used to determine the number of sentiments and keywords in social media.
iii. Develop a predictive model to help business to predict sales

1.4.1 Research questions

The study was guided by the following questions:

i. What are the models that have been previously used in prediction?
ii. Which techniques that are used in determining number of sentiments in social media?
iii. How will the predictive model help businessmen to predict sales?

1.5 Proposed predictive model

In this study the research has compared the existing models and come up with Naïve Bayes classification method to predict sales in e-commerce. The prediction model has been chosen in preference to other models because it has the following advantages as compared to the other models that are used for prediction.

1. **Consistent, theoretically solid mechanism for processing uncertain information**

   Probability theory provides a consistent calculus for uncertain inference, meaning that the output of the system is always unambiguous.

2. **Smoothness properties**

   Bayesian network models have been found to be very robust in the sense that small alterations in the model do not affect the performance of the system

dramatically. This means that maintaining and updating existing models is easy since the functioning of the system changes smoothly as the model is being modified. For sales and marketing systems this is a crucial characteristic, as these systems need to be able to follow market changes rapidly without complex and time consuming remodelling.

3. **A theoretical framework for handling expert knowledge**

   In Bayesian modelling, expert domain knowledge is coded as prior distributions, prior meaning that the probability distributions are defined before and independently of processing any possible sample data. This allows for combining expert knowledge with statistical data in a very practical way. Using suitable prior distributions, the priors are given a semantically clear explanation in terms of the data (expert knowledge can be interpreted as an unseen data-set of the same form as the training data).

4. **A clear semantic interpretation of the model parameters**

   Unlike neural network models, which usually appear to the user as a ``black box", all the parameters in Bayesian networks have an understandable semantic interpretation. It is for this reason that Bayesian networks can be constructed directly by using domain expert knowledge, without a time-consuming learning process.

5. **Flexible applicability.** Bayesian networks model the problem domain as a whole by constructing a joint probability distribution over different combinations of the domain variables. This means that the same Bayesian network model can be used for solving both discriminative tasks (classification) and regression problems (configuration problems and prediction). Besides predictive purposes, Bayesian networks can also be used for explorative data mining tasks by examining the conditional distributions, dependencies and correlations found by the modelling process.

1.6 Justification of the study

Social media has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter and JISC listservs on the academic side. 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 that range from the environment and politics to technology and the entertainment industry (Asur & Huberman ,2010). An emerging community of researchers has utilized social media data for a wide variety of purposes, for example, to predict stock market movements e.g., (Thelwall, et al, 2011), to predict announcements of flu outbreaks (Lampos & Cristianini, 2010), to forecast box-office revenues for movies (Asur & Huberman, 2010) and even to predict election outcomes (Gayo-Avello, 2012) found out that real world events have an impact in online systems and trails left by users on such systems have been used to perform early event detection in an automatic fashion.

One can also build models to aggregate the opinions of the collective population and gain useful insights into their behaviour, while predicting future trends. Moreover, gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns (Jansen et al, 2009).

1.7 Scope of the study

1) The objective is to study the predictive models
2) The research covers binary classifications and Naive Bayes algorithm
3) The research uses the Natural Language Processing and sentiments analysis.

1.8 Limitation of the study

A number of limitations were faced in this study as listed:

1. The use of positive words preceded by negations such as ‘not’ in negative sentiments led to erroneous classifications since the classifier uses the bag of words model, which assumes every word is independent. It cannot therefore learn that "not great" is a negative.
2. Based on the data obtained from Kenyan opinions we observed that the language used is mostly informal (slang). Kenyan slang is constantly changing coupled with the idea of lack of inadequate literature on it since it is informal. This made it challenging during classification of data.
3. Consequently, some of the keywords searched for would fetch sentiments done in irrelevant and unknown languages.

1.9 Organization of the thesis

Chapter one provides an orientation to the study on embedding semantic web into sales prediction using Bayesian Network Model. In this regard, the background to the study and its justification and scope was discussed. The problem was stated as a research question and broken down into research sub-questions to guide the researcher during the inquiry.

Chapter two involves an elucidation of the literature review of the existing work and that formed the basis of the study.

Chapter three provides a description of the methodological approach to the study. Chapter four describes how data was analysed and interpreted as well as feature selection for the model building.

Chapter five presents the model design and the training module dissection. Chapter six illustrates the description of the prediction model in terms of intelligent interface and the knowledge base.

Chapter seven summarizes the work presented in this thesis, drawing the main contribution. An outline of future research work is also provided.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter conceptualizes the literature review that defines the study. It analyses the models that have been previously used in prediction, techniques that are used in determining number of sentiments in social media and conceptualizes a predictive model useful for future trends of sales.

2.1.1 Social media

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinions on all kinds of topics and events. Federal Departments and Agencies defines social media as, “Web 2.0” or “Gov 2.0,” web-based tools, websites, applications, and media that connect users and allow them to engage in dialogue, share information, collaborate, and interact. Social media websites are oriented primarily to create a rich and engaging user experience. In social media, users add value to the content and data online; their interactions with the information (e.g., both collectively and individually) can significantly alter the experiences of subsequent users (Federal Departments & Agencies, 2009).

Social media is an umbrella term that describes websites that connect individuals somehow. A hallmark of social media is the user generated content. This model contrasts with the editorially controlled style of old media. Social media is sometimes called Web 3.0. The best way to define social media is to break it down (Wanyama, 2012). Media is an instrument on communication, like a newspaper or a radio, so social media is a social instrument of communication. In Web 3.0 terms, this would be a website that does not just give you information, but interacts with you while giving you that information. This interaction can be as simple as asking for your comments or letting you vote on an article, or it can be as complex as the process of recommending movies to a user based on the ratings of other people with similar interests (Wanyama, 2012).

Regular media is synonymous to a one-way street where you can read a newspaper or listen to a report on television, but you have very limited ability to give your thoughts on the matter. Social media, on the other hand, is a two-way street that gives you the ability to communicate too. Web 3.0 is a category of new Internet tools and technologies created around the idea that the people who consume media, access the Internet, and use the Web should not passively absorb what is available; rather, they should be active contributors, helping customize media and technology for their own purposes, as well as those of their communities. These new tools include, but are by no means limited to, blogs, social networking applications, RSS, social networking tools, and Wikis (Corbo, 2012).

2.2 Growth of social media

According to (Wanyama, 2012), the 1990s ushered in a number of innovations that were exciting to Internet communities at the time. The earliest versions of web pages and webcams emerged in 1991, as did the adoption of mp3 as a standard audio file. Geocities surfaced in 1997, providing a platform for individuals and businesses to operate their own websites within Geocities' virtual communities. Also debuting in 1997 were AOL's Instant Messaging and Six degrees, a social networking service that introduced the concepts of friends in social networking communities. This era's advances would prove to be foundational to social media platforms and activities down the road (Corbo, 2012).

The 2000s brought a tremendous growth in social networking platforms. Friendster opened its virtual doors in 2002, followed quickly in 2003 by Myspace, originally a Friendster clone. 2003 also brought the world Skype and Linkedin, representing the vast potential of VoIP and business networking respectively. The opening of Facebook in year 2004 represented one of the most well-known events in social media history. It was the year the term social media was accepted into mainstream consciousness. Twitter, was born in 2006 and is credited with facilitating monumental social change across the world. 2007 gave the world the iPhone, which is also now etched into cultural consciousness. It launched a fervent and more lasting interest in mobile Internet activities, a great many of them with a social element.

New options for Internet-based social interactions continue to arise, the recent launch of Google+ being one of them.

Analytics have changed over time to keep pace with the innovations in online community platforms. Marketing analysis has evolved to take advantage of the constantly growing reach of social media. Website statistics have expanded from simple analysis of Internet user behaviour to contemporary analysis that includes social analytics. By 2010, even URL shortening services such as bit.ly were providing analytics, to help individuals and businesses track link distribution (Corbo, 2012).

Over the years, people have demonstrated a strong desire to utilize the Internet to connect with others, to be social and find others with whom to relate and share interests. Businesses have studied these wishes and adapted to them, to tap into social networking tendencies as a way to find and cultivate customers, just as customers find and cultivate relationships amongst one another and also to get feedback from customers over services and goods offered. Whatever the future of Internet-based communication holds, it is assured that social media elements will remain significant. Wanyama (2012).

The Internet and other digital networks are the driving forces behind a dramatic change in the way business dealings are conducted. Increasingly, networks and information technology are being used to electronically design, market, buy, sell, and deliver products and services worldwide. This Ie has seen tremendous growth and with latest fastest internet connectivity technologies has made internet service to be available to almost all consumers; this has presented all marketers to trade via e-commerce websites (Maury & Kleiner, 2002).

2.3 Prediction through social media

The expansion of the Internet in the past decade has given researchers new avenues to explore the art of prediction. First, online forums and blogs allowed individuals to share thoughts; opinions and information with one another on any imaginable subject as a result of social media platforms. Websites like Twitter, Facebook, YouTube, Flickr, and others make it easy to reach large numbers of people.

Social media allows anyone who uses information to also create making it, an ideal platform for sharing information, starting conversations, and exchanging knowledge. The growth in size and popularity of social media sites like Facebook and Twitter has enabled researchers to use it as another source of data for prediction. Examing tweets from the site Twitter (Asur & Huberman, 2010) made predictions about the financial success of various movies using only tweets that preceded the release of a movie, there was a strong correlation between the amounts of attention a movie is given and its future financial success.

The predictive ability of Internet chatter from user-generated content on blogs and forums had impact on music album sales (Dhar & Chang, 2007). Their research it was concluded that online chatter is predicted album sales during the first two weeks following the album release and the week preceding the album release. Taking a similar approach, (Mishne & Glance, 2006) found correlation between references to movies in weblog posts, both before and after their release and the movies' financial success.

Stock market investors have increasingly turned to Twitter as an investment tool. Twitter's appeal as an investment tool lies in the user's ability to relay company information, investment ideas and market sentiment in a short, concise manner. As the Internet became more ubiquitous, various avenues have been pursued estimating future music sales where word-of-mouth had an impact effects on music blogs on sales figures, (Dewan & Ramaprasad, 2009) whereas (Hann, et al., 2011) considered how widely a forthcoming album circulates in P2P networks.

Tumarkin and Whitellaw (2001), focused their research on using chatter only from the once popular website RagingBull.com. Unlike previous research, they concluded that the message board activity of RagingBull.com did not have any predictive capabilities but did, however, show correlation between message volume and next day trading volume. Examining 1.5 million messages focused on 45 companies that were posted on Yahoo! Finance and RagingBull.com (Antweiler & Frank, 2004), concluded that stock market messages help to predict market volatility and that while economically small; stock market message boards do affect stock market returns in a statistically significant manner.

Previous research by (Das & Chen, 2007) developed a formal sentiment analysis tool and applied it to stock market message boards. They found no significant correlation between sentiment and stock price movements. They acknowledge this in their research and attribute it to the large amount of noise in stock market message boards as well as the lack of market power that many investors participating in online message boards have. Previous research has primarily used online message boards as a mean to aggregate investor sentiment.

Recent research has taken advantage of the emergence of social media and applied this towards financial markets. Bollen et al.,(2011) aggregated "tweets" from Twitter as a whole, choosing not to focus stock market specific "tweets," and examined the "mood" states of Twitter users and corresponding stock market movements. Signaling six different mood types that would reflect the mood of an individual (Calm, Alert, Sure, Vital, Kind & Happy), the conclusion showed the collective "mood" of Twitter users successfully predicted the upward and downward movement of the stock market.

An approach by (Welpe & Sprenger, 2010) used a more direct approach to examine the relation between Twitter messages and stock market movement by filtering out all non-market related "tweets." They did this by using messages only pulled from StockTwits.com. Using the S&P 100 as the market index for analysis, their research suggests that public sentiment conveyed through StockTwits.com aligns itself with the movement of the S&P 100 and is positively related to the volume of trading.

2.4 Sentiment analysis

In his research, (Wanyama, 2012) argues that other people's opinions are usually consulted when conducting research or making decisions. We consult political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. And now Internet has made it possible to find out the opinions of millions of people on everything from latest gadgets to political philosophies (Mejova, 2009).

The internet today contains huge quantities of data that keeps growing day by day. A large number of websites have come up that provide people with an opportunity to

express their opinion about certain things. These include product review sites, forums, blogs and social network sites. Consequently as a response to the growing availability of informal, opinionated texts like blog posts and product review websites, a field of Sentiment Analysis has sprung up in the past decade to find out what people feel about a certain issue. At eBay research labs expressed that properly conducted sentiment analysis provides more usable information than surveys and focus groups, where participants wanted to complete the process quickly or expressed less than honest opinions due to the influence of others. Sentiment Analysis is an active area of ongoing research. It is a field of study that aims touncover the attitude of the author on a particular topic from the written text.

