MSc in Information Technology

Analyze Quality of Products in E-commerce Systems with Sentimental Analysis

Prepared by

P. M. A. U. Bandara

Index No : 158751L

Supervised by

Mr. SamindaPremaratne

Faculty of Information Technology

University of Moratuwa

March 2019

MSc in Information Technology

Analyze Quality of Products in E-commerce Systems with Sentimental Analysis

Prepared by

P. M. A. U. Bandara

Index No: 158751L

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

March 2019

Declaration

I declare that this research is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and the list of references is given.

P. M. A. U. Bandara

(158751L)

2019/04/

I have supervised and accepted this thesis for the submission of the degree.

Mr. Saminda Premaratne

(Main Supervisor)

2019/04/

Acknowledgements

I would like to express my genuine gratitude to my advisor Mr. S.C Premaratne for the continuous support of my research, for his patience, motivation, and immense knowledge. His direction helped me in all the period of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my research. Also I thank my associates and staff in the University of Moratuwa who helped me carry out this research. Last but not the least, I would like to thank my family, my parents for supporting me spiritually throughout writing this thesis.

Abstract

E-commerce websites getting extra significant and popular today since the vast differentiated and diversified information that is presented. Studies says that more than 80% of the world population is using these websites to purchase goods and services online. For these online customers, comments / feedbacks play a major role in decision making when buying the products from the market space. Hence the diversity and the popularity of the Online space, sales of these online products get increased with time. Therefore, it is not practical to review all the given product feedback and come to conclusion on purchasing the product for a consumer. Focusing on this, this study is urges to observe the success factors of online websites and how those aspects influence on online marketing to sales growth in any organization. Therefore, in this research a Analyze Quality of Products in E-commerce System is focused on analyzing the online consumers feedbacks or comments on various products using data mining techniques such as Sentimental and filtering analysis. The outcome from the study will show feature wise relativeness in the mobile phone domain. All procedures were based on the features extracted through a thorough literature review and existing apparatuses. This will aid to calculate a "Trust Score" for the online products and a general overview to achieve a higher trust score for e-commerce organization.

Table of Contents

Declaration I

Acknowledgements II

Abstract III

Table of Contents IV

List of Figures VII

List of Tables VII

Chapter 1 – Introduction 1

1.1 Prolegomena 1

1.2 Identifying the Research Problem 2

1.2.1 Symptoms of the Problem 2

1.2.2 Justification of the problem 3

1.2.3 Defining Research Problem 3

1.3 Research Question 5

1.4 Aim 5

1.5 Objectives of the Study 5

1.6 Study Limitations 5

1.7 Chapter Outline 5

1.8 Chapter Summary 6

Chapter 2 – Literature Review on Developments and Challenges in Product Quality Analysis 7

2.1 Introduction 7

2.2 Early development and Modern Trends 7

2.3 Existing Models 8

2.4 Lexicon-Based Approach 8

2.5 Confusion Matrix 9

2.6 Summary 12

Chapter 3 – Adopted Technology in Data Mining Tools and Techniques 13

3.1 Introduction 13

3.2 Comparison of Machine Learning and Lexicon Based Method 13

3.3 What is Data Mining? 14
3.4 R Data Mining Tool 15
3.5 Summary 15
Chapter 4 - Approach to Analyze Quality of Products in a System 16
4.1 Introduction 16
4.2 Users 16
4.3 Input 16
4.4 Output 16
4.5 Process 16
4.6 Technology 17
4.7 Summary 17
Chapter 5 – Design to Analyze Quality of Products in System 18
5.1 Introduction 18
5.2 Top level Architecture of the System 18
5.3 Data Collection 19
5.4 Pre-Processing 19
5.4.1 Removal of duplicate 20
5.4.2 Removal of HTML tags , hyperlinks 20
5.4.3 Stop words removal 20
5.4.4 Tagging Speech 20
5.5 Tokenization 21
5.6 Categorization 21
5.7 Lexicon Module 21
5.7.1 Lexicon Dictionary 21
5.8 Generating Product Rating 22
5.9 Rating Module 22
5.10 Validate Model 22
5.10.1 Confusion matrix for positive reviews 23
5.10.2 Confusion matrix for negative reviews 23
5.10.3 Confusion matrix for neutral reviews 23
5.11 Summary 24

Chapter 6 – Implementation of Analyzing the Quality of the Product in a System 25
6.1 Introduction 25
6.2 R Language 25
6.3 Data Collection 26
6.4 Data set Preprocessing 26
6.5 Categorization of the Dataset 26
6.6 Implementation of the Application 26
6.6.1 Lexicon Module 26
6.6.2 Lexicon Dictionary 27
6.6.3 Scoring Algorithm 27
6.7 Summary 27
Chapter 7 - Evaluation 28
7.1 Introduction 28
7.2 Analysis 28
7.3 Model Validation with Existing Products 31
7.3.1 TheySay API 31
7.3.2 Google Cloud Natural Language 32
7.3.3 Paralleldots API 33
7.4 Overview of the Research 34
Chapter 8 – Conclusion and Recommendations 35
8.1 Introduction 35
8.2 Summary of the research 35
8.3 Conclusions 35
8.4 Limitations of the Study 35
8.5 Future Work 36
Appendixes 41
Appendix A – R code for Web Scrape from Amazon 41
Appendix B – R code for Web Scrape Parse Amazon html pages for data 42
Appendix C – R code for Review categorized by features 46
Appendix D – R code for Sentimental Analysis using Lexicon Dictionary 48

List of Figures

Figure 1.1.1: Retail e-commerce sales worldwide from 2014 to 2021 1
Figure 2.4.1: Example of Native words 9
Figure 3.3.1: Steps of data mining Process 14
Figure 5.2.1: Top level architecture of the system 19
Figure 7.2.1: Positive Accuracy Comparison for Lexicon 30
Figure 7.2.2: Negative Accuracy Comparison for Lexicon 30
Figure 7.3.1: Result of theysay API 31
Figure 7.3.2: Google Cloud Natural Language Results 32
Figure 7.3.3: Google Cloud Natural Language Results 33
Figure 7.3.4: Paralleldots Results 33

List of Tables

Table 5.11.1: Confusion Matrix for Positive Reviews 23
Table 5.11.2: Confusion Matrix for Negative Reviews 23
Table 5.11.3: Confusion Matrix for Neutral Reviews 23
Table 7.2.1: Lexicon Performance 28
Table 7.2.2: Tweaked for Mobile Phone Domain 29
Table 7.2.3: Lexicon VS Tweaked Lexicon 29
Table 7.3.1: Comparison with TheySay API 31
Table 7.3.2: Comparison with Google Cloud Natural Language 33
Table 7.3.3: Comparison with Paralleldots API 34

Chapter 1 – Introduction

1.1 Prolegomena

The rapid development of Information and Communication Technology has changed customer purchase behavioural pattern from bricks and mortar to e-commerce purchases owing to its vast range of options, low cost due to higher discounts and convenience.

In the year 2017 e-commerce retail sales achieved $2.3 trillion by making an increase of 25% compared with the previous year and $4.88 trillion growth is projected to the year 2021 [1].

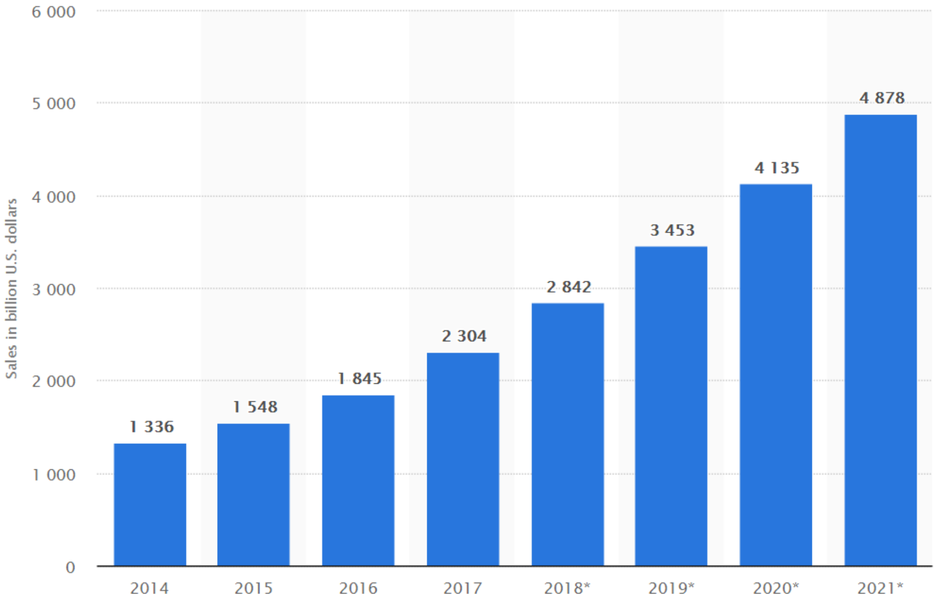


Figure 1.1.1: e-commerce sales growth year by year from 2014 to the projected year 2021

At present electronic word of mouth (e-WOM) is one of the most important factors for e-commerce websites due to the current explosion of online tools and social networking sites [2]. As per e-WOM by Ahrens, Coyle, and Strahilevitz, 70% of customers have checked the online websites and social media for information and reviews about the product and the company (seller) prior making the purchase online [3].

In order to facilitate customer decision making regarding the online purchase, e-commerce websites have allowed users to comment and to provide a star rating for products and for sellers.

Customers can provide their negative, positive or neutral feedback towards each transaction in an unbiased manner by expressing their true feelings about the product quality, price, delivery speed and many other related factors which can have an impact on a potential buyer's purchasing decision.

1.2 Identifying the Research Problem

1.2.1 Symptoms of the Problem

Online customers are exhausted by overloaded product review information due to a high number of user comments on e-commerce websites making it difficult to make a purchasing decision. Customers are required to decide what subset of available information to use. The solution for this can be highlighted as the star rating system, since it provides a quick indication of the tone of the overall review. However, the star rating scheme is not detailed enough to have a clear understanding of the product or the seller.

Expansion and the fast growth of the web allowed people to put their views on diverse websites. This has become a booming to a digital age which enabled the old styled WOM to transfer into a varied communication. Hence this enabled to discover different sentiments of those people who have put their opinion and thoughts. Therefore, lots of people use these online platforms to share their thoughts and well as receive other people thoughts in the same regards. This makes a large number of opinions and thoughts of data collected via various types of differentiated mediums such as blogs, forums, reviews in social media etc [4].

Moreover, numerous companies which uses online marketing has connected social media to their platforms as a strategic move so that it will encourage online users to give their thoughts about the products that the organization sell or the services that they provide via online [5]. Hence this makes many organizations work towards understanding these reviews to adjust and for use for the betterment of the organization product improvement.

Hence, in overall a systematic method should be used to automatically perceive the complexity, scope, deepness and extensiveness of users' reviews and to create more inclusive solo rating by combining the above useful reviews into one to display to the customer at a glance.

1.2.2 Justification of the problem

Many online users make their decisions based on online reviews that are stated in an online platform. This not only makes users to make decisions but act as the base platform to rethink about their selections which impact to customers buying behavior and buying patterns. This will eventually make users to help in their decision making to buying quality products as well as buying various diverse products [6].

According to these facts along with various studies,

- It is said that 92% of online users read products reviews in an online product catalog when buying the product beforehand and 40% of users make a final decision by just reading the 1- 3 reviews of the product [7].
- One of the studies that was done by Nielsen expresses that 66% of online consumers believes in reviews and rating which are displayed against in online products [8].
- Another research study reveals that one to three bad reviews to a product is sufficient to discourage 67% of online customers [9].
- The study reveals that 63% of online users are more encouraged to buy online products which have user reviews [10].
- One of the studies which was conducted on US internet users discloses that the online users are more biased towards customer reviews rather that the product descriptions given by the product manufacturer [11].
- Another survey discloses that customer pays more than 20% for a 5 Star rated product than a lesser rated (4 star) online products [12].

Above stated information expresses the concentration that customer puts on the online product reviews before making a final decision on recommending purchase the online goods and services. Hence the sentiment analysis is vital for these kinds of decision making and helpful for the customers to make better choices in terms of acknowledging the negative and positive traits of a product overall and feature wise.

1.2.3 Defining Research Problem

The expansion of the World Wide Web contains a load of prejudiced information stored in a digitalized way. By examining this large info, any organization can make great opportunity by recognizing the sentiments of the community. But large set of data means it is hard to extract helpful information. One of the surveys shows that applications such as Twitter generates 21 million tweets per hour and Facebook creates 20 million posts per hour [4]. Hence it is revealed that most of the online users frustrate while roaming in a shopping space which hard to come to a purchase decision by going through the comments.

Furthermore, data flows in millions everyday via the web which makes organizations also in trouble in dealing with the large scale of opinionated information. One of the surveys illustrated that nearly 700 companies from 2100 companies do not have an idea of what most of their respected customers thinks about their products and services and 31% of them having problems in measuring the opinions of the customers [13]. It is understandable that they do not lack of any valuable customer insight data but having problems in extracting the valuable opinioned data out of large scale complicated data sources.

This leads to an exploration into the research field of sentiment analysis in unstructured data. Semantic orientation (SO) describes the degree of opinions and subjectivity which states the positivity, negativity or neutrality of the opinion and strength of words, phrases, sentences or documents [14].

At present one of the sentiment analysis approaches is a machine learning technique with data mining. This enables to select correct attributes from texts and algorithms from characterized text examples of building text classifiers. In many large tools of sentiment analyzing been developed to mine to user generated opinions in a web platform.

So that the problem address here is how to automate large scale of opinioned data or user reviews with public sentiments so that the outcome rating can accommodate people in decision making in purchasing online products. Furthermore, this can be also helpful for product manufacturers to improve their goods and services based on the given outcome from this sentiment analysis.

1.3 Research Question

How can the online product reviews can be effectively and efficiently categorized with according to the given feedbacks?

1.4 Aim

The aim of this project is to analyze and design an accurate way to develop sentiments by refining the sentiment classification presentation and mining characteristics related to sentiments and evaluate the model in e-commerce systems to make better, effective and efficient user experience when purchasing online mobile phones.

1.5 Objectives of the Study

1. Identify customer satisfaction on existing e-commerce customer feedback monitoring frameworks
2. Describe In-depth analysis of technologies related to sentiment analysis
3. Design, Develop and Evaluate a new user feedback analysis technique using sentiment analysis for mobile purchases at Amazon.com

1.6 Study Limitations

This project aims at providing a more refined analytic technique to evaluate mobile phone product related user feedback at a glance in the Amazon ecommerce website.

1.7 Chapter Outline

The rest of the thesis is structured as follows. Chapter 2 critically reviews the literature on Analysis for the product reviews on e-commerce system and identify the research problems. Chapter 3 is about the technology used for the analysis of reviews on products. Chapter 4 present our new approach to use to calculate product ratings for the e-commerce system. Chapter 5 and Chapter 6 describe the design and implementation respectively. Chapter 7 is a evaluation of the new solution. Chapter 8 concludes the research with a note on further work.

1.8 Chapter Summary

This chapter gave an overall picture of the entire project presented in this thesis. As such we described the background/motivation, problem definition, objectives, and a brief overview of the solution. Next presents a critical review of the literature on analysis for the product reviews on e-commerce systems.

Chapter 2 – Literature Review on Developments and Challenges in Product Quality Analysis

2.1 Introduction

Chapter 1 consists of a comprehensive description of the overall project. This chapter provides a literature review in relation to developments and challenges in Data mining which is presented under three main segments namely; early developments, modern trends and future challenges. At the latter part, models available to solve the problem has been discussed.

2.2 Early development and Modern Trends

A recent research shows that 79% of online customers will not return if their first experience is not at a satisfactory level [15] e-commerce websites are performed on a virtual platform making harder to create the customer relationship since there is no physical contact between the buyer and the seller. Therefore, e-commerce websites should make sure that the customer receives the expected convenience and user experience during their purchasing period in order to make him a loyal customer.

Web crawlers are used to easily access reviews, thus it is difficult for an end user to go through all listed reviews for a product prior making the purchasing decision. As per [16] classifier can be shown as the best solution for this requirement. However, few authors observed that negative comments affect the sale of a product than the positive comments [17]. Further, e-word of mouth (WOM) created a boundless influence on purchase power and decision making even the process is time consuming.

When an online product purchasing decisions are made, customers have to go through a various comments and user reviews. However, examining each review is a tedious task. Sentiment analysis techniques allow analysing sums of information data and abstract sentiments which are useful to make purchasing decisions. Sentiment analysis is the field of computational study which analyses customer’s opinions expressed in digitally written language by processing text to identify opinionated information.

At present, star rating is considered as the deciding factor for product purchase in e-commerce websites. When the Star rating is higher, it is believed to be a great product. But these ratings are not consistent and dependable as these ratings can be manipulated and also can be vague using an automated process or a system [18].

Current online product recommendations are based on direct parameters (Eg: Price, rate) without focusing on customers' experience by analyzing reviews. Therefore, it is required to develop a method to review comments based on sentiment analysis.

2.3 Existing Models

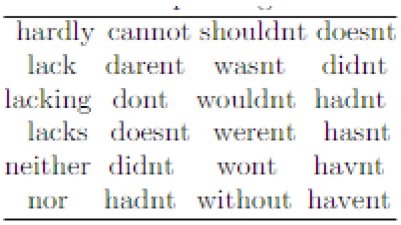
As per the literature review conducted, techniques suitable to solve this problem as in below,

a) Lexicon Based Model – a model which used to extract product features using a syntactic pattern tree-based classifier [19].
b) Unsupervised Word Alignment Model– model that used to check the positioning of the words and frequency of the occurrence of words. [20].
c) Semi supervised Word Alignment Model– a model that is used to differentiate and distinct informal and formal texts [20].

Based on above models, Lexicon and Word alignment models can be combining to develop a new demi-controlled model based on lexicon which is implement using the lexicon data sources (Eg: SentiWordNet, WordNet and Attempto Controlled English Lexicon). To process formless data and extract valuable information from those data SentiWordNet can be used. WordNet can be useful to create sets of English words of synonyms and speech classification can be derived using Attempto model [21].

2.4 Lexicon-Based Approach

Lexicon-based methodology is also known as a dictionary approach and depend on a lexi-con or dictionary of words with pre-calculated polarity. At times this technique is considered to be part of the Machine Learning Unsupervised method, however we quality of classification in the lexicon-based method depends exclusively on the quality of the lexicon.



hardly	cannot	shouldnt	doesnt
lack	darent	wasnt	didnt
lacking	dont	wouldnt	hadnt
lacks	doesnt	werent	hasnt
neither	didnt	wont	havnt
nor	hadnt	without	havent

Figure 2.4.1: Example of Negative words

Ever since the creation of a lexicon is the central part of the lexicon-based method, many academics produced their particular sentiment lexicons. Among them:

1. Opinion Lexicon [22]

2. SentiWordNet [23]

3. AFINN Lexicon [24]

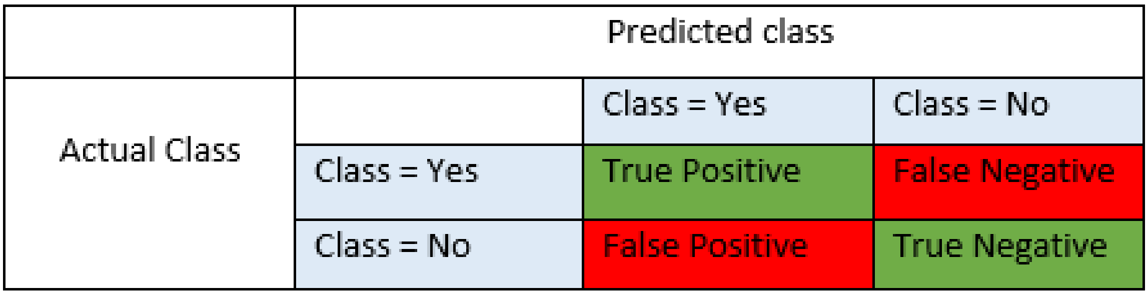
4. LoughranMcDonald Lexicon

5. NRC-Hashtag [25]

6. Harvard Inquirer Lexicon

2.5 Confusion Matrix

A confusion matrix is a tool that can be used for the evaluation of the developed model. This matrix is a board that is frequently used to measure the effectiveness of a classification model with a data set to identify the true values. The table matrix is a good way of identifying the accuracy of the model based on answers at a given time by using the classifiers.



	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Based on the above matrix, classifications that are properly forecasted are shown by True positive and negative classifiers. Hence this research is focused on minimizing both the negative (false positive and False negative) which illustrate in red color in the graph.

Positive but True (TP)– this is the correctly foreseen positive values. In general, the actual values and the predicted values is Yes or same.

Negative but True (TN) – these are the correctly foreseen negative values. In general, the actual values and the predicted values are No.

When the actual class having a contradiction with the predicted class, it generates False positives and false negatives values.

Positive but False (FP) – generally this is because when the predicted class is Yes but the actual class is No

Negative but False (FN) – generally this is because when the predicted class is No and the actual class is Yes.

Finally, we can use the above factors to calculate the accuracy and other elements.