2.4.1 Studies on sentiment research

Most research work in Sentiment analysis has been done on product and movie reviews, where it is easy to identify the topic of the text. This is because of the ready availability of product review datasets (Mejova, 2009). Conducting an extensive experiment on movie reviews using three traditional supervised machine learning methods (i.e., Naive Bayes (NB), maximum entropy classification (ME), and support vector machines (SVM)), (Pang et al., 2002) in his results indicated that standard machine learning techniques definitively outperform human produced baselines. However, he found that machine learning methods could not perform as well on sentiment classification as on traditional topic based categorization. Their results indicated that use of machine learning methods is more superior to the use of human generated baselines. The advantage is that reviews from such sites already have a clearly specified topic which is often assumed that the sentiments expressed in the reviews have to do with the topic.

Sentiment analysis on movie review data comments from the popular social network Digg as their data set and classified text by subjectivity/objectivity and negative/positive attitude (Yessenov & Misailovic, 2009). They used different approaches in extracting text features such as bag-of-words model, use of large movie reviews corpus, restricting to adjectives and adverbs, handling negations,

bounding word frequencies by a threshold, and using WordNet synonyms knowledge. They then evaluated their effect on accuracy of four machine learning methods - Naive Bayes, Decision Trees, Maximum-Entropy, and K-Means clustering. Their results showed that simple bag-of-words model performed relatively well and could be further refined by the choice of features based on syntactic and semantic information from the text.

Training a classifier emoticons to identify tweets that were to be used to train the classifier (Pak & Paroubek, 2009). Their tests were able to classify texts as positive, negative or neutral. However, the only language they supported in their text was English. Sentiment analysis explored the use of methodologies in the classification of web forum opinions written in multiple languages (Abbasi et al., 2007), evaluated the utility of stylistic and syntactic features in sentiment classification of English and Arabic content on movie review dataset, U.S and Middle Eastern web forum postings. Specific feature extraction components were integrated to account for the linguistic characteristics of Arabic. The Entropy Weighted Genetic Algorithm (EWGA) was also developed, which is a hybridized genetic algorithm that incorporates the information gain heuristic for feature selection. Their experiments indicated that stylistic features significantly enhanced performance across all test beds while EWGA also outperformed other feature selection methods, indicating the utility of these features and techniques for document level classification of sentiments.

Naive Bayes model was used by (Gitau & Miriti, 2011) to research on sentiment analysis using unigrams, emoticons and bigrams that were mined from Twitter on Kenyan issues. The model was able to perform polarity classification of tweets into positive and negative. The results achieved from their work suggest that the Naïve Bayes Model of classification could be used as a starting point with relative ease to perform Sentiment Analysis with good results on Social Media. They suggest further work to be done on additional Social Media players such as Facebook and Youtube.

2.4.2 Elements used in sentiment analysis

Five elements were identified by (Liu, 2010) in sentiment analysis include the object, attribute, opinion holder, opinion orientation and strength and time. An object refers to the target entity that has been commented on. It may have a set of components (or parts) and a set of attributes (or properties) which are collectively called the features of the object. We can use an example of a particular brand of cellular phone as an object. It has a set of components (e.g., battery and screen), and also a set of attributes (e.g., voice quality and size), which are all called features. An opinion can be expressed on any feature of the object and also on the object itself. For example, in “I like iPhone. It has a great touch screen”, then first sentence expresses a positive opinion on “iPhone” itself, and the second sentence expresses a positive opinion on its “touch screen” feature. Opinion holder: The holder of an opinion is the person or organization that expresses the opinion.

An opinion holder is the person who expresses opinion about the object. For product reviews and blogs, opinion holders are usually the authors of the posts. Opinion holders are more important in news articles because they often explicitly state the person or organization that holds a particular opinion. An opinion on a feature f (or object o) is a positive or negative view or appraisal on f (or o) from an opinion holder. Positive and negative are called opinion orientations.

2.5 Models used in predicting future sales

2.5.1 Linear regression model

Linear regression is used as a predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, and purchases of a country's exports, spending on imports, the demand to hold liquid assets, labour demand, and labour supply. Linear regression from text and metadata features was used to predict earnings for movies (Joshi et al., 2010).

Carrying out research on prediction (Sharda & Delen, 2006) have treated the prediction problem as a classification problem and used neural networks to classify movies into categories ranging from ’flop’ to blockbuster’. Apart from the fact that

they are predicting ranges over actual numbers, the best accuracy that their model could achieve was fairly low. While linear regression is a very powerful modelling tool, it assumes that the response variable (the log odds, not the event itself) is linear with respect to the predictor variables (Kharya, 2012). However the model has the following limitations:

1. **Only looks at linear relationships**

By its nature, linear regression only looks at linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them. Sometimes this is incorrect.

2. **Only looks at the mean of the dependent variable**

Linear regression looks at a relationship between the mean of the dependent variable and the independent variables. However, sometimes you need to look at the extremes of the dependent variable. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

3. **Sensitive to outliers**

Outliers are data that are surprising. Outliers can be univariate (based on one variable) or multivariate. If you are looking at age and income, univariate outliers would be things like a person who is 18 years old, or one who made Ksh.12 million last year. A multivariate outlier would be an 18-year-old who made Ksh.100, 000. In this case, neither the age nor the income is very extreme, but very few 18-year-old people make that much money. Outliers can have huge effects on the regression. You can deal with this problem by requesting influence statistics from your statistical software.

4. **Data must be independent**

Linear regression assumes that the data are independent. That means that the scores of one subject (such as a person) have nothing to do with those of another. This is often, but not always, sensible. Two common cases where it does not make sense are clustering in space and time. Regression cannot elegantly handle missing values on a variable - by - variable basis. This means that data must be lost, some assumption made about the missing data to give it a value. The model assumes fixed

increments/decrements in the score values for variables on an interval scale. May not capture, or at least make readily apparent, interactions in data. Categorical variables may have to be represented by dummy variables, i.e., multiple variables which represent the absence or presence of each component attribute in the predictor variable.

2.5.2 Neural network model

Neural networks outperform decision trees for prediction of churn (Au et al, 2003). They state the biggest disadvantage of neural networks is that they do not uncover patterns in an easily understandable form, categorizing them as a 'black box' model. The basic idea behind neural networks is that each attribute is associated with a weight and combinations of weighted attributes participate in the prediction task. During learning the weights are constantly updated, thus correcting the 'effect' which an attribute has. Given a customer data set and the set of predictor variables the neural network tries to calculate a combination of the inputs and to output the probability that the customer is a churner. Neural networks need a large volume of data set and a lot of time in order to calculate a reasonable weight age for the predictor attributes.

2.5.3 Decision trees

Decision Trees is the most popular data mining technique for classification problems. The principal idea of decision tree is to split your data recursively into sub-sets so that each subset contains more or less homogenous states of your target variable. At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When the recursive process is completed, a Decision Tree is formed (Choudhury et al, 2013). The development of such trees is done in two major steps: building and pruning. During the first phase the data set is partitioned recursively until most of the records in each partition contain identical value. The second phase then removes some branches which contain noisy data (those with the largest estimated error rate).

CART, a Classification and regression tree, is constructed by recursive splits of an instance into subgroups until a specified criteria has been met. The tree grows until the decrease of impurity falls below a user-defined threshold. Each node in a decision

tree is a test condition and the branching is based on the value of the attribute being tested. The tree is representing a collection of multiple rule sets. When evaluating a customer data set the classification is done by traversing through the tree until a leaf node is reached. The label of this leaf node (Churner or Non Churner) is assigned to the customer record under evaluation Lazarov and Marius (2007).

2.5.4 Bayesian Network Model

Bayesian networks (BNs), also known as belief networks (or Bayes nets for short), belong to the family of probabilistic graphical models (GMs). These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science, and statistics (Ben-Gal, 2007).

The Bayesian Network uses probability that is called Bayes' theorem or rule which is used to calculate the posterior probability which is a conditional probability of future uncertain event based on relevant evidence relating to it historically. Bayesian Networks is a directed acyclic graph (Pearl, 1988) and it is composed of qualitative parts and quantitative parts. The qualitative part is a directed graph consists of domain related variables and relationships between variables.

The quantitative part is consisted of aggregate probability distribution of domain related variables(Charniak, 1991). In these directed graphs, each node represents a variable, while each link indicates the relationship between two variables. In other words, such directed graph is an illustration of the distribution of aggregate probability of these variables. Bayesian prediction has been used in predicting music sales and has been proved to be meaningful approach in this regard. The model developed by (Lee, 2003), is based on prelaunch data such as the success of previous records and gets updated sequentially when the first sales data of the participant record are available.

Bayesian Network Model uses Naive Bayes which is a type of supervised-learning module that contains examples of the input-target mapping the model tries to learn. Such models make predictions about new data based on the examination of previous data. The Naïve Bayes algorithm uses the mathematics of Bayes' Theorem to make its predictions. The theorem is denoted by:

$$P(A/B) = P(B/A) * P(A)/P(B) \quad \text{(i)}$$

Bayes' Theorem states that the probability of a particular predicted event, given the evidence in this instance, is computed from three other numbers: the probability of that prediction in similar situations in general, ignoring the specific (the so called prior probability) multiplied with the probability of seeing the evidence we have here, given that the particular prediction is correct divided by the probability of that prediction in general.

2.5.5 Application areas of Bayesian Networks

Bayesian networks have been applied in various fields such as data mining, troubleshooting, bioinformatics/computational biology and medical diagnosis. Bayesian network has been used in cost minimizing troubleshooting strategies, (Jayech & Mahjoub, 2010), used Bayesian Networks to cluster the vectors of images to improve classification in pattern recognition. Bayesian networks are particularly interesting because they are able to encode data from experts and patient history. They also have the benefit that the results are easily explainable as compared to other methods (Madsen, 2010).

2.6 Techniques that are used to define sentiments and keywords in social media

2.6.1 Using lexicon based Vs learning based technique

Lexicon based techniques use a dictionary to perform entity-level sentiment analysis. This technique uses dictionaries of words annotated with their semantic orientation (polarity and strength) and calculates a score for the polarity of the document. Usually this method gives high precision but low recall. The lexicon-based approach (Taboadaet et al, 2010) determines the sentiment or polarity of opinion via some function of opinion words in the document or the sentence.

Learning based techniques require creating a model by training the classifier with labeled examples. This means that you must first gather a dataset with examples for positive, negative and neutral classes, extract the features/words from the examples and then train the algorithm based on the examples.

Using lexicon based techniques with large dictionaries enables us to achieve very good results. Nevertheless they require using a lexicon, something which is not always available in all languages. On the other hand learning based techniques deliver good results nevertheless they require obtaining datasets and require training.

2.6.2 Using Statistical Vs Syntactic techniques

Some approaches go beyond word-level, e.g., Wilson el at, (2008) used special features to model the existence of polarity modifiers in the syntactic context of a sentiment word. (Choi & Cardie, 2008) used syntactic patterns to treat content negators, an integrated polarity reversing words into a dependency tree based method.

Syntactic techniques can deliver better accuracy because they make use of the syntactic rules of the language in order to detect the verbs, adjectives and nouns (Liu, 2010). Unfortunately such techniques heavily depend on the language of the document and as a result the classifiers cannot be ported to other languages. Statistical techniques have probabilistic background and focus on the relations between the words and categories. Statistical techniques have two significant benefits over the Syntactic ones: they are used in other languages with minor or no adaptations and use Machine Translation of the original dataset and still get quite good results, (Nakagawa et al, 2010).

2.6.3 Feature Selection algorithm

Popular algorithms, including support vector machine (SVM) and reinforcement learning, have been reported to be quite effective in tracing the stock market and help maximizing the profit of stock option purchase while keep the risk low (Moody & Shaffel , 2001). In learning based techniques, before training the classifier, you must select the words/features that you will use on your model. You can’t just use all the

words that the tokenization algorithm returned simply because there are several irrelevant words within them (Yulan & Deyu, 2010).

Two commonly used feature selection algorithms in Text Classification are the Mutual Information and the Chi-square test. Each algorithm evaluates the keywords in a different way and thus leads to different selections. Also each algorithm requires different configuration such as the level of statistical significance, the number of selected features etc. Again you must use Trial and error to find the configuration that works better in your project (Gaurangi at el.,2014)

2.6.4 Natural Language Processing

Natural Language Processing (NLP) is the use of computers to process written and spoken language. NPL translates languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines (Rada et al., 2006). NLP relates to computer systems that process human language in terms of its meaning. Apart from common word processor operations that treat text like a sheer sequence of symbols, NLP considers the hierarchical structure of language: many words make a phrase, many phrases make a sentence and, ultimately, sentences convey messages. By analyzing language for its meaning, NLP systems have many other useful roles, such as correcting grammar, converting speech to text and automatically translating texts between languages (Sattikar & Kulkarni, 2012).

Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will (hopefully) be able to talk to the computer in their own language, rather than learn a specialized language of computer commands. For programming, however, the necessity of a formal programming language for communicating with a computer has always been taken for granted (Rada et al., 2006).

Many Natural Language Processing (NLP) techniques have been used in information retrieval. The results are not encouraging. Simple methods (stopwording, porter-style stemming, etc.) usually yield significant improvements, while higher-level processing (chunking, parsing, word sense disambiguation, etc.) only yield very

small improvements or even a decrease in accuracy. At the same time, higher-level methods increase the processing and storage cost dramatically. This makes them hard to use on large collections (Rada et al., 2006).

2.7 Summary

People perceive social media as a way of communication of information, ideas and opinions to one another. Social media allow users to interact, share and create content collectively. The Bayesian network model is the best suited for this research because Naïve Bayes Classifiers are fast, accurate, allow rich structure support for missing data during learning and classification, mixes expert data and opinion and are simple to implement.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the research methodology and specific approach that was adopted to investigate the appropriate predicting techniques for predicting future product sales and incorporating the results in the Bayesian prediction model as presented in Figure 3.1.

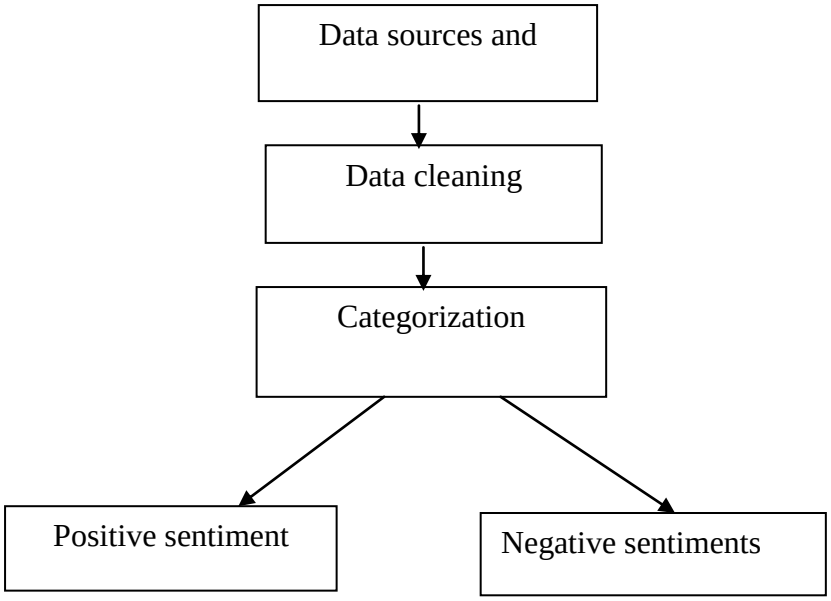

Figure 3. 1 Illustration of the predictive model

The developed model uses naïve Bayes algorithm to predict sentiments based on user satisfaction or dissatisfaction. Feature selection for the model was based on positive and negative sentiments posted by users on two social media platforms (Facebook and Twitter). The messages go through intelligent interface in order to be analyzed for prediction.

3.2 Research design

The research methodology that was employed for this research is qualitative and quantitative research. The research design involved the following activities: data sources and collection mechanisms, data cleaning and categorization into two

categories (positive and negative), classifier development, system design, and system implementation, run some test and evaluate the results and discuss results.

3.3 Target population

The sample population for this project were the social media websites. This is because people post sentiments on the social platform. In the research at least 300 sentiments were used. The research sought the use of two social media types namely Facebook and Twitter where 150 samples from each social media were collected. When developing machine learning systems, it is a requirement to have an annotated data for training and evaluation of the system and hence we considered data that was rich in both positive and negative posts. The target population for the study was 300 sentiments from both Facebook and Twitter. Five brands of mobile phones were used (Samsung, Tecno, Nokia, LG and Sony) each having 60 sentiments to have an equal representation. The researcher purposely targeted (Samsung, Tecno, Nokia, LG and Sony) because these are commonly used brands and the researcher believed the brands would have an adequate amount of posts on social media hence reliable for the study.

3.4 Sampling and sampling technique

In qualitative research, sampling can be based on probability, that is, a random sample, or on non-probability, that is, a purposive sample. Non-probability sampling requires the researcher to purposively select a section of the wider population to include or exclude from the sample because they illustrate some feature or process in which the researcher is interested, with the aim being for the sample to represent it rather than to seek generalizability, Silverman (2005). The research employed purposive sampling since the study used sentiments from specific social media websites. A purposive, or judgmental, sample is one that is selected based on the knowledge of a population and the purpose of the study. In a purposive sample those being interviewed fit a specific purpose or description. The sample that will be selected in this case is positive and negative sentiments (Cooper & Schindler, 2008).

3.5 Data collection method

The source of data for this study was Facebook and Twitter. Facebook and Twitter provide millions of users an opportunity to express their personal opinions about any

topics. Thus the data available on this site and other social network sites has immediate applicability in business environments such as gaining information from summarized views of users on certain products or services (Yessenov & Misailovic, 2009).

Data for this study was collected from Facebook and Twitter using Natural Language Processing Tool kit API because it generates a feature set for training the classifier using a bag of words, trains a Naive bayes classifier imported from nltk using the training set and reads in training data from files (Nielsen, 2011). These APIs are readily available and free for use in accessing data from the two social networks. In this study, Twitter and Facebook were selected since they have a large community of users. It is a fact that the style and nature of writing posts on Facebook and Twitter is not strictly formal.

3.6 Data cleaning criterion

The data collected therefore had to go through some cleaning exercise before it was used in training the classifiers. It was necessary to remove irrelevant characters such as URL links and repeated characters using php based natural processing language toolkit (alchemy nltk). This was meant to ensure that the classifiers were trained on appropriate data for better performance. This research used the positive and negative sentiments and the keywords posted on social media platforms.

3.7 Data analysis

Data analysis involved collection cleaning and categorization of data into positive and negative sentiments. The sentiments were derived from five brands of mobile phones (Samsung, Tecno, Nokia, LG & Sony). In addition, the study focused on online posts from Twitter and Facebook, mainly positive and negative sentiments.

3.8 Chapter summary

In conclusion this chapter highlighted the methodology that was applied in data analysis. It conclusively pointed at the research design, population and data collection and analysis criteria.

CHAPTER FOUR
DATA ANALYSIS

4.1 Introduction

This chapter describes the entire process of data selection and the process of defining the features used in model building. The study utilized the existing data of 300 posts from five brands of mobile phones (Samsung, Tecno, Nokia, LG & Sony) each having 60 sentiments to have an equal representation. The study focused on online posts from Twitter and Facebook, mainly positive and negative sentiments.

4.2 Data description

The sentiments of posts from the five brands of mobile phones (Samsung, Tecno, Nokia, LG and Sony) were further classified into operating system, features and hardware.

4.2.1 Brand name of the mobile phones

In terms of brand name, all the five leading brands had 20% of the sentiments. These percentages were used as prior probabilities in the bayes' model.

4.3 Operating systems of the mobile phones

The operating system was classified as either android or windows. Majority of the sentiments 239 (79.7%) were commented on the android phones with the other 61(20.3%) sentiments commenting on windows. The testing for os=1&os=2 was done for the operating system.

4.3.1 Features of the mobile phones

The features of the five brands of mobile phones (Samsung, Tecno, Nokia, LG & Sony) were classified into three categories; bad, moderate and good. There were 52(17.3%) sentiments describing features as bad, 184 (61.3%) moderate and 64(21.3%) describing features as good. These were sentiments for the five brands. The results are shown in Table 4.1.

Table 4. 1 Features of mobile phones

Features	Bad	Moderate	Good
Android	16.90%	65.40%	17.70%
Windows	19.0%	46%	35%

The features were further tested on where;

os=1 &features=1, os=1&features=2 and os=1&features=3 and os=2 &features=1, os=2&features=2 and os=2&features=3.

Most of the sentiment (android, 65.40% and windows, 46%) were moderate with (android, 16.9% and windows 19.0%) of the sentiments describing features as bad and the other (android, 17.70% and windows, 35%) describing the features as good.

4.3.2 Hardware of the mobile phones

Hardware of the mobile phones was classified into three categories which are bad, moderate and good. There were 92(30.7%) describing hardware as bad, 92(30.7%) moderate and 116(38.6%) describing hardware as good. These were sentiments for the five brands. The results are presented in Table 4.2.

Table 4. 2 Hardware of mobile phones

Hardware	Bad	Moderate	Good
Android	33.70%	28.70%	37.60%
Windows	19.0%	38%	43%

We further tested individual cases for hardware where;

os=1 & hardware=1, os=1 & hardware=2 and os=1 & hardware=3

os=1 & hardware =1, os=1 & hardware =2 and os=1 & hardware =3

Most of the sentiments (37.60%) described hardware as good for android, while (43%) of sentiments described hardware as good for windows. On the other hand 28% and 38% as moderate for android and windows respectively.

4.3.3 Class of mobile phones

The sentiments on class of mobile phones portrayed satisfaction and dissatisfaction with the three classifications of the data i.e. operating system, hardware and features. For classification we tested for both yes 160 (53.4%) and no 140(46.6%). These represented sentiments for all the five brands. For the operating system we were testing whether it was Windows or Android phones and features and hardware tested on whether the brand was good, moderate and bad. The Yes sentiments indicated the probability of the brand making good sales while the No indicated the probability of the brand not making sound sales. 1 and 2 indicated the ranking of an opinion. The higher the opinion the higher the value. The results are presented in Table 4.3.

Table 4. 3 Class of mobile phones

Brand	Features	Hardware	Yes	No	Yes	No
Android	Bad	Bad	1	15	6.30%	93.70%
		Moderate	1	25	3.80%	96.20%
		Good	2	2	50%	50%
	Moderate	Bad	2	64	3%	97%
		Moderate	22	14	61%	39%
		Good	58	1	98%	2%
	Good	Bad	3	1	75%	255
		Moderate	10	2	83%	17%
		Good	31	1	97%	3%
Windows	Bad	Bad	1	3	25%	75%
		Moderate	1	9	10%	90%
		Good	3	1	75%	25%
	Moderate	Bad	1	11	8%	92%
		Moderate	2	5	29%	71%
		Good	15	1	93.30%	6.3%
	Good	Bad	1	1	50%	50%
		Moderate	12	1	92%	8%
		Good	12	1	92%	8%

To avoid getting zero probabilities, 1 had to be added due to laplace smoothing since there were some cases where probabilities were zero. Majority of the sentiments were android because four brands had (98.0%) described Yes preference as moderate and good and for windows (93.3%) described No preferences as moderate and good.

4.4 Mining product features

In predicting product market adoption this study will borrow from Tuarob and Tucker (2013) who predicted a precision of 50% in their model. They looked at features of various smartphones products using Term frequency - inverse document frequency model (TFIDF) and Latent Dirichlet Allocation model (LDA) respectively, and they both models yielded a precision of 50%.

4.4.1 Sample tables with features

The features extracted from tweets related to each selected smartphone model using the TFIDF based feature extraction algorithm are shown in Table 4.4.

Table 4. 4 Features extracted from tweets

Features	iPhone 4			Samsung Galaxy S II			Motorola Droid RAZR			Sony Ericsson Xperia Play		
	Strong	Weak	Controversial	Strong	Weak	Controversial	Strong	Weak	Controversial	Strong	Weak	Controversial
1	case	phone	case	touch-screen	touch-screen	touch-screen	battery-life	touch-screen	android	game	network	game
2	camera	case	face-time	update	update	update	commercial	update	battery-life	play	play	play
3	face-time	face-time	camera	ics	screen	screen	update	screen	commercial	control	game	network
4	app	battery-life	battery-life	battery-life	video	battery-life	screen	video	update	controller	video	game
5	screen	camera	app	screen	note	sensation	ics	sensation	screen	playstation	control	control
6	life	screen	screen	sensation	sensation	ics	app	battery-life	cream	commercial	picture	commercial
7	battery-life	app	texting	support	battery-life	contract	keyboard	case	ics	freebie	style	picture
8	price	life	update	message	case	camera	message	ics	app	bootloader	data	battery-life
9	update	update	video	upgrade	ics	internet	picture	carrier	price	battery-life	emulator	playstation
10	video	video	price	camera	function	app	price	function	release	carrier	cpu	integration
Pr@50	0.22	0.22	0.22	0.2	0.22	0.16	0.28	0.14	0.24	0.28	0.14	0.2

The features extracted from tweets related to each selected Smartphone model using the LDA based feature extraction algorithm are shown in Table 4.5.