Accuracy – this can be taken as a relation of all opinions which are expected correctly to the sum of opinions so that this will be the most natural measurement of the model. But when the false positive and false negative data sets are the same identical, accuracy measurement is hard to measure. Therefore, other constraints need to be integrated into the model to improve the performance.

Hence accuracy can be noted as (TP+TN/TP+FP+FN+TN).

Precision – this can be described as the relation of correctly expected positive comments to the sum of all comments. When the false positive rates are going down, it automatically higher the precision rate.

Therefore, the Precision rate can be describe as (TP/TP+FP).

Recall – in other words this can be defined as the user sensitivity. Also, its illustrated the relation between the correct predicted positive opinions and the sum of all opinions. Recall also known as sensitivity.

Hence this (Recall) can be noted as (TP/TP+FN)

F1 score – Outcome of this is generated using the false positives as well as the false negatives together. Therefore, this gives the prejudiced average between recall and precision. This score usually appropriate than the accuracy feature when the dataset class distribution is uneven.

In general, F1 Score can be classified as {2*(Precision x Recall) / (Precision + Recall)}.

2.6 Summary

This chapter contains a comprehensive literature review on the identified problem and the available solutions to fulfill the mentioned requirement. Chapter 3 contains technology for the solution.

Author	Year	Method	Data set	Achieved accuracy in %
Khan and Bashir	2016	Only adjectives	Reviews on Movies	76%
Park *et al.*	2015	Machine learning and domain-specific lexicon	Reviews for Amazon	-
Liu and Li	2010	Machine learning	Movie reviews	74.7%

Chapter 3 – Adopted Technology in Data Mining Tools and Techniques

3.1 Introduction

In the previous chapter, different types of existing methods and attributes for analyzing opinionated reviews using sentiment analysis were discussed and some study gaps were identified. This chapter will present the technologies and framework that is used to achieve and fill the research gap identified in the earlier chapter. To begin with, chapter will explain data mining technology which can be used effectively to examine user opinion data.

There are generally two categories of sentiment analysis,

- Lexical Methods
- Machine Learning Methods

For this study the Lexicon-based Method is used.

3.2 Comparison of Machine Learning and Lexicon Based Method

Lexical Methods: These methods service dictionaries of words annotated with their semantic polarity and sentiment strength. This is then used to compute a total for the polarity and/or sentiment of the document. Frequently this technique gives high precision but low recall.

Machine Learning Methods: Such methods involve creating a model by teaching the classifier with labeled samples. This means that you must first collect a dataset with samples for positive, negative and neutral classes, mine the features from the samples and then train the algorithm built on the samples. These approaches are used mostly for calculating the polarity of the document.

Choice of the technique deeply depends on the application, area and language. By lexicon based techniques with large dictionaries permits us to attain very decent results. However, they need using a lexicon, rather which is not always accessible in all languages.

On the other hand Machine learning centered methods provide good results, but they require gaining training on labeled statistics. For this study, it's very tough to obtain training on labeled data.

3.3 What is Data Mining?

The process of Data mining is a technique that is used to extract useful information from a large variety of data. This concept integrates with many domains such as machine learning, patterns recognition, statistics and many more. Currently, at the present world, most of the firms or business entities use this technology as a part of their marketing strategy to increase their growth in profit as well as in customer engagement. Data mining as a technology has numerous outcomes which are useful such as classification, clustering, mining of text, analysis of sequences and predictions. At present there is a new addition to the data mining domain which is sentimental analysis and analysis on social networks.

The whole procedure of discovery useful information in raw data includes in the sequential line up of steps such as developing and understanding of the application area, creating aim data set centered on an intelligent way of choosing, cleaning, integration, transform data and do data mining to evaluate and present data.

Data mining process is shown in Figure -3.3.1.

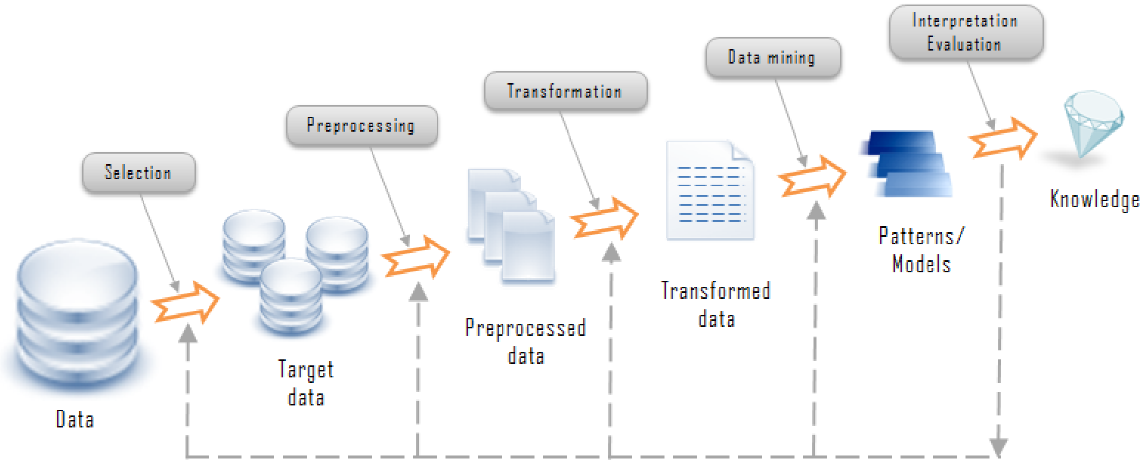


Figure 3.3.1: Steps of data mining Process

3.4 R Data Mining Tool

R offers a wide-ranging stats and graphs approaches. R comprises of numerous packages (more than 7324 packs) access through CRAN3 and provide for websites like Bitconductor., R Forge and GitHub. Moreover, in academics and industry this R is heavily used. R comprises different tasks vies and also some package collections are focused on data mining such as;

1. Learnings related techniques
   - Machine learning methods
   - Statistical learning methods
   - Multivariate Statistics
2. Analysis techniques
   - Cluster Analysis
   - Time Series Analysis,
   - Spatial data analysis
3. Processing of natural languages,

It is important to state that R can run on several operating systems like Linux and Windows and the IDE of R is joint development using RStudio. Therefore, RStudio can be notified as a useful tool to program R so that it is strongly recommended that to use RStudio for learning or development purposes.

3.5 Summary

This chapter presented data mining as the technology proposed to analyze consumer reviews to rate the products. In this sense, it is pointed out how data mining offers an efficient and accurate solution for consumer review analysis. The next chapter shows a novel approach to analyzing consumer records through the technology presented here.

Chapter 4 - Approach to Analyze Quality of Products in a System

4.1 Introduction

Chapter 3 discussed the technology for analyzing the consumer reviews. This chapter presents our approach to analyze consumer reviews in detail using data mining under several headings namely input, output, process, users. This chapter describes the selected approach for consumer reviews analysis. Here we describe our novel approach for analyzing Quality of Products in E-commerce

4.2 Users

A number of users who can be benefited by the Analyze Quality of Products in E-commerce systems in multiple ways. More importantly, e-commerce System Company, customers who are buy the products can be directly benefited by this solution.

4.3 Input

This will include all specific product details such as the identification number of the product, the title of the feedback, review, Rating for the feedback the number of products of the retailer in the ecommerce system.

4.4 Output

The output evaluating model will act as the process used to determine the accuracy of the classification. Based on the output next it will calculate the prediction value. Classification is done in this research for positive, negative, neutral reviews. So, the proposed model is to evaluate as a 3x3 matrix with the confusion matrix. So, the output will be the accuracy of the developed model.

4.5 Process

In this process of analyzing consumer reviews with data mining all the standard steps in the knowledge discovery process. In this fetch data, data selection, creating a target data set, Data cleaning and preprocessing. We will achieve this using Lexicon based approach, opinion word, opinion target and pinion analysis for excluding some basic limitation of sentiment analysis using R language.

4.6 Technology

Lexicon based method marks sentiment analysis further fitting. In an existing implementation using Sentiment analysis with lexicon-based approach using R. In the landscape of R, the sentiment R package and the additional common text mining package have been well developed. You can check available sentiment packages and the fanciful RTextTools packages in R.

4.7 Summary

This chapter presented our novel approach to analyze the user reviews to rate the product base on the feedbacks from consumers. It is pointed out how the novel approach offers an efficient and accurate solution using machine learning algorithms in data mining. The next chapter shows the design of the novel approach presented here.

Chapter 5 – Design to Analyze Quality of Products in System

5.1 Introduction

The chapter which was preceding gave a full picture of the entire proposed solution. This chapter describes the design of the process presented in the approach. Here we describe the top-level architecture of the design by elaborating the role of each component of the architecture.

5.2 Top level Architecture of the System

Aligning to proposed model consumers can make decisions easily about the retailers as well as the products which are best. Following are the 4 phases which we going to implement in order to achieve our goal. This shows the basic structure of the design. There are 4 phases. Figure- 5.2.1 illustrate the high-level architecture of the system.