Table 4. 5 Features of smart phone mobile

Features	iPhone 4			Samsung Galaxy S II			Motorola Droid RAZR			Sony Ericsson Xperia Play		
	Strong	Weak	Controversial	Strong	Weak	Controversial	Strong	Weak	Controversial	Strong	Weak	Controversial
1	camera	battery-life	battery-life	touch-screen	touch-screen	touch-screen	battery-life	keys	picture	game	game	game
2	battery-life	face-time	face-time	update	function	email	screen	price	price	battery-life	accessories	video
3	screen	app	camera	battery-life	email	video	picture	browser	browser	control	video	commercial
4	app	video	app	screen	video	bootloader	android	bootloader	webpage	fun	battery-life	control
5	price	jailbreak	video	ics	bootloader	photo	glass	warranty	life	hardware	commercial	gaming
6	music	wifi	update	sensation	photo	texting	app	microphone	music	performance	style	battery-life
7	face-time	bug	voice-control	display	gallery	price	camera	delay	update	experience	control	baseball
8	message	charge	wifi	video	button	jelly bean	keyboard	bloatware	screen	wifi	app	hardware
9	voice-contr	location	screen	app	texting	app	network	fixes	touch-screen	video	size	experience
10	case	touch-screen	case	picture	price	network	noise	email	android	controller	carrier	controller
Pr@50	0.62	0.56	0.58	0.52	0.1	0.52	0.36	0.26	0.22	0.38	0.16	0.1

The top 10 strong/weak/controversial features extracted from the four smartphone models using the LDA based approach provides useful information that matches with the actual product specification. For example, the Apple iPhone 4 features 5MP and dual (back and front) cameras, longer battery life. However, some users still complain about the battery time while on 3G mode, harder to jailbreak, and the bug about signal occasionally drop while touching the antennae sideline. The controversial features are mostly new features missing in the predecessors. The Sony Ericsson Xperia Play features the combination of smartphone and game console. Thus, most of its strong features involve gaming. It controversial features are mostly about gaming which are newly integrated into this model.

4.5 Training dataset

In this study, the whole population was included in the training set due to the small sample size. Training is the use of data which we already know the type of preference (i.e. Yes or No) in the proposed model and establish whether the model is able to correctly predict the type of preference given the three classifications of data.

4.5.1 Positive and negative messages

The messages were drawn from online tweets and posts by users of specific phone brands. It was a requirement for this study that the messages collected had either positive or negative information concerning the selected smart phone brands. The results are presented in Table 4.6.

Table 4. 6 Examples of messages

Category	Sample
I can't achieve this clarity if I am running a movie on TECNO	Moderate
LG smartphone is bad not good for heavy games as the processor is slow	Bad
Samsung is good but battery life is so bad	Moderate
Ah SONY Experia Z has a very good big screen, music sounds very good. It is a good options I recommend it	Good

4.6 Chapter summary

This chapter has highlighted and discussed the analysis of the five brands of mobile phones. The analysis was done based on the features, operating system and hardware of the five brands. A final classification of the five brands was done to determine the sales of the mobile phones in the market.

CHAPTER FIVE

MODEL DESIGN IMPLEMENTATION AND EVALUATION

5.1 Introduction

This chapter describes the bayes theorem and its applicability. It also deals with the pseudo code and the classifier implementation in the prediction.

5.2 An Overview Bayes Theorem:

The Bayes theorem is presented in this section.

$$P(A/B) = P({}^{B}/_{A}) * P(A)/P(B)$$

P (positive/message) = P (message/positive). P (positive)/P (message)

Dropping the denominator

P (good/message) =P (message/good). P (good)

The message is split to a number of features: x1-xn represents features

P(x1…..xn)

Return the class with the Maximum value.

cMAX = argmaxc.

5.3 The Training Model Algorithm

TrainBayesClassifier($TrainingSet, TargetPhoneBrands$)

$Vocabulary$ ←all distinct keywords and other phrases and tokens in$TrainingSet$

For each target phone brandb_j in $TargetPhoneBrands$ do

$docs_j$ ←subset of the $TrainingSet$ for which the Target phone brand is b_j

$$P(b_j) \leftarrow \frac{|docs_j|}{|TrainingSet|}$$

$Text_j$← a single document created by concatenating all members of $docs_j$

n← total number of keywords in $Text_j$ counting duplicate keywords multiple times.

For each keyword w_k in vocabulary

n_k ← number of times keyword occurs in $Text_j$

$$P(w_k|b_j) \leftarrow \frac{n_{k+1}}{n + |Vocabulary|}$$

5.4 The Classifier Algorithm

ClassifySentiment ($SentimentDocument$)

$Positions$←all keyword positions in *SentimentDocument* that contains tokens found in *Vocabulary*

Return a list L of keywordssorted in descending order such that

$$L_0 \leftarrow b_{NB} = \underset{b_j \in TargetP\ honeBrands}{\operatorname{argmax}} P(b_j) \prod_{i \in positions} P(a_i|b_j)$$

Explanation of the Training Algorithm

The sentiment document can either be positive or negative. We represent each document by vector of keywords, one attribute per keyword position in the document. In training the classifier, we use training examples to estimate $P(+)$, $P(-)$, $P(doc|+)$and $P(doc|-)$. We utilize the Naïve Bayes conditional independence assumption such that

$$P(doc|b_j) = \prod_{i=1}^{lengt\ h(doc)} P(a_i = w_k|b_j)$$

Where $P(a_i = w_k|b_j)$is the probability that the keyword in position i is w_k given target phone brandb_j. We also assume that $P(a_i = w_k \backslash b_j) = P(a_m = w_k), \forall i, m$

In order to train the classifier, we loaded disambiguated training set (we clean our data using the Alchemy API NLPTK) and target phone brands. We then collected all the keywords and tokens that occur in the training set followed by iterative calculation of the required $P(b_j)$ and$P(w_k)\backslash b_j$ probability terms.

Explanation of the Classifier

Classifying a sentiment involves determining the keyword positions in the sentiment that contain tokens found in the dictionary. To achieve the comparative advantage, we return a list of the keywords sorted in descending order.

5.5 System architecture

The system architecture consist of presentation layer, workflow controller, classifier, classifying subsystem and Alchemy API NLTK. Figure 5.1 presents the system architecture.

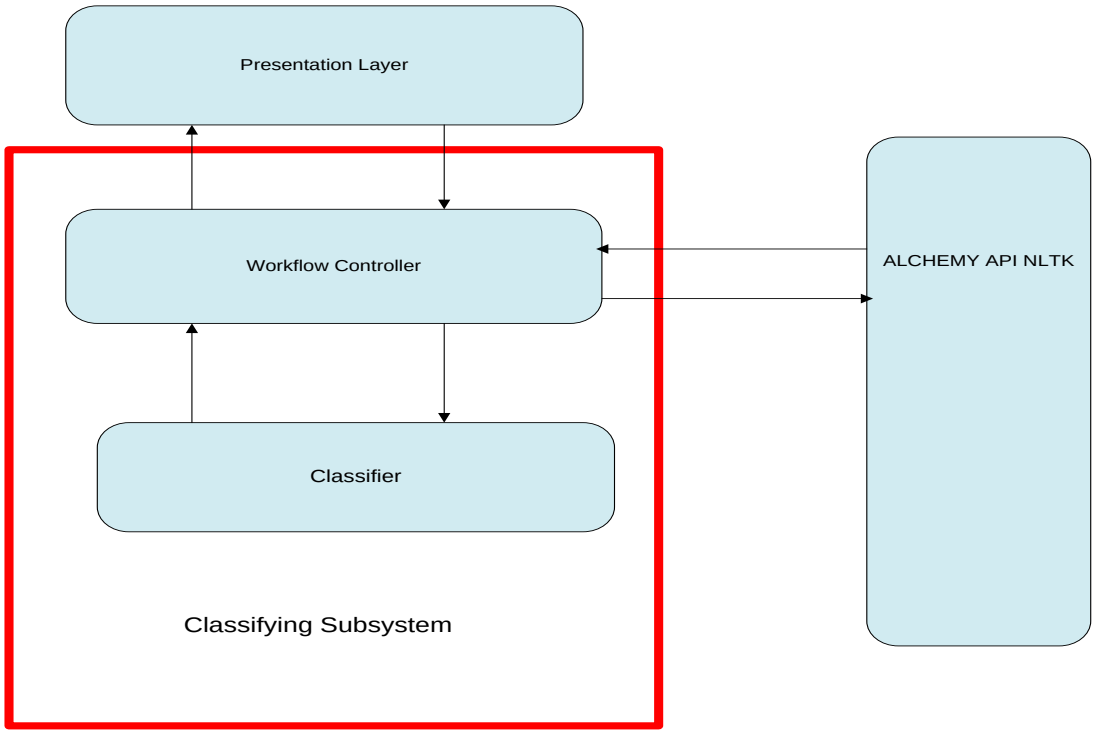

Figure 5. 1 System architecture

The Presentation layer captures sentiment and makes a classification request to the classifying subsystem. It also presents the classification results. The classifying subsystem uses Naïve Bayes principles to discover keywords, phones brands and features with associated probabilities in the sentiment document. It is also able to determine the polarity of the sentiment document. Alchemy API Natural processing ToolKit helps disambiguate the sentiments and detect the language used. The workflow controller makes disambiguation requests to the Alchemy API and passes clean data to the classifier then classifier passes the results back to the presentation layer.

5.5.1 User interface

This is the user interface where the user enters the sentiments. Figure 5.2 shows the interface.

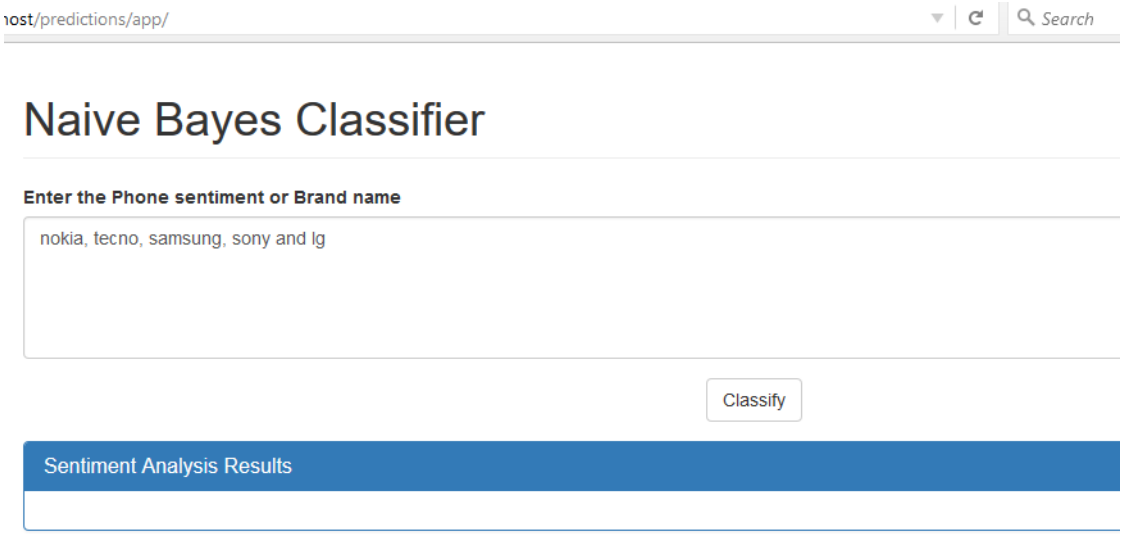

Figure 5. 2 User Interface

The user interface is a web based interface developed using php. In this figure, when a user key in the sentiments, the interface classifies the brand.

5.5.2 Nokia prediction

The results outline the comparative analysis of the prior probability for Nokia brand against the posterior probability realized after subjecting the sentiments to the predictive model. The Nokia brand had a 'moderate' prediction of 0.924077. The implication is that the probability of the Nokia brand selling in the market is average based on its class performance. The prediction is shown in Figure 5.3.

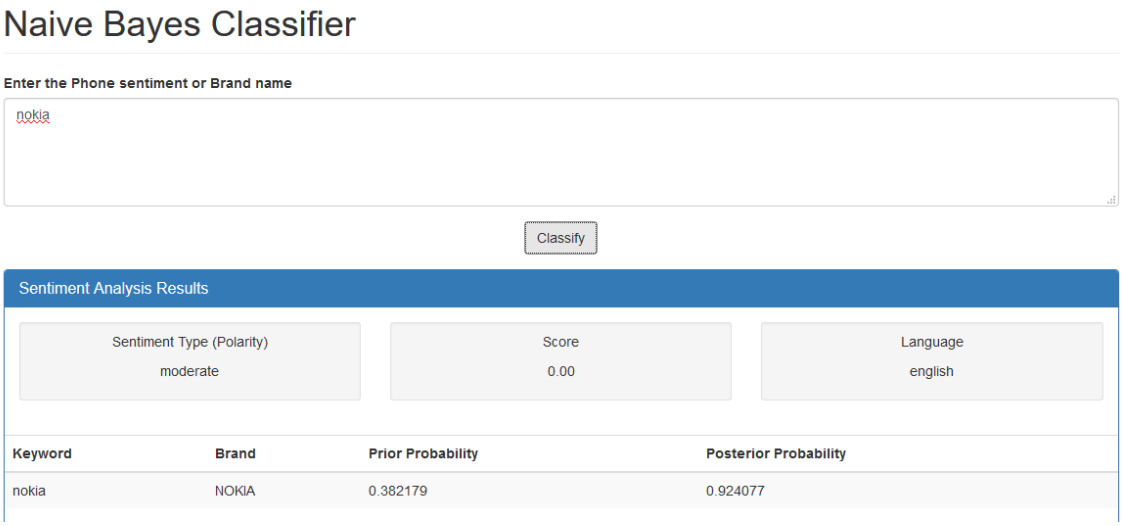

Figure 5. 3 Nokia predictions

5.5.3 Samsung prediction

Comparative analysis of the prior probability for Samsung brand against its posterior probability realized a 'good' prediction after being subjecting to the predictive model. The brand had a 'good' prediction of 0.969435. The implication is that the probability of the Samsung brand selling in the market is good based on its class performance.

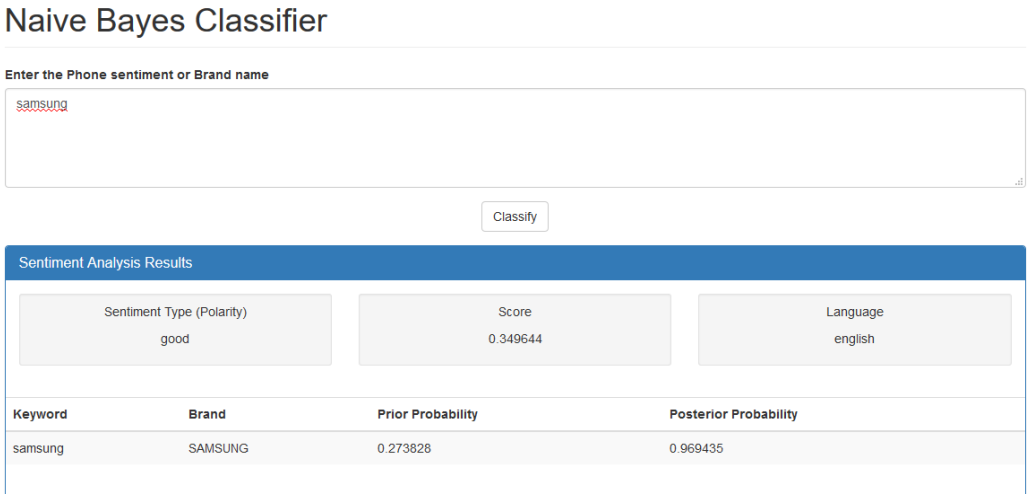

Figure 5. 4 Samsung predictions

5.5.4 Sony prediction

The results indicate the comparative analysis of the prior probability for Sony brand against the posterior probability realized after subjecting the sentiments to the

predictive model. Sony brand had a ‘moderate’ prediction of 0.973398. The implication is that the probability of the Sony brand selling in the market is average based on its class performance.

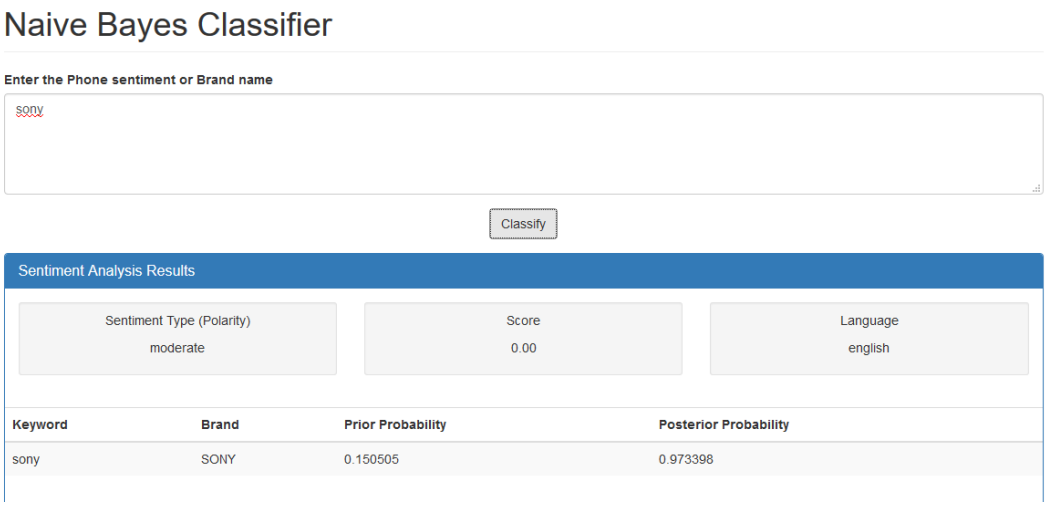

Figure 5. 5 Sony predictions

5.5.5 Tecno prediction

A comparative analysis of the prior probability for Tecno brand against its posterior probability after being subjected to the predictive model was tested. The Tecno brand had a ‘moderate’ prediction of 0.94901. The implication is that Techno brand has an average chance of selling in the market as indicated by its class performance.

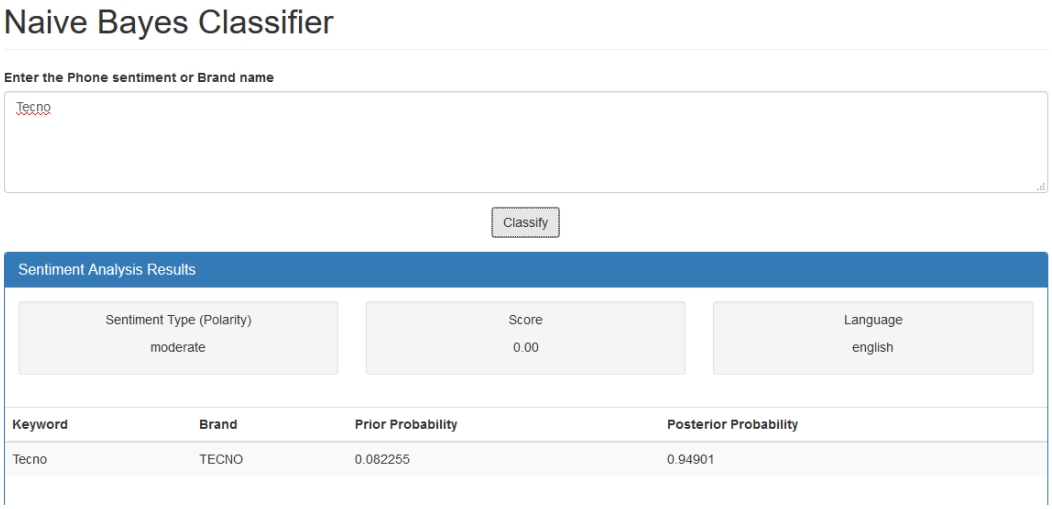

Figure 5. 6 Tecno predictions

5.5.6 LG prediction

The results outline the comparative analysis of the prior probability for LG brand against the posterior probability realized after subjecting the sentiments to the predictive model. The LG brand had a 'good' prediction of 0.988024. The implication is that the probability of the LG brand performing well in the market is good based on its class performance.

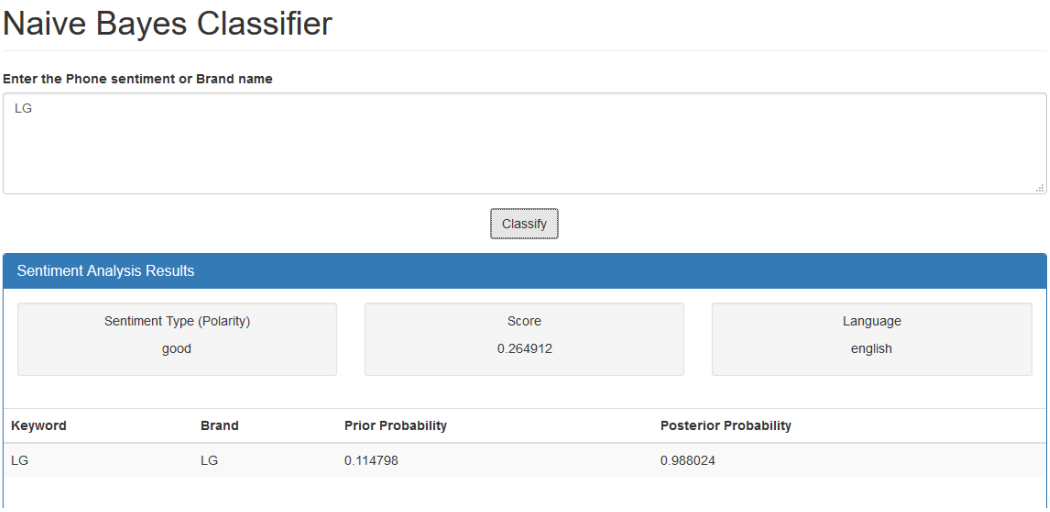

Figure 5. 7 LG predictions

5.6 Capturing notable aspects of the brand

The predictive system could further highlight other notable aspects of analysis of a brand. This enables the user to move from the level of interpretation of the brand to a higher level of searching for extra information. E.g. when a user enters the word Xperia the system is able to Identify Xperia as a Sony Brand

5.6.1.1 Lumia prediction

The Figure 5.8 shows that, in actual comparison of Nokia and Nokia, the prior probability varies for the two models. On searching further into other notable features of the brand, the posterior probability for Nokia significantly registering 0.99246 and Nokia was 0.924077. The variation is due to the fact that the predictive model is capable of highlighting the notable brand specifications. The classification

for Nokia is ‘good’ with a score of 0.888659. This implies that the other aspects predicted correctly enable the user/buyer to identify Nokia as an ideal brand.

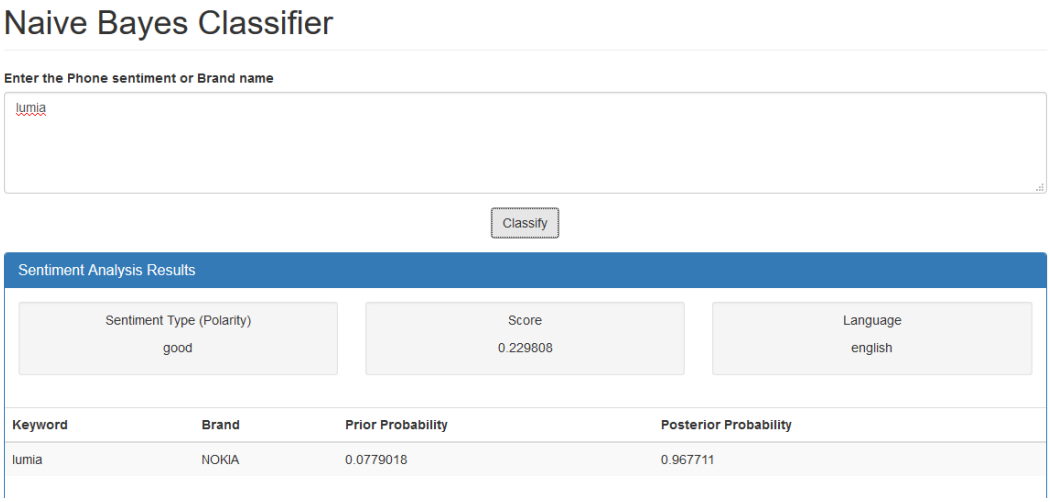

Figure 5. 8 Lumia prediction

5.6.1.2 Galaxy prediction

The Figure 5.9 shows that, in actual comparison of Samsung and Samsung without. However, on searching further into the other notable aspects of the brand, the posterior probability significantly variance with Samsung registering 0.922311 and Samsung had a higher score of 0.969435. The variation is due to the fact that the predictive model is capable of highlighting the notable brand specifications. The sentiment polarity for Samsung is ‘good’. This implies that the notable aspects correctly enable the user/buyer to identify Samsung as a good brand.