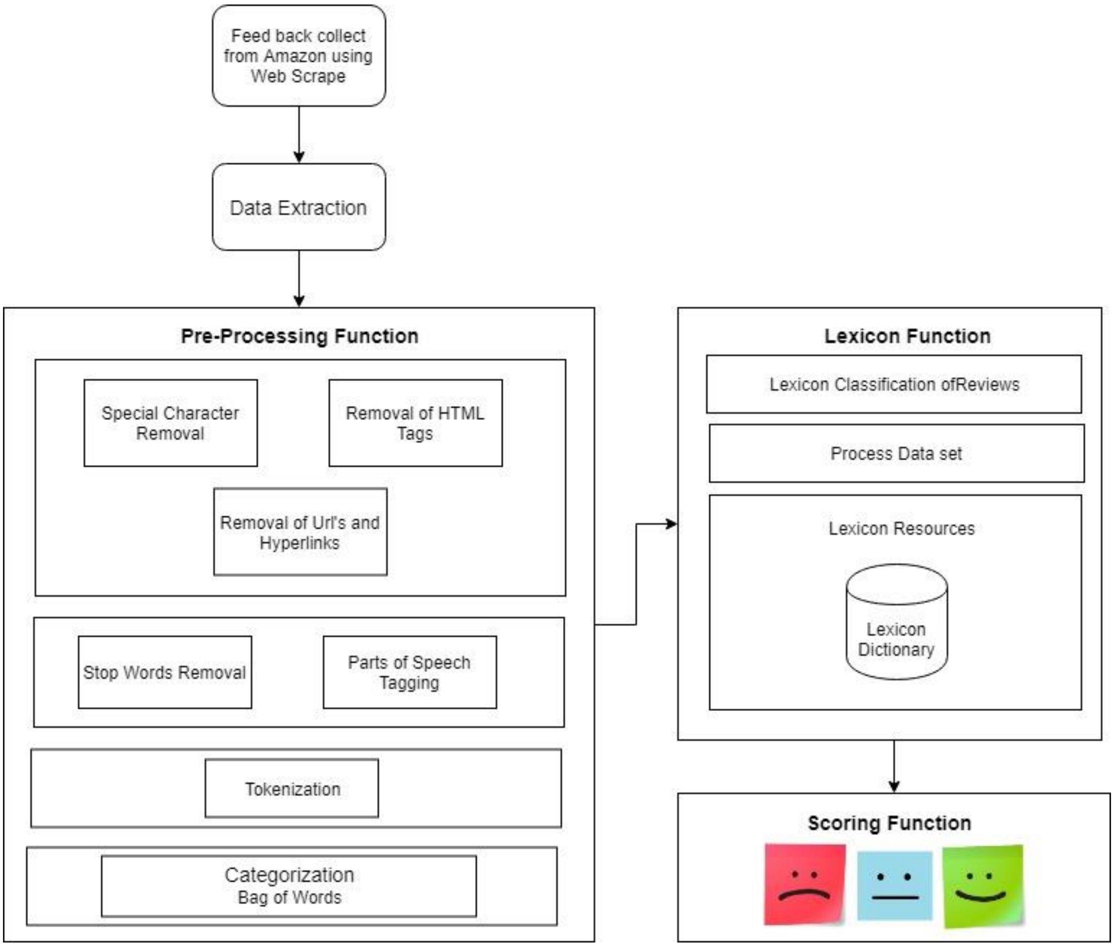


Figure 5.2.1: Proposed system high level architecture

5.3 Data Collection

As the initial step, Amazon was selected as the ecommerce platform and reviews were collected. These data obtain only from the mobile phone category using web scraping. Collected reviews contain large set of information's about the mobile detail reviews. Collected reviews is not in an appropriate format for data analyzing hence preprocessing was required to determine the useful info out of the reviews.

5.4 Pre-Processing

Since the composed data were not in a format for analyzing it was required to do a preprocessing on the collected reviews. These pre-processing steps help to convert noise from high dimensional features to the low dimensional space to obtain as much accurate information's from reviews as possible from the data obtain. Pre-processing data can consist of many steps on the data and situations.

5.4.1 Removal of duplicate

Data set obtain contain duplicate data. So, in pre-processing those duplicate data will be removed. The duplicated function will remove the duplicated reviews from the data set. Removing these duplicated reviews will remove unnecessary processing data and time in the model.

5.4.2 Removal of HTML tags , hyperlinks

Removal of HTML tags, url, hyperlinks from the reviews. From the ‘rm_url’ function these words will be removed.

5.4.3 Stop words removal

These Stop words are frequently used in words in any language. Now a days any application used this techniques to filter text. So eliminating stop words and can focus on important words.

For example, in search engine query is like ‘Who is the President in Srilanka’. This type of query in search engine searches for the words ‘who’, ‘president’, ‘Srilanka’ then it will retrieve more pages containing ‘Srilanka’, ‘president’ word. So it’s not related document which we exactly want because any document contains this stop word. By removing these stop words can get proper correct efficient data to analysis.

‘rm_words’ function can be used to remove stop words. Stop word removal process can be used to improving and standardizing the processing speed of the data for better results.

5.4.4 Tagging Speech

POS tagging stands for part of speech tagging. POS annotates of the word with particular word according to a document in which its use. For example verb (V), adjective (Adj), noun (N), verb phase (VP), noun phrase (NP), preposition (PREP), conjunction (CONJ). POS tagging is necessary for identification of noun, conjunction, adjective and adverb.

Eg: ‘The product is good.’

5.5 Tokenization

Tokenization is the procedure of changing text into tokens before transforming it into vectors. It is also easier to filter out unnecessary tokens. For eg, a document into sections or sentences into words. In this case we are tokenizing the reviews into words. From 'word_tokenize' function can be used for this.

5.6 Categorization

After pre-processed is done with all the reviews, these will be categorizing to features of the mobile phone. This task is done using by Bag of Words. Following categories will be used for this purpose.

- Display
- Performance
- Camera
- Storage
- Battery
- Ram
- Network

5.7 Lexicon Module

In this review can classify in to three aspects in the sentimental analysis, those are positive, negative and neutral reviews. By using previous opinion word identification step, it will extract the option word then like an adjective, adverb, which present in comments. Then Check for available words of positive, negative, neutral for classifications. Following Dictionary used in this process.

5.7.1 Lexicon Dictionary

This covers a list of positive and negative words. This study uses a dictionary of 1980 quantity of positive words and 4800 number of negative words.

If retrieved words equal with the positive list of words then that review is classified as a positive review.

If retrieved word equal with negative word then that review is classify as negative review otherwise neutral review.

5.8 Generating Product Rating

This result of this method will be score generated by mining user review by applying the Lexicon Dictionary. This is using the Lexicon centered method wherever the semantic orientation of the text is calculated by summing the SO of the words and phrases in the document. Before evaluating text will be clean from text preprocessing jobs. It will be a numerical value among ‘0-5’ based on the product review contents. This rate will be calculated by pre-stored words in the databank and their particular assigned values.

Product Rating will be shows as Overall and Category wise.

5.9 Rating Module

The Score is modified by assigning +1 for positive words and -1 for negative words in the lexicon dictionary.

5.10 Validate Model

In order to authenticate the perfect model developed, Amazon product reviews are manually branded as positive, negative or neutral. So confusion matrix needs to be developing as a 3X3 matrix.

A confusion matrix is also called a table of confusion and is a summary of calculation results on a classification problematic. The quantity of correct and incorrect calculations are brief with amount values and broken down by each class. This is the key to the confusion matrix. Following metrics will be generated.

5.10.1 Confusion matrix for positive reviews

Table 5.10.1: Confusion Matrix for Positive Reviews

	Positive Reviews (Predicted)		
True Class for **Positive** Reviews (Actual)		Positive	Non-Positive
	Class Positive	TP I	FN I
	Class Non Positive	FP I	TN I

5.10.2 Confusion matrix for negative reviews

Table 5.10.2: Confusion Matrix for Negative Reviews

	Negative Reviews (Predicted)		
True Class for **Negative** Reviews (Actual)		Negative	Non-Negative
	Class Negative	TP I I	FN I I
	Class Non-Negative	FP I I	TN I I

5.10.3 Confusion matrix for neutral reviews

Table 5.10.3: Confusion Matrix for Neutral Reviews

	Neutral Reviews (Predicted)		
True Class for **Neutral** Reviews (Actual)		Neutral	Non-Neutral
	Class Neutral	TP I I I	FN I I I
	Class Non-Neutral	FP I I I	TN I I I

This can support in computing further advanced classification metrics such as,

- Accuracy (ACC)
- Precision (Positive Predictive Value)
- Recall (Sensitivity, Hit Rate, True Positive Rate)
- F1-Measure (harmonic mean of precision and recall)

5.11 Summary

This chapter provided details on research design and applicability of selected research method for the research. Furthermore, this chapter focuses on top level design for the research and how research question are structured with in the research. Subsequent section will be discussed about implementation details according to this design.

Chapter 6 – Implementation of Analyzing the Quality of the Product in a System

6.1 Introduction

In chapter 5 the top-level design of the solution has been described in terms of what attributes are used to represent customer reviews analysis. This chapter describes the implementation of the problem regarding software, algorithms, method,etc.

6.2 R Language

R language is one of the most famous language is Statistical Programming which is developed by Rss Ihaka and Robert Gentleman in 1993. This language is armed by wide range functions of statistical as well as graphical function approaches. R has main advantage for machine leaning algorithm as well as statistical programming. R has a really good community worldwide and there are public libraries are written that other users can use the features easily.

There are famous applications also using R which is Facebook, Uber, Google as well as Airbnb. Which is given trusted to other users to use this language to development. One of the R’s main feature is to use data analysis. To achieve this R recommended to follow a path of sequence of steps. First step is programming. Second step is transmuting and discovering, followed by modeling the solution and communicate the final output.

- Program: Famous statistical program tool.
- Transform: Data science is a main feature that is powered with wide range of library’s developed by community.
- Discover: Inspect the data which inputs, it helps to improve your hypothesis and then analyses for the end of this step.
- Model: R will give to find the correct model for the input records by using wide range of tools.
- Communicate: Outputting the final results with Codes as well as Graphs will be done by easily.

6.3 Data Collection

Data for this analysis is collected from Amazon (Mobile phone category). Using web scrape for amazon product reviews. Fetch records like Product Id, Product Name, Review Heading, Review, Rating, Sellers stock.

6.4 Data set Preprocessing

Since the factors effecting the problem is identified after the collecting all the related domain knowledge, data needs to be preprocessed as the next step to create the predictive models.

6.5 Categorization of the Dataset

Once pre-processed is done, it will break down to feature wise for the mobile phone domain. So in next level when doing the sentimental analysis, final outcome can be showing with overall and feature wise (Display, Performance, Camera, Storage, Battery, Ram, Network).

6.6 Implementation of the Application

Application is implemented using R language, R studio. After fetching data using web scrappers and pre-processed it will process with our sentimental analysis applying Lexicon based approach. Then after apply Tweaked Lexicon Dictionary for the reviews obtain from the amazon. Then processing all reviews finally showing in text and graphically about the product rating. Product Rating will be showing as Overall and Category wise (Display, Performance, Camera, Storage, Battery, Ram, Network).

6.6.1 Lexicon Module

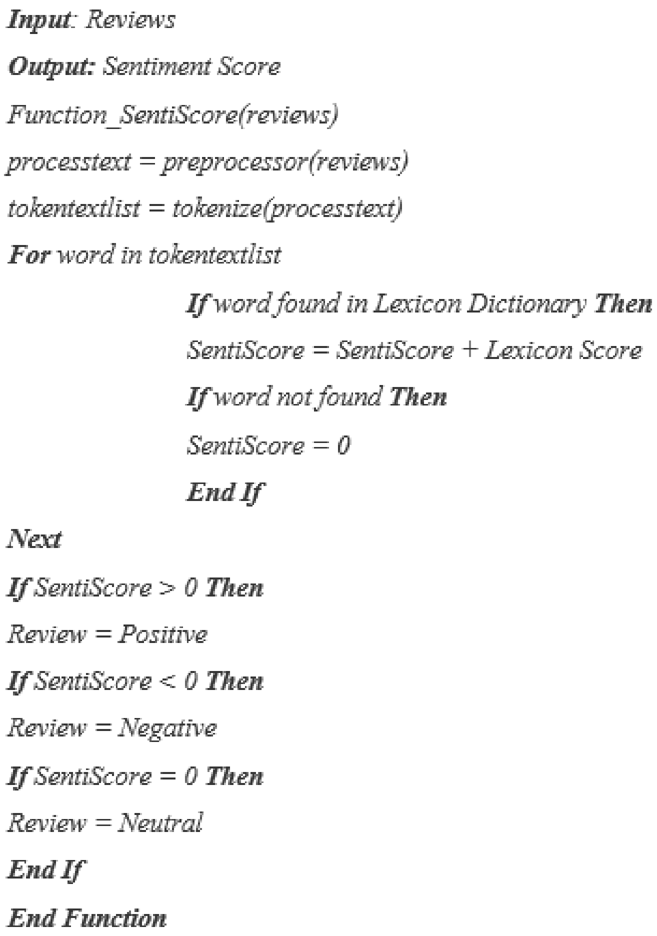
This module is based on the integration of opinion lexicons and dictionary resources for sentimental evaluation. Following dictionaries are used for this research.