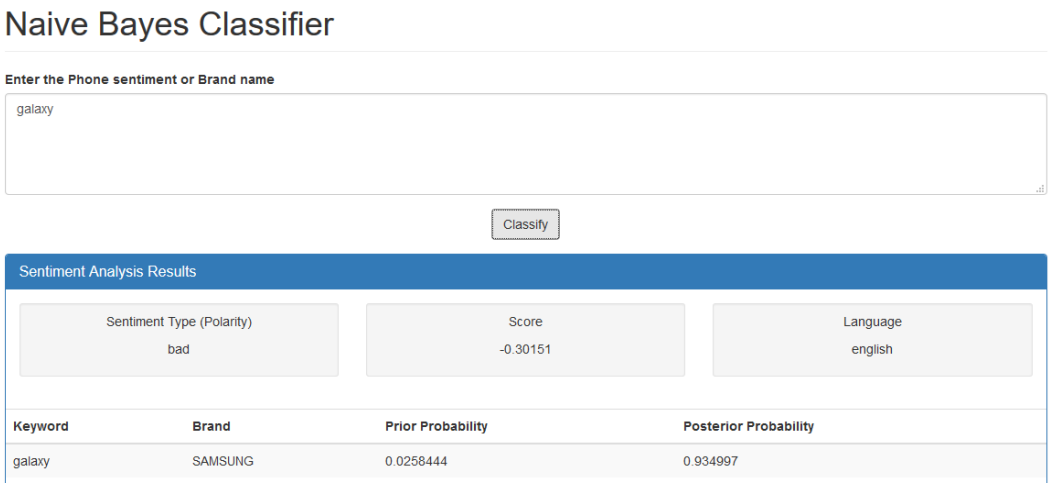

Figure 5. 9 Galaxy predictions

5.6.1.3 Xperia prediction

In the actual comparison of Sony and Sony with other notable aspects, the prior probability and posterior probability are not similar for the two models. The posterior probability varies since the predictive model is capable of highlighting the notable brand specifications. The classification for Sony was 0.956942. This implies that the other notable aspects had a higher degree of accuracy enabling the user to identify Sony as a good brand.

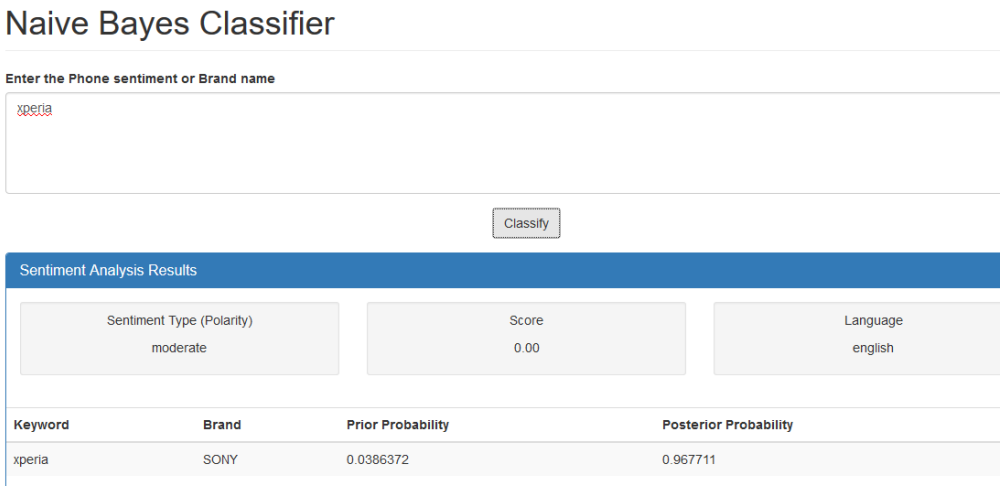

Figure 5. 10 Xperia prediction

5.6.1.4 Nokia battery prediction

A comparative analysis of the prior probability for Nokia battery against its posterior probability indicates a ‘moderate’ prediction after being subjecting to the predictive model. The brand had a ‘good’ prediction of 0.953477. The implication is that the Nokia brand has an average chance probability of selling in the market based on its class performance.

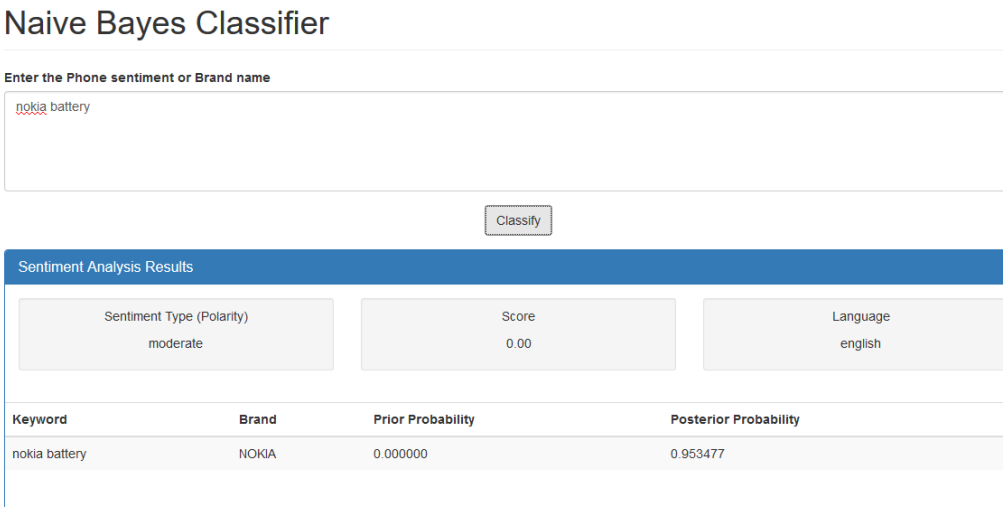

Figure 5. 11 Nokia battery prediction

5.6.1.5 Tecno battery prediction

A comparative analysis of the prior probability for Tecno battery against its posterior probability indicates a 'moderate' prediction after being subjecting to the predictive model. The brand had a posterior prediction of 0.962678. The implication is that the Tecno brand has a significant average chance probability of selling in the market based on its class performance.

Figure 5. 12 Tecno battery prediction

5.6.1.6 Samsung battery prediction

A comparative analysis of the prior probability for Samsung battery against its posterior probability indicates a 'good' prediction after being subjecting to the predictive model. The brand had a posterior prediction of 0.979604. The implication is that the Tecno brand has a significant good chance probability of selling in the market based on its class performance.

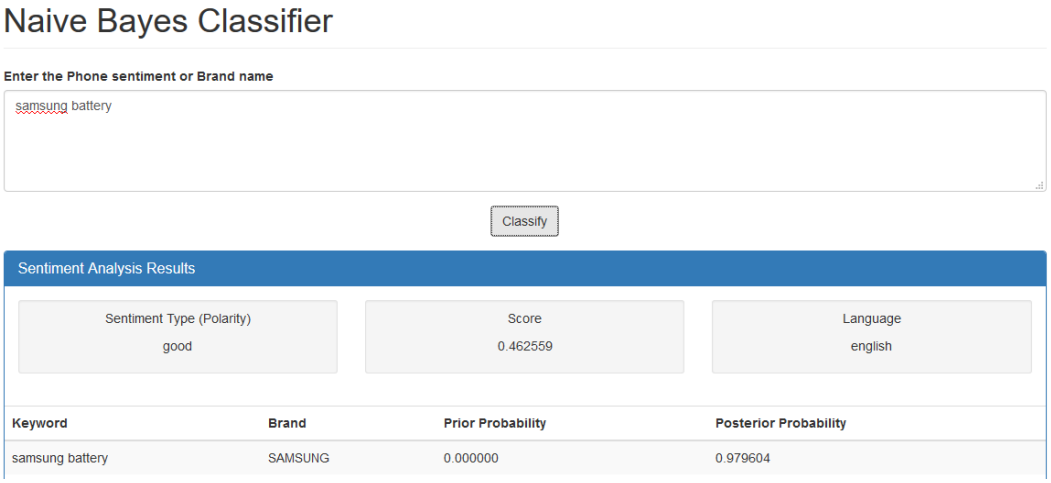

Figure 5. 13 Samsung battery prediction

5.6.1.7 Sony battery prediction

A comparative analysis of the prior probability for Sony battery against its posterior probability indicates a 'good' prediction after being subjecting to the predictive model. The brand had a posterior prediction of 0.967155. The implication is that the Sony brand has a significant good chance probability of selling in the market based on its class performance.

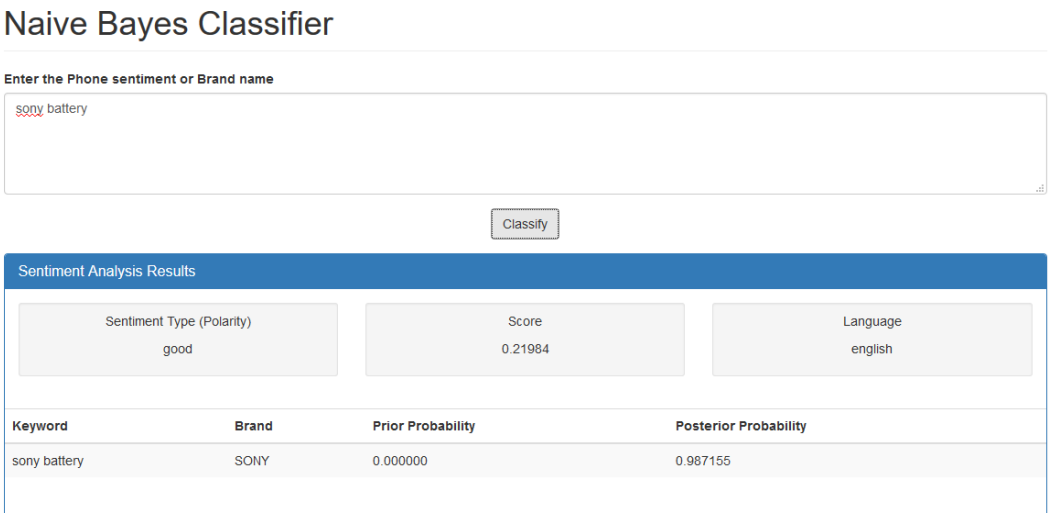

Figure 6. 1 Sony battery prediction

5.6.1.8 LG battery prediction

A comparative analysis of the prior probability for LG battery against its posterior probability indicates a 'good' prediction after being subjecting to the predictive model. The brand had a posterior prediction of 0.943396. The implication is that the LG brand has a significant good chance probability of selling in the market based on its class performance.

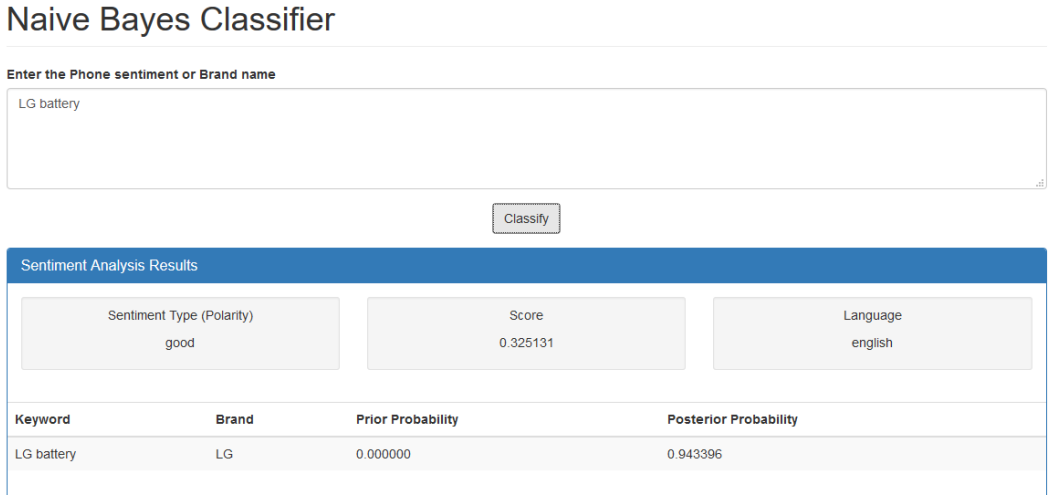

Figure 5. 14 LG battery prediction

5.7 Performance of mobile phone brands in the market

Evaluation of the sentiments for each of the five brands of mobile phones went further into assessing the performance of the brands in the market. The results indicate that on average, all the five brands registered a sentiment of 'good' performance in the market. Tecno had the highest performance measure in the market of 0.999824 followed by LG (0.903687), Nokia (0.887728), Samsung (0.880043) and finally Sony with a score of 0.753349. The implication is that based on the prediction of users sentiments, Tecno brand is generally performing considerably well in the market, and on the other hand, a sizeable population feels that the performance of Sony is bad, either in terms of features, hardware or class. The representation of performance is shown in Figure 5.5.