6.6.2 Lexicon Dictionary

This contains a list of positive and negative words.

6.6.3 Scoring Algorithm

Sentiment scoring algorithms for the approach is following



```
Input: Reviews
Output: Sentiment Score
Function_SentiScore(reviews)
processtext = preprocessor(reviews)
tokentextlist = tokenize(processtext)
For word in tokentextlist
            If word found in Lexicon Dictionary Then
            SentiScore = SentiScore + Lexicon Score
            If word not found Then
            SentiScore = 0
            End If
Next
If SentiScore > 0 Then
Review = Positive
If SentiScore < 0 Then
Review = Negative
If SentiScore = 0 Then
Review = Neutral
End If
End Function
```

6.7 Summary

Implementation chapter provide the full path in constructing data model for addressing the research questions. Furthermore this chapter gives detail description about using R language to build the model .Next chapter will be on discussion about evaluation.

Chapter 7 - Evaluation

7.1 Introduction

This chapter focuses on how testing strategies carried out for the research question in terms of the evaluation measurements for the selected data mining techniques. The system is tested in terms of users, and how the selected model tested.

7.2 Analysis

Gathered data after web scraping from amazon and preprocessing the amazon reviews from Lexicon dictionaries were applied on the feeds independently and classified the feeds as positive, negative or neutral. Table 7.2.1 shows the performance measures for developed approach and as per that and how Lexicon dictionary performs on this classification.

Table 7.2.1: Lexicon Performance

Rating	Measure	Lexicon Dictionary							
		Overall	Display	Performance	Camera	Storage	Battery	Ram	Network
Positive	Accuracy	85.96%	82.60%	81.80%	83.96%	80.00%	80.67%	81.58%	82.48%
	Precision	79.16%	57.14%	62.54%	58.63%	55.53%	61.13%	59.63%	60.74%
	Recall	86.36%	80.00%	82.46%	80.72%	79.26%	81.00%	80.28%	81.63 %
	F1 Score	82.60%	66.66%	67.00%	71.46%	68.00%	67.35%	67.54%	68.45%
Negative	Accuracy	80.70%	69.56%	71.68%	72.78%	68.14%	68.66%	69.23%	70.00%
	Precision	68.01%	60.00%	62.46%	61.44%	60.62%	61.78%	61.44%	60.86%
	Recall	85.62%	66.66%	68.22%	70.32%	68.62%	69.77%	71.00%	65.82%
	F1 Score	75.55%	63.15%	65.98%	67.25%	62.00%	64.17%	63.00%	64.43%
Neutral	Accuracy	77.19%	60.86%	62.13%	62.51%	61.74%	62.00%	61.00%	60.01%
	Precision	62.50%	55.00%	56.10%	56.31%	55.62%	54.26%	56.27%	55.23%
	Recall	33.33%	31.33%	32.62%	33.00%	31.63%	32.98%	33.00%	33.62%
	F1 Score	43.47%	40.00%	43.27%	42.83%	42.22%	41.96%	43.00%	42.67%

Lexicon dictionary was tweaked more in order to align to Mobile phone field and some of the alterations done were shown on Table 7.2.2.

Table 7.2.2: Tweaked for Mobile Phone Domain

Positive Words		Negative Words	
Added	**Removed**	**Added**	**Removed**
speeds	smart	overheated	swipe
no-scratches		Slow	toxic
facial-id		Heavier	
smart-hdr		high-price	
highest-capacity		couldnt	
		few-hours	
		shaking	
		turn-on	

After tweaked Lexicon Dictionary, the accuracy increased by following (Refer Table 7.2.3),

Positive = 5.26%

Negative = 3.50%

Neutral = 3.50%

Table 7.2.3: Lexicon VS Tweaked Lexicon Overall

Rating	Measure	Lexicon Dictionary	Tweaked	Accuracy
Positive	Accuracy	85.96%	91.22%	5.26%
	Precision	79.16%	86.95%	7.79%
	Recall	86.36%	90.90%	4.54%
	F1 Score	82.60%	88.88%	6.28%
Negative	Accuracy	80.70%	84.20%	3.50%
	Precision	68.01%	74.52%	6.51%
	Recall	85.62%	89.57%	3.95%
	F1 Score	75.55%	84.58%	9.03%
Neutral	Accuracy	77.19%	78.94%	1.75%
	Precision	62.50%	66.66%	4.16%
	Recall	33.33%	40.00%	6.67%
	F1 Score	43.47%	50.00%	6.53%

Accuracy comparison for Lexicon with Tweaked Lexicon for the same data set showing in Figure 7.2.1.

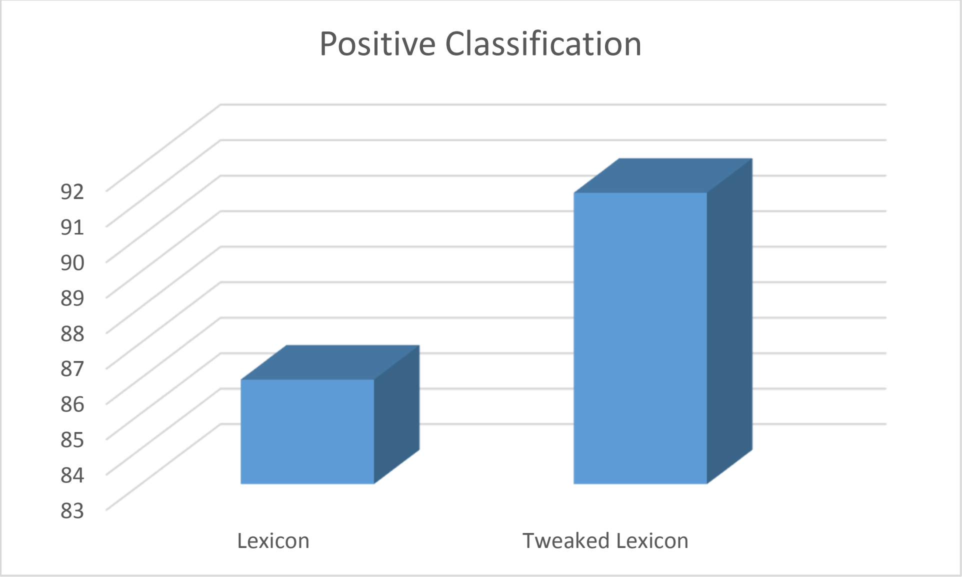


Figure 7.2.1: Positive Accuracy Comparison for Lexicon

We need to analyze the negative reviews specially for getting an understanding on the mobile phone provided by the amazon website. From this customers can choose whether to choose the device to buy or not. As in figure 7.2.2, negative comments accuracy comparison is shown.

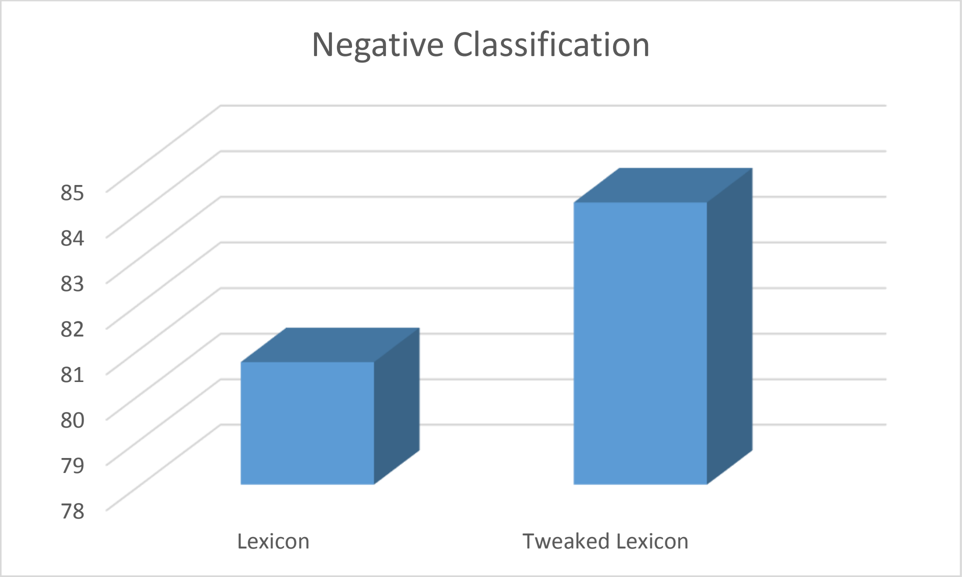


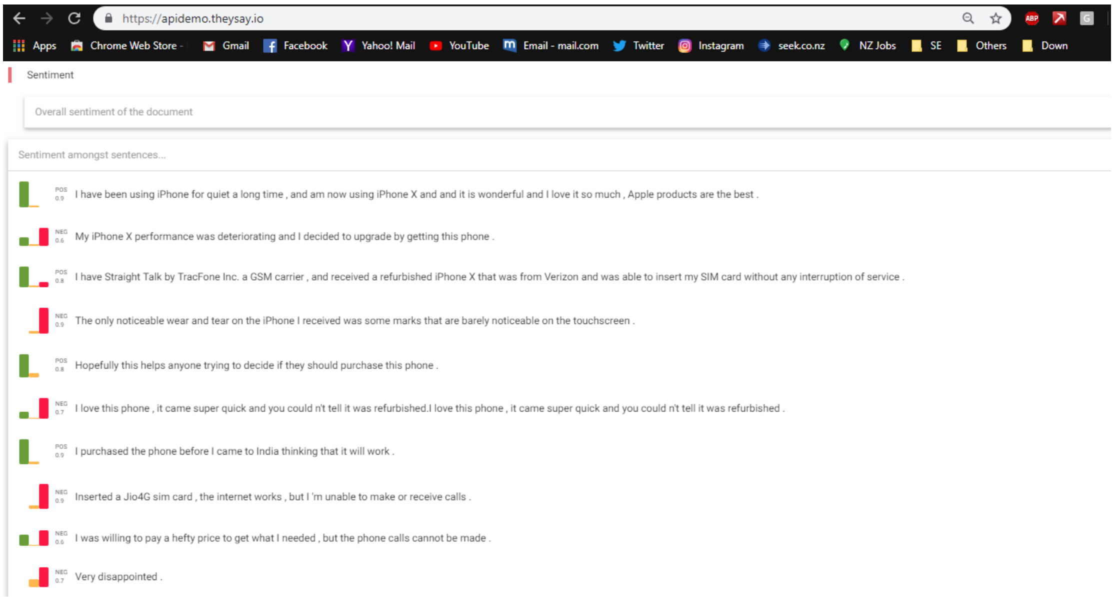
Figure 7.2.2: Negative Accuracy Comparison for Lexicon

7.3 Model Validation with Existing Products

There are existing sentimental analysis products can be found in the internet. Therefor going to compare the accuracy of these products against the proposed model.

7.3.1 TheySay API

TheySay's is a real-time Sentiment Analysis API provides you access to advanced sentiment analysis algorithm through accessible and safe RESTful API facility. The appliance is covered in a platform REST API service that allows software applications, workflows, and services to obtain rich TheySay JSON metadata with insignificant combination work, API service is called PreCeive. They give a web interface to do the sentimental analysis website itself (Figure 7.3.1).



Sentiment

Overall sentiment of the document

Sentiment amongst sentences...

POS 0.9 I have been using iPhone for quiet a long time , and am now using iPhone X and and it is wonderful and I love it so much , Apple products are the best .

NEG 0.6 My iPhone X performance was deteriorating and I decided to upgrade by getting this phone .

POS 0.8 I have Straight Talk by TracFone Inc. a GSM carrier , and received a refurbished iPhone X that was from Verizon and was able to insert my SIM card without any interruption of service .

NEG 0.9 The only noticeable wear and tear on the iPhone I received was some marks that are barely noticeable on the touchscreen .

POS 0.8 Hopefully this helps anyone trying to decide if they should purchase this phone .

NEG 0.7 I love this phone , it came super quick and you could n't tell it was refurbished.I love this phone , it came super quick and you could n't tell it was refurbished .

POS 0.9 I purchased the phone before I came to India thinking that it will work .

NEG 0.9 Inserted a Jio4G sim card , the internet works , but I 'm unable to make or receive calls .

NEG 0.6 I was willing to pay a hefty price to get what I needed , but the phone calls cannot be made .

NEG 0.7 Very disappointed .

Figure 7.3.1: Result of theysay API

As in showing below suggested model is showing higher accuracy than the TheySay API.

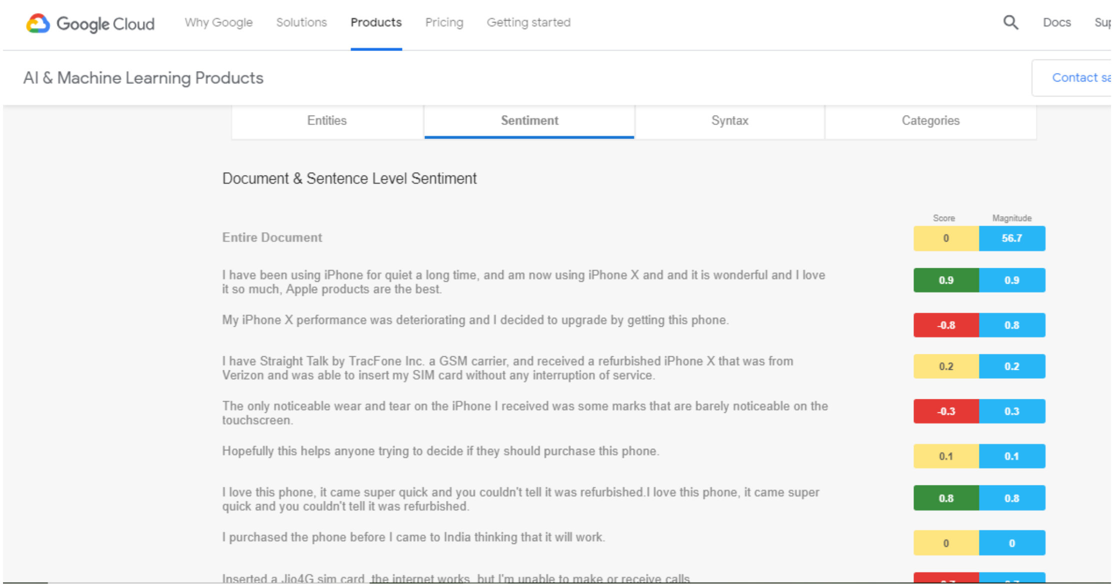
Table 7.3.1: Comparison with TheySay API

Rating	Suggested Model	TheySay
Positive	91.22%	80.27%
Negative	84.20%	75.91%
Neutral	78.94%	69.53%

7.3.2 Google Cloud Natural Language

Google Cloud Natural Language is a famous tool used in by the community. This uses pre-trained machine learning models. This tool can be used by REST API and easy to set with AutoML Natural Language. Sentimental positive score indicates with value greater than zero. Sentimental negative score shows with value less than zero.

- Score: This is the sentiment score that is overall emotion leaning of text shows with range between negative (-1.0) and positive (1.0).
- Magnitude: This is not regularized, and both positive and negative expression followed by emotions include in the sentence. This shows strength of sentiment followed by positive and negative. Magnitude will be grater for longer text also.



Document & Sentence Level Sentiment

	Score	Magnitude
Entire Document	0	56.7
I have been using iPhone for quiet a long time, and am now using iPhone X and and it is wonderful and I love it so much, Apple products are the best.	0.9	0.9
My iPhone X performance was deteriorating and I decided to upgrade by getting this phone.	-0.8	0.8
I have Straight Talk by TracFone Inc. a GSM carrier, and received a refurbished iPhone X that was from Verizon and was able to insert my SIM card without any interruption of service.	0.2	0.2
The only noticeable wear and tear on the iPhone I received was some marks that are barely noticeable on the touchscreen.	-0.3	0.3
Hopefully this helps anyone trying to decide if they should purchase this phone.	0.1	0.1
I love this phone, it came super quick and you couldn't tell it was refurbished.I love this phone, it came super quick and you couldn't tell it was refurbished.	0.8	0.8
I purchased the phone before I came to India thinking that it will work.	0	0
Inserted a Jio4G sim card, the internet works, but I'm unable to make or receive calls		

Figure 7.3.2: Google Cloud Natural Language Results

Using Google Cloud Natural Language, three class classification were prepared on the identical dataset which has been used in this study and assess the accurateness of the classification with that API. As per Table 7.3.2, it shows that the model built for this research work has been shown higher accuracy level compared with the google cloud natural language API.

As in showing below suggested model is showing higher accuracy than the Google Natural Language API.

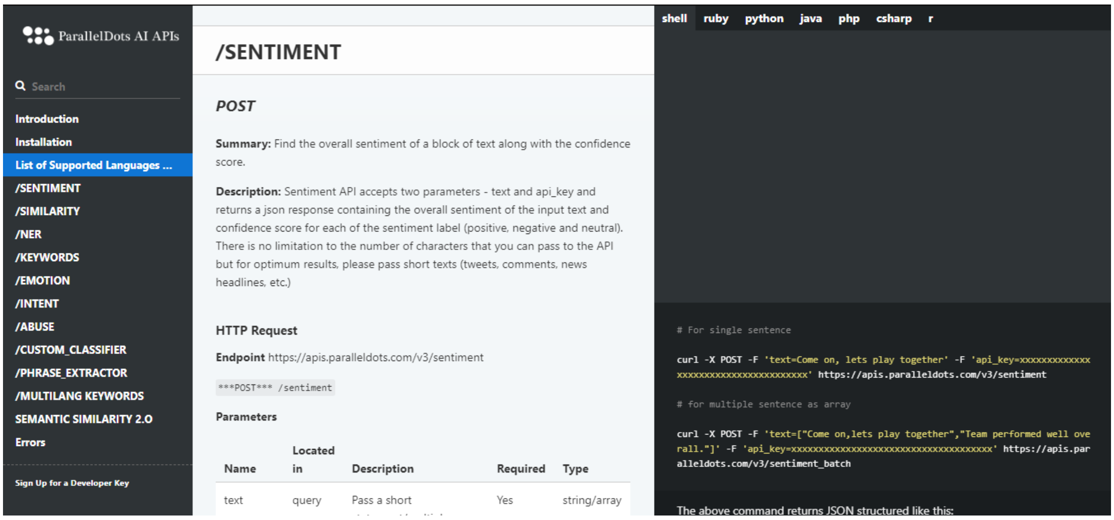
Table 7.3.2: Comparison with Google Cloud Natural Language

Rating	Suggested Model	Google Natural Language
Positive	91.22%	84.67%
Negative	84.20%	78.36%
Neutral	78.94%	71.82%

7.3.3 Paralleldots API

Paralleldots API also a best text analysis tool which using customary patterns for the text. This API support to solve the complex problems which is most famous these days social media analytics , chatbots as well as automation the process.

Paralleldots provides an API to perform sentimental analysis as in Figure 7.3.3.



/SENTIMENT

POST

Summary: Find the overall sentiment of a block of text along with the confidence score.

Description: Sentiment API accepts two parameters - text and api_key and returns a json response containing the overall sentiment of the input text and confidence score for each of the sentiment label (positive, negative and neutral). There is no limitation to the number of characters that you can pass to the API but for optimum results, please pass short texts (tweets, comments, news headlines, etc.)