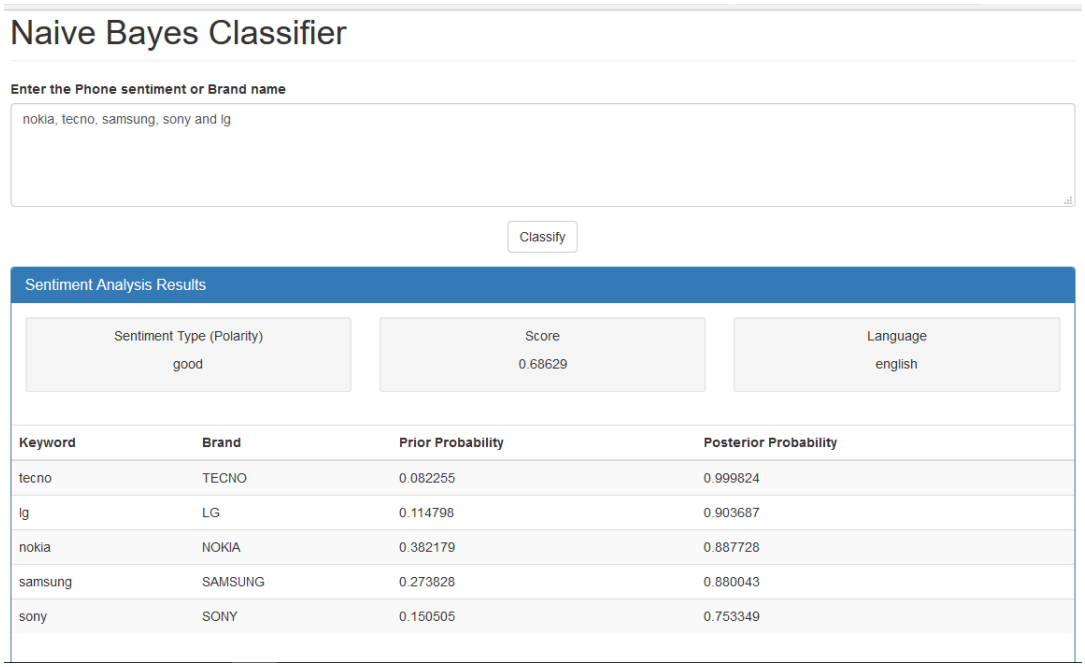

Figure 5. 15 Market performance of mobile brands

5.7.1 Pie chart representation of brands performance

A pie chart presentation of the posterior probability of mobile performance is shown in this section.

The results of a pie chart representation of the average performance of all the five brands registered a significantly even performance in the market. Though Tecno had the highest performance measure in the market and Sony with the lowest score, the

user's sentiments indicate a tight competition for all the five brands. The implication is that based on the prediction of user's sentiments, the features, hardware or OS determines the market performance of each brand. The representation is indicated in Figure 5.16.

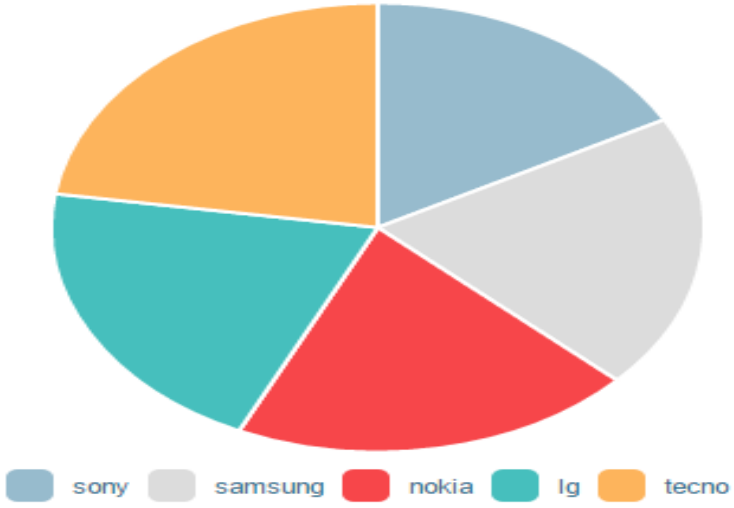

Figure 5. 16 Performance of market brands

5.8 Evaluation of the model using Receiver Operating Characteristic curve

Bayesian lab software was further used to evaluate our predictive model. The ROC (Receiver Operating Characteristic) curve is used to measure the performance of the model used in prediction of the class of preference given the three categories of our data. The area gives sensitivity (true positive rate) versus specificity (false positive rate). It measures the ability of the used model to correctly classify the sentiments into the correct category of preference (either yes or no). The curve shows the trade-off between sensitivity and specificity. The closer the curve is to the left (sensitivity border) and the top border the more accurate the model performance.

Markov model recorded a total precision of 91.67% while the ROC was 97.86%. The model predicted 161 of the sentiments belong to preference class No with precision of 98.57%. This means that the model correctly predicted the sentiments to be in class No with 98.57% accuracy. Also, the model predicted 139 sentiments to belong to preference class Yes with precision of 85.62%. From the plot above we also find that the model predicts sentiments to be in preference class No while they actually

belong to class Yes at a rate of 14.37% and also predicts sentiments to be in preference class Yes while they actually belong to class No at a rate of 1.43% which is acceptable. Comparison of naïve bayes is better because it has a high precision of 93.33% while Markov had a precision of 91.67% as seen in Figure 5.17.

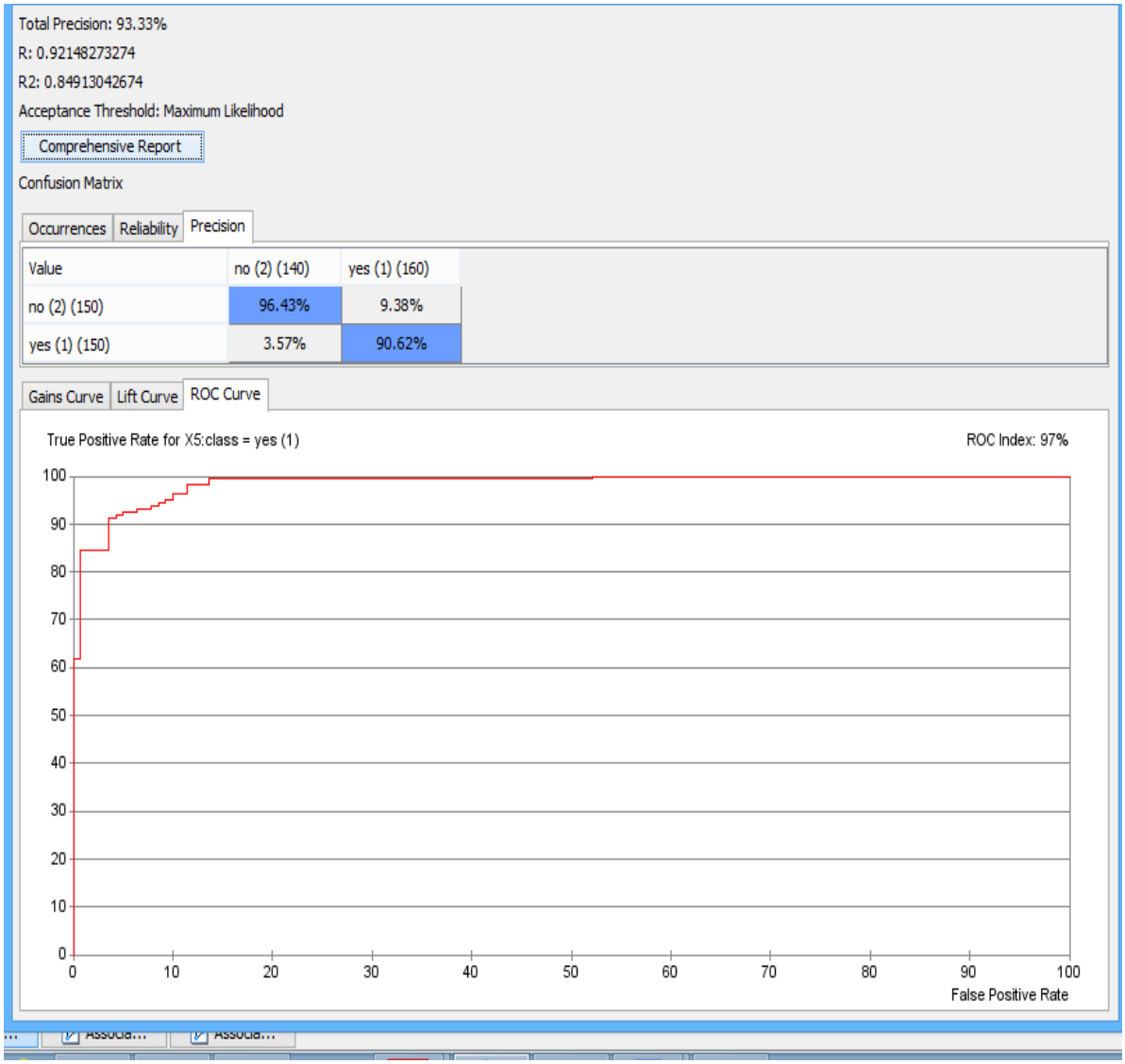

Figure 5. 17 Bayesian Receiver Operating Characteristics Curve

5.8.1 Bayesian Receiver Operating Characteristics Curve

Markov model recorded a total precision of 91.67% while the ROC was 97.86%. The model predicted 161 of the sentiments belong to preference class No with precision of 98.57%. This means that the model correctly predicted the sentiments to be in

class No with 98.57% accuracy. Also, the model predicted 139 sentiments to belong to preference class Yes with precision of 85.62%. From the plot above we also find that the model predicts sentiments to be in preference class No while they actually belong to class Yes at a rate of 14.37% and also predicts sentiments to be in preference class Yes while they actually belong to class No at a rate of 1.43% which is acceptable. When we compare the two models naïve bayes is better because it has a high precision of 93.33% while Markov had a precision of 91.67%. The prediction is presented in Figure 5.18.