HTTP Request

Endpoint https://apis.paralleldots.com/v3/sentiment

```
***POST*** /sentiment
```

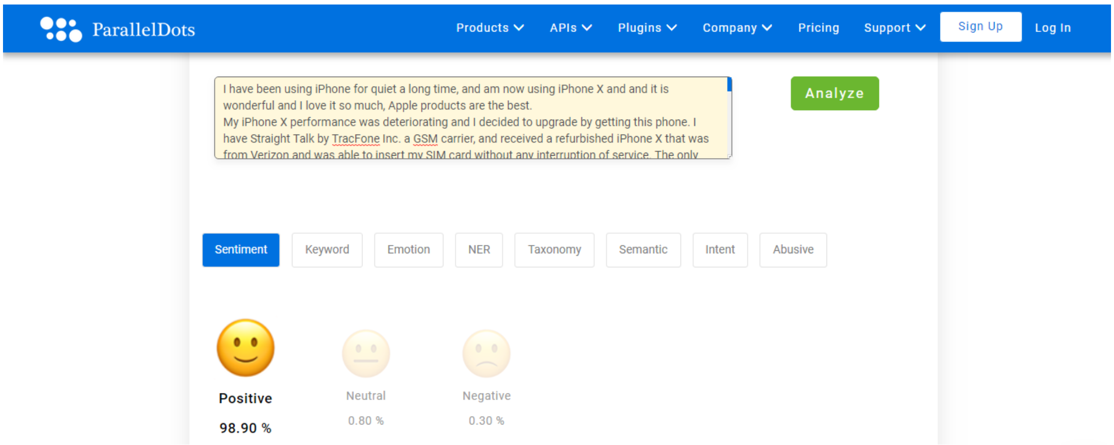
Parameters

Name	Located in	Description	Required	Type
text	query	Pass a short	Yes	string/array

```
# For single sentence

curl -X POST -F 'text=Come on, lets play together' -F 'api_key=xxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxx' https://apis.paralleldots.com/v3/sentiment

# for multiple sentence as array

curl -X POST -F 'text=["Come on,lets play together","Team performed well ove
rall."]' -F 'api_key=xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx' https://apis.par
alleldots.com/v3/sentiment_batch
```

The above command returns JSON structured like this:

Figure 7.3.3: Paralleldots API Details



I have been using iPhone for quiet a long time, and am now using iPhone X and and it is wonderful and I love it so much, Apple products are the best.
My iPhone X performance was deteriorating and I decided to upgrade by getting this phone. I have Straight Talk by TracFone Inc. a GSM carrier, and received a refurbished iPhone X that was from Verizon and was able to insert my SIM card without any interruption of service. The only

Positive	Neutral	Negative
98.90 %	0.80 %	0.30 %

Figure 7.3.4: Paralleldots Results

Comparison between Paralleldots API and suggested model is showing in following Table 7.3.3. As in showing below suggested model is showing higher accuracy than the paralleldots API.

Table 7.3.3: Comparison with Paralleldots API

Rating	**Suggested Model**	**Paralleldots**
Positive	91.22%	78.64%
Negative	84.20%	71.64%
Neutral	78.94%	65.19%

7.4 Overview of the Research

This study will shows which is the best method for carrying out sentiment analysis for the reviews. This will proved mention best method is the most efficient way to do the sentiment with comparing to the third party apps. To achieve this sentiment analysis will be done for the input sentence level applying the methods of Lexican Dictionary and Tweak the Dictionary for the specific domain it self.

Chapter 8 – Conclusion and Recommendations

8.1 Introduction

This chapter focuses on what are conclusion and recommendations for the research. What are the objectives are covered in this research for the sentimental analysis.

8.2 Summary of the research

The goal of this research study is to determine an accurate way to analyze sentiments by refining the sentiment classification presentation and mining characteristics related to sentiments. Hence design and develop new user feedback analysis using sentiment analysis using tweaked Lexicon dictionary-based system. Then processing all reviews scraped from Amazon related to mobile phone finally showing product base rating.

8.3 Conclusions

Through the research it was revealed that the lexicon-based approach was intended to have the highest accuracy compared to other methods which was determined using the reviews from amazon. Furthermore, it was also revealed that the polarity score which was used in all lexicon-based methods mostly rely on the quantity and the quality of the dictionaries that were used in the analysis. From the proposed approach final outcome showing as category wise for mobile phone domain, even it will more accurate when showing the final outcome feature wise.

8.4 Limitations of the Study

This research study is focused to create and framework that can automatically cooperate large scale of prejudiced review data effectively. The proposed and developed framework is tested and evaluated using real data. Hence the proposed framework illustrates efficiency and credibility in accuracy, few research limitations are still needing to be considered.

The main limitation is that this framework is only focused in mobile phone domain. So that the accuracy for other domains are yet to unveil. Next focused limitation is creation of sentiment lexicons. Since to come up with a good set of lexicons it takes more time and human effort. This is the main influencer which impact the accuracy of the proposed model. Moreover, value of the positioning of the semantics which was focused in sentiment terms could be biased based on the approach used. So that some aspects are not considered due to these limitations of the lexicon. Hence there are some common negatives of the sentiment study.

Next limitation is the online reviews or the opinionated data that is used in the proposed system. This is one of the main limitations as this only covers the scope of informal texts which are generated by user reviews that are stated in an online platform. But other opinionated data from other sources or product-based reports are not considered and used to evaluate the model.

8.5 Future Work

Due to inspiring and challenging problems, study of research in sentiment analysis was very dynamic and active for the last couple of years. Although this research is based on a specific domain to fill few gaps related to the study, more study is necessary for more developments and enhancements. The most important enhancement will be the development of the sentiment lexicon as in present lexicon is very limited as in sentiment terms.

Moving further, it is vital do develop lexicons for other domains as well rather than Mobile phones. Hence the developed framework can be appraised in other different fields as future work. Moreover, developing techniques that can auto generate techniques can be used in future study so that the limitation of the study field will be limited and will be less time consuming and reduce the human error.

In addition, sarcastic texts are rare in mobile phone reviews as the users are directly complaining or praising its features. But still negotiating sarcasm text is necessary as this is an erudite speech form which used in other domains. So, in that case it is better to do more study in terms of sarcasm speech terms, hence to facilitate sarcasm text in reviews in the future.

Furthermore, need to do a study based on review classifications and characteristics on mobile phones with different operating systems, different memory and POS terminals which are developed using Java.

In overall the research is only based on English language. Since nowadays users used Sinhala and Singlish fonts to describe reviews, the proposed framework will be not eligible. Hence sentiment lexicons should also develop for other languages so that scope of this sentiment analysis study can be stretched further.

Reference

[1] Statista, "Global retail e-commerce sales 2014-2021," 1 January 2019. [Online]. Available: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[2] M. Meuter, D. McCabe and J. Curran, "Are all forms of Word-of Mouth equally influential?," in *Electronic Word of Mouth Versus Interpersonal Word of Mouth*, Service Marketing Quartely, 2013, pp. 240-256.

[3] J. Ahrens, J. Coyle and M. Strahileyitz, "Electronic Word of Mouth: The effects of incentives on e-referrals by senders and receivers.," *European Journal of Marketing*, pp. 1034-1051, 2013.

[4] S. George, "How Much Data Is Generated Every Minute On Social Media," 2015. [Online]. Available: http://wersm.com/how-much-data-is-generated-everyminute-on-social-media/. [Accessed 26 January 2019].

[5] A. Joshi, T. Finin, A. Java, A. Kale and P. Kolari, "In Proceedings of the NSF symposium on next generation of data," in *Web 2.0 mining: Analyzing*, 2007, p. 28.

[6] A. S. Cantallops and S. F. , "New consumer behavior: A review of research on eWOM and hotels," *International Journal of Hospitality Management,*, vol. 36, pp. 41-51, 2014.

[7] K. Shrestha, "'50 Stats You Need to Know About Online Reviews," 2016. [Online]. Available: https://www.vendasta.com/blog/50-stats-you-need-to-know-aboutonline-reviews/. [Accessed 27 Janurary 2019].

[8] Z. Stone, "A Surprisingly Large Amount of Amazon Reviews Are Fake," 2015. [Online]. Available: http://thehustle.co/a-surprisingly-large-number-of-amazon-reviewsare-scams-the-hustle-investigates. [Accessed 27 January 2019].

[9] G. Charlton, "Ecommerce consumer reviews: why you need them and how to use them," 2012. [Online]. Available: Econsultancy. com.. [Accessed 27 January 2017].

[10] iPerceptions Releases Retail, "E-Commerce Industry Report," 2011. [Online]. Available: : https://uk.finance.yahoo.com/news/iPerceptions-ReleasesRetail-iw-1564944333.html. [Accessed 27 January 2019].

[11] Emarketer, "Mons place trust in other consumers," 2010. [Online]. Available: http://www.emarketer.com/Article/Moms-Place-Trust-Other-Consumers/1007509. [Accessed 27 January 2019].

[12] comScore and the Kelsey group , "Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior," 2007. [Online]. Available: https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior?cs_edgescape_cc=UShit. [Accessed 27 January 2019].

[13] S. Michael, "Social Media Marketing Industry Report," 2012. [Online]. Available: :http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2012.pdf. [Accessed 27 January 2019].

[14] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

[15] Yottaa, "Key factors to accelerating e-commerce success in 2012," 7 December 2012. [Online]. Available: http://blog.yottaa.com/wpcontent/uploads/2011/12/7/ecommerce-2012-Ebook.pdf. [Accessed 27 January 2019].

[16] M. Arun Monicka Raja , S. Godfrey Winster and S. Swamynathan , "Review Analyzer: Analyzing Consumer Product Reviews from Review Collections," in *IEEE International Conference on Recent Advances in Computing and Software System*, 2012.

[17] S. Ansari, R. Kohavi and L. Mason, "Integrating Ecommerce and Data Mining," in *IEEE International Conference on Data Mining*, 2011.

[18] Dr.S.Murugavalli1, U.Bagirathan, R.Saiprassanth and S.Arvindkuma, "Feedback analysis using Sentiment Analysis for E-commerce," 2017.

[19] Y. S. HuLi, "WordNet based lexicon model for CNL," in *IEEE proceeding at 2009 International Conference on Test and Measurement*, 2009.

[20] K. Liu, L. X. and J. Z. , "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Mode," in *IEEE proceeding at IEEE Transactions on Knowledge and Data Engineering* , 2014.

[21] V. Singh, R. Priyani, P. Uddin and P. W. Marisha , "Sentiment Analysis of Textual Reviews ,Evaluating," in *proceeding in IEEE 2013 5th International Conference on Knowledge and Smart Technology* , 2013.

[22] M. H. a. B. Liu, "Opinion lexicon," 2004.

[23] A. E. a. F. Sebastiani, "Sentiwordnet: A publicly available lexical resource foropinion mining," 2006.

[24] F. Nielsen, "A new anew: evaluation of a word list for sentiment analysis inmicroblogs," p. pp. 93–98, 2011.

[25] S. K. a. X. Z. S. M. Mohammad, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," 2013.