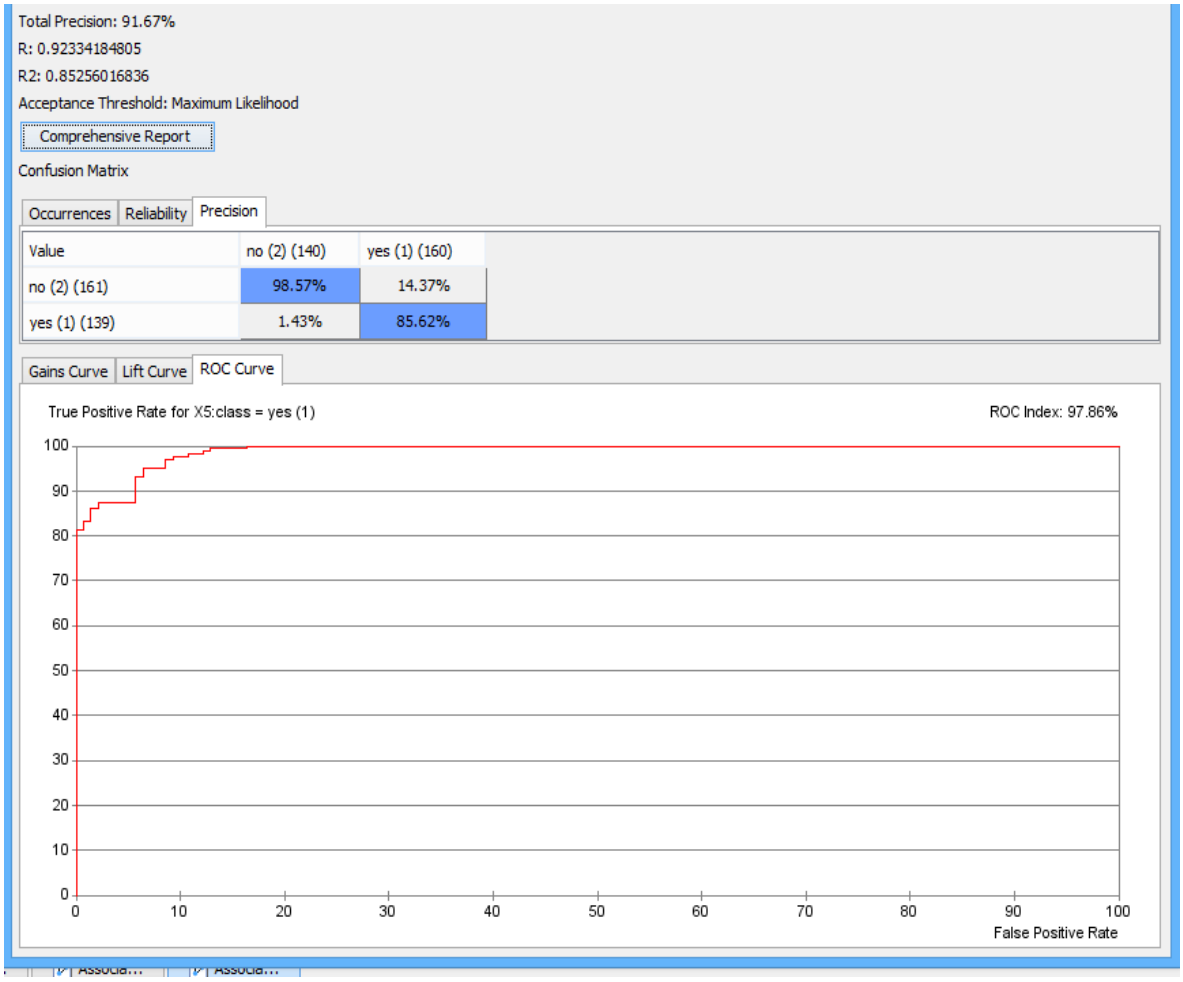

Figure 5. 18 Markov Receiver Operating Characteristics Curve

5.9 Summary

This chapter has presented the prediction model and demonstrated its practical application through the comparative analysis of the prior probability against the

posterior probability of the five mobile phone brands. The semantic aspect of the predictive model on the mobile phone brands has been explained as well as the application of Bayesian lab software to generate the ROC curve and Markov curves.

CHAPTER SIX

CONCLUSION AND FUTURE RESEARCH

6.1 Introduction

This chapter is aimed at discussing and summarizing the main findings from the study, drawing relevant conclusions and where necessary making some vital recommendations.

6.2 Conclusion of the study

This study was successful in building an application that has the ability to fetch and store data searched on various products and services from two most popular social media sites (Facebook & Twitter). A prototype using Naïve Bayes predictive model was developed that integrates an information gain heuristic using the Natural Language Tool Kit and trained it on dataset from Social Media. The results obtained from experiments with the model indicate that it is capable of performing classification with an accuracy of 93.33% for sentiments obtained from Social Media. This is near human accuracy, as apparently people agree on sentiment only around 80% of the time. Most of the sentiments in this data are expressed partly in informal language.

It can therefore be concluded that the prototype model of classification selected is ideal for the kind of data collected from social media on e-commerce. Finally, integration of the techniques and methods developed into a web based application for use in providing Sentiment Analysis with respect to products e-commerce. The application provides user friendly interface that can assist mobile phone companies in determining the competitiveness of their various brands in the market.

6.3 Future work

The following are some of the work that can be done in the future:

1. Further research can also focus on other prediction techniques such as Decision Trees, Structural Equation Models, Neural Networks and Analytical Hierarchical Processes in analyzing data from customer satisfaction surveys in E-Commerce.

REFERENCES

Abbasi, A., Chen, H., and Salem A. (2007). *Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums.* Retrived from :http://ai.arizona.edu/intranet/papers/ahmedabbasi_sentimenttois.pdf.

Allison, P.D. (2012). *Handling missing Data by Maximum Likelihood*. USA: Statistical Horizons Haverford, P. A.

Asur, S. and Huberman, B. (2010). Predicting the future with social media, in Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI IAT).

Au, W.H., Chan K. C. C. and Yao X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. *Evolutionary Computation,. 7(6), 532–545.*

Antweiler, W, and Frank M. Z. (2004). Is all that talk just noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance. 59(3), 1259-1294.*

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

Ben-Gal I. (2007). *Bayesian Networks*, in Ruggeri F., Faltin F. &Kenett R., Encyclopedia of Statistics in Quality & Reliability,New York: Wiley & Sons

Charniak, E. (1991). Bayesian networks without tears, *AI Magazine, 12 (4), 50-63.*

Choudhury, T. Vashisht, V. Kumar, V. And Srivastava, H. (2013).Data Mining Using Decistion Tree for Sales Force Optimization

International journal of advanced Research in Computer Science and Software Engineering (IJARCSSE) *3 (4).*

Choi, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for sub sentential sentiment analysis. In Proceedings of EMNLP '08, 793–801.

Chen K-Y. Leslie R. and Bernardo A. Huberman. (2003). *Predicting the Future*. Information Systems Frontiers, 5(1),47–61.

Das S.R. and Chen M.Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science, 53(9),1375–1388.

Dewan, S., & Ramprasad, J. (2009). Chicken and egg? Interplay between music blog buzz and album sales. *PACIS 2009 proceedings,* 87.

Dhar, V. and Chang, E. (2007). Does chatter matter? *The impact of user-generated content on music sales.* NYU Working Paper No. CEDER-07-06.

Dhar, V. and Chang, E.A. (2009). Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing;. 23(4), 300-307.*

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, 0894439313493979.

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

Gaurangi, P., Varsha G.,Vedant, K. and Kalpana, D. (2014).Sentiment Analysis

Using Support Vector Machine. *International Journal of Research in computer and communication Engineering.(2)*

Gruhl, D. Guha, R. Kumar, R. Novak, J. and Tomkins, A. (2005). The predictive power of online chatter, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 78-77.

Gustavsson, T. (2006). Troubleshooting using Cost Effective Algorithms and Bayesian Networks. Masters' Degree Project Stockholm, Sweden, XR-EE-RT 2007:002.

Huang**,** W**.** Nakamori, Y. and Wang**,** S. Y.**(***2005***).**Forecasting stock market **...**vector machine," Computers & Operations Research**,** 32, 2513-***2522.***

Hann, I. H. Oh J. and James, G. (2011). Forecasting the Sales of Music Albums: A Functional Data Analysis of Demand and Supply Side P2P Data. Working paper.

International Symposium on Distributed Computing and Artificial Intelligence (2010). Advances in Intelligent and Soft-Computing. Springer, 79, 613-620.

Jungherr, A. Jürgens, P. and Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., and Welpe, I. M. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment, *Journal of Social Science Computer Review 30(2), 229-234.*

Leskovec, J. Lada A. Adamic and Bernardo A. Huberman (2006). The dynamics of viral marketing. In Proceedings of the 7th ACM Conference on Electronic Commerce.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*, 627-666.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, *60*(11), 2169-2188.

Joshi, M. Das, D. Gimpel, K. and Noah A. S. (2010). Movie Reviews and Revenues: An Experiment in Text Regression NAACL-HLT.

Jayech K. and Mohamed A. M. (2010). New approach using Bayesian Network to improve content based image classification systems. *IJCSI International Journal of Computer Science Issues, 7 (6),122-134.*

Kharya S. (2012). Using Data Mining Techniques for Diagnosis of Cancer Disease. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT) 2(2).*

Lampos, V., and Cristianini, N. (2010). Tracking the flu pandemic by monitoring the Social Web. In the Proceedings of the 2nd IAPR Workshop on Cognitive Information Processing, pp. 411-416, IEEE Press.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., & Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, *19*(1), 90-97.

Liu, B. (2010). *Sentiment Analysis: A multi-faceted problem.* Retrieved from http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf.

Loughlin, C., & Harnisch, E. (2014). The viability of StockTwits and Google Trends to predict the stock market.

Madsen, A. (2010). *Bayesian Networks for Disease Diagnosis*. Computer Science

Maury, M. D., & Kleiner, D. S. (2002). E-commerce, ethical commerce?. *Journal of Business Ethics*, *36*(1-2), 21-31.

Mejova, Y. (2009). *Sentiment Analysis – an overview.* Retrieved from http://www.divms.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf.

Mishne, G. and Glance, N. (2006). Predicting movie sales from blogger sentiment, In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.

Mary,M.D. and Deborah K.S. (2002). *Journal of Business Ethics Kluwer Academic Publishers. 36,21–31,*

Myers, M. D. (1997). Qualitative Research in Information Systems. *MIS Quarterly 21(2),* 241-242.

Moody J. and Saffell M. (2001). Learning to Trade via direct Reinforcement, IEEE Trans Neural Networks, 12 (4) ,123-126.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Retrieved from *arXiv preprint arXiv:1103.2903*.

Nakagawa, T. Inui, K. and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In Proceedings of HLT (10), 786–794.

Neck, C. P., & Manz, C. C. (2010). *Mastering self-leadership: Empowering yourself for personal excellence*. Pearson.

Pak, A. and Paroubek, P., 2009. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, Cedex, France: Universit´e de Paris-Sud.

Pang, B., Lee, L. and Vaithyanathan, S, (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Pp.79–86

Phua, Y. L. (2013). *Social media sentiment analysis and topic detection for Singapore English*. *Singapore: N*aval postgraduate school monterey ca.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco, CA: Morgan Kaufmann Publishers Inc.

Pennock D.M, Lawrence S. Giles C.L and Nielsen F.A (2001). *The real power of artificial markets.* Science, 291(5506):987–988 New Delhi: Pearson Education (2010), Inc. Publishing as Prentice Hall.

Rada, M., Hugo, L. and Lieberman, H. (2006). *NLP (Natural Language Processing) for NLP (Natural Language Programming).* Computer Science Department, University of North Texas rada@cs.unt.edu 2 Media Arts and Sciences, Massachusetts Institute of Technology. Retrieved {hugo, henry}@media.mit.edu

Sharda, R. and Delen, D. (2006). *Predicting box-office success of motion pictures with neural networks.* Expert Systems with Applications, 30, pp 243–254.

Sattikar, A. A. and Kulkarni, R. V. (2012). A Role of Artificial intelligence in Security and Privacy Issues in Social Networking: *International Journal of Computer Science and Engineering Technologies, 2 (1), 792- 79.*

Socialbakers. (2012). *Kenya Facebook Statistics.* Retrieved from http://www.socialbakers.com/Facebook-statistics/kenya#chart-intervals.

Tumarkin R. Whitelaw R.F. (2001). News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal. 57(3), 41-51.*

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology 62*(2), 406-418.

Silverman, D. (2005). Doing Qualitative Research: A Practical Handbook. Oliveryard, London: Sage Publications Ltd.

Welpe, I. M., & Sprenger, T. O. (2010). Tweets and Trades: The Information Content of Stock Microblogs. Chicago: *SSRN eLibrary*.

Whitman, M. E., & Woszczynski, A. B. (Eds.). (2004). *The handbook of information systems research*. Igi Global.

Stock, W. G., Peters, I., & Weller, K. (2010). Social semantic corporate digital libraries: Joining knowledge representation and knowledge management. *Advances in Librarianship*, *32*, 137-158.

Wanyama, E. N. (2012). Social media sentiment analysis for local Kenyan products and services. *International Journal of Computer Science and Engineering Technologies, 3 (1), 62- 69.*

Wilson, T. Wiebe, J. and Hoffmann, P. (2008). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of HLT ’05, 347–354.

Yessenov, K. and Misailovic, S. (2009). *Sentiment Analysis of Movie Review Comments.*Retrieve from :http://people.csail.mit.edu/kuat/courses/6.863/report.pdf.

Yulan H. and Deyu Z.(2010).Self-training from labeled features for sentiment analysis.Knowledge Media Institute, Open University, Walton Hall, Milton Keynes MK6 6AA, UK. Walton: University of Walton Hall.

Zhang, X. Fuehres, H. and Gloor, P.A. (2011). *Predicting stock market indicators through Twitter*. I hope it is not as bad as I fear, Procedia - Social and Behavioral.