Appendixes

Appendix A – R code for Web Scrape from Amazon

```
pacman::p_load(XML, dplyr, stringr, rvest, audio)

#Remove all white space
trim <- function (x) gsub("^\\s+|\\s+$", "", x)

#prod_code = "B0043WCH66"
prod_code = "B01N9YOF3R"
url <- paste0("https://www.amazon.com/dp/", prod_code)
doc <- read_html(url)

#obtain the text in the node, remove "\n" from the text, and remove white space
prod <- html_nodes(doc, "#productTitle") %>% html_text() %>% gsub("\n", "", .)
%>% trim()
prod

#Source function to Parse Amazon html pages for data
source("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\amazon_scraper.R")

pages <- 100
amazon_reviews_list <- NULL
for(page_num in 0:pages){
  url <- paste0("http://www.amazon.com/product-
reviews/",prod_code,"/?pageNumber=", page_num)
  doc <- read_html(url)

  amazon_reviews <- amazon_scraper(doc, reviewer = F, delay = 2)
  amazon_reviews_list <- rbind(amazon_reviews_list, cbind(prod, amazon_reviews))
}
write.csv(amazon_reviews_list,"amazon_reviews.csv")
```

Appendix B – R code for Web Scrape Parse Amazon html pages for data

```
#Parse Amazon html pages for data
amazon_scraper <- function(doc, reviewer = T, delay = 0){

  if(!"pacman" %in% installed.packages()[,"Package"]) install.packages("pacman")
  pacman::p_load_gh("trinker/sentimentr")
  pacman::p_load(RCurl, XML, dplyr, stringr, rvest, audio)

  sec = 0
  if(delay < 0) warning("delay was less than 0: set to 0")
  if(delay > 0) sec = max(0, delay + runif(1, -1, 1))

  #Remove all white space
  trim <- function (x) gsub("^\\s+|\\s+$", "", x)

  title <- doc %>%
    html_nodes("#cm_cr-review_list .a-color-base") %>%
    html_text()

  author <- doc %>%
    html_nodes(".review-byline .author") %>%
    html_text()

  date <- doc %>%
    html_nodes("#cm_cr-review_list .review-date") %>%
    html_text() %>%
    gsub(".*on ", "", .)

  ver.purchase <- doc%>%
    html_nodes(".review-data.a-spacing-mini") %>%
    html_text() %>%
    grepl("Verified Purchase", .) %>%
    as.numeric()

  format <- doc %>%
    html_nodes(".review-data.a-spacing-mini") %>%
    html_text() %>%
    gsub("Color: |\\|.*|Verified.*", "", .)
    #if(length(format) == 0) format <- NA

  stars <- doc %>%
    html_nodes("#cm_cr-review_list  .review-rating") %>%
    html_text() %>%
    str_extract("\\d") %>%
    as.numeric()

  comments <- doc %>%
    html_nodes("#cm_cr-review_list .review-text") %>%
    html_text()

  helpful <- doc %>%
    html_nodes(".cr-vote-buttons .a-color-secondary") %>%
    html_text() %>%
    str_extract("[:digit:]+|One") %>%
    gsub("One", "1", .) %>%
    as.numeric()

  if(reviewer == T){
```

```
rver_url <- doc %>%
  html_nodes(".review-byline .author") %>%
  html_attr("href") %>%
  gsub("/ref=cm_cr_othr_d_pdp\\?ie=UTF8", "", .) %>%
  gsub("/gp/pdp/profile/", "", .) %>%
  paste0("https://www.amazon.com/gp/cdp/member-reviews/",.)

#average rating of past 10 reviews
rver_avgrating_10 <- rver_url %>%
  sapply(., function(x) {
      read_html(x) %>%
      html_nodes(".small span img") %>%
      html_attr("title") %>%
      gsub("out of.*|stars", "", .) %>%
      as.numeric() %>%
      mean(na.rm = T)
  }) %>% as.numeric()

rver_prof <- rver_url %>%
  sapply(., function(x)
    read_html(x) %>%
      html_nodes("div.small, td td td .tiny") %>%
      html_text()
  )

rver_numrev <- rver_prof %>%
  lapply(., function(x)
    gsub("\n  Customer Reviews: |\n", "", x[1])
```

```
      ) %>% as.numeric()

    rver_numhelpful <- rver_prof %>%
      lapply(., function(x)
        gsub(".*Helpful Votes:|\n", "", x[2]) %>%
          trim()
      ) %>% as.numeric()

    rver_rank <- rver_prof %>%
      lapply(., function(x)
        gsub(".*Top Reviewer Ranking:|Helpful Votes:.*|\n", "", x[2]) %>%
          removePunctuation() %>%
          trim()
      ) %>% as.numeric()

    df <- data.frame(title, date, ver.purchase, format, stars, comments, helpful,
                     rver_url, rver_avgrating_10, rver_numrev, rver_numhelpful, rver_rank, stringsAsFactors = F)

  } else df <- data.frame(title, author, date, ver.purchase, format, stars, comments, helpful, stringsAsFactors = F)

  return(df)
}
```

Appendix C – R code for Review categorized by features

```
FeatureList=c('display','performance','camera','storage','battery', 'ram', 'network')


Display=c('touchscreen','display aspect ratio', 'flexible
display','crysta','notch','screen','flickering','scratches','screen protector')

Performance =c('mobile computing','micro
processor','upgrade','horsepower','tasks','slower','lags','hardwear', 'overheated')

Camera =c('camera phone', 'picture quality', 'pictures', 'focus')

Storage=c('storage','gb')

Battery=c('battery','battery life', 'charger','cable')

Ram=c('flash memory','memory')

Network =c('cellular network', 'service', 'sim card', 'internet','calls','recived
calls','VoLTE','lte','voicemail','call','sim')

data <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC Project\\sentimental_source\\Mobile
Review.csv", header = TRUE, select=c("X","review","Rating"))

############Split string ##########################################

library(stringr)

word.list = str_split(data$review, '\\s+')

word.list


for (feature in FeatureList){

  #####Search word#####

  library(qdap)

  WordWrapOne<-word.list[grep("camera",word.list)]

  WordWrapOne


  lapply(WordWrapOne, bag_o_words)




  WordWrapTwo<-word.list[grep("picture",word.list)]
```

```
 WordWrapTwo

 WordWrapThree<-word.list[grep("focus",word.list)]
 WordWrapThree

 #####Bind wordss#####
 ReviewList<-cbind(WordWrapOne,WordWrapTwo,WordWrapThree)
 ReviewList

 url <- "C:\\Users\\asanka_09441\\Desktop\\MSC Project\\sentimental_source\\"
 url <- paste(url, feature, "_reviews.csv", sep="")

 #####Write to the file#####
 Cam<-write.csv(ReviewList,url)

}
```

Appendix D – R code for Sentimental Analysis using Lexicon Dictionary

```
#Appendix A -  R code for using only Lexicon Dictionary


#Install & Load Required R packages

#install.packages("data.table")

library(data.table)

library(qdapRegex)

library(plyr)

library(stringr)

library(qdap)


#Import Amazon Feeds

reviews <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\Mobile Review.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsDisplay <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\display_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsCamera <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\camera_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsPerformance <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\performance_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsStorage <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\storage_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsBattery <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\battery_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsRam <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\ram_reviews.csv", header = TRUE,
select=c("x","review","Rating"))

reviewsNetwork <- fread ("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\network_reviews.csv", header = TRUE,
select=c("x","review","Rating"))
```

```
FeatureList=c('overall','display','camera', 'performance', 'storage', 'battery','ram', 'network')
#FeatureList=c('overall', 'display','performance','camera','storage','battery', 'ram', 'network')

data <- 1
elementName <- "Overall Results"
finalResults <- list()

for (feature in FeatureList){

  if(feature == 'overall'){
    data <- reviews
    elementName <- "Overall Results"
  } else if(feature == 'display'){
    data <- reviewsDisplay
    elementName <- "Display Results"
  }else if(feature == 'camera'){
    data <- reviewsCamera
    elementName <- "Camera Results"
  }else if(feature == 'performance'){
    data <- reviewsPerformance
    elementName <- "Performance Results"
  }else if(feature == 'storage'){
    data <- reviewsStorage
    elementName <- "Storage Results"
  }else if(feature == 'battery'){
    data <- reviewsBattery
    elementName <- "Battery Results"
  }else if(feature == 'ram'){
    data <- reviewsRam
    elementName <- "Ram Results"
  }
  else if(feature == 'network'){
    data <- reviewsNetwork
    elementName <- "Network Results"
  }



  #Remove Duplicate reviews
  data<- data[!duplicated(data), ]



  #Duplicate "review" field column
  data$OriginalText = data$review



  #Lower all the letters
  data$review<-tolower(data$review)




  #Text Preprocessing
  ##Removal of Reviews
  data$review <- gsub("RT @[a-z,A-Z]*:", "", data$review)


  ##Removal of HTML Links - Need qdapRegex Package to use rm_url
  data$review <- rm_url(data$review, pattern=pastex("@rm_twitter_url", "@rm_url"))
```

```
##Removal of @People
data$review <- gsub("@\\w+", "", data$review)


##Removal of Special Characters ? & .
data$review <- gsub("?", " ", data$review, fixed = TRUE)
data$review <- gsub(".", " ", data$review, fixed = TRUE)
data$review <- gsub("!", " ", data$review, fixed = TRUE)
data$review <- gsub("\"", " ", data$review, fixed = TRUE)


#Text Refinement
#Remove Stop words
Remove_Words <- function(string, words) {
  stopifnot(is.character(string), is.character(words))
  spltted <- strsplit(string, " ", fixed = TRUE) # fixed = TRUE for speedup
  vapply(spltted, function(data) paste(data[!tolower(data) %in% words], collapse = " "), character(1))
}


data$review <-Remove_Words(data$review, tm::stopwords("en"))

#Part Of SPeech Tagging
```

```
score.pos = function(sentences)
{
 # sentences <- tokenize_sentences(text, simplify = TRUE)
 # unipostagger <- rdr_model(language = "UD_English", annotation = "UniversalPOS")
  #unipostags <- rdr_pos(unipostagger, sentences)
  #unipostags$word.type <- unipostag_types[unipostags$word.type]
  return (sentences)
}



#Function for Positive & Negative Words match
score.sentimentGenerator = function(reviews, positive.words, negative.words,
.progress='none')
{
  ##require libs
  require(plyr)
  require(stringr)

  #function for score
  scoresReviews = laply(reviews, function(review, positive.words, negative.words) {

    # review to lower case:
    review = tolower(review)

    # words split from stringr package
    word.list = str_split(review, '\\s+')
    # unlist the list because its too much
    words = unlist(word.list)

    # checking tokenize words with positive and negative from the dictionaries
    pos.matches = match(words, positive.words)
    neg.matches = match(words, negative.words)


    # returns positive of matched or NA
    # True or False will be return
    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)


    # true or false will be match as 1/0 by sum():
    score = sum(pos.matches) - sum(neg.matches)


    return(score)
   }, positive.words, negative.words, .progress=.progress )


  scoresReviews.df = data.frame(score=scoresReviews, text=reviews)
  return(scoresReviews.df)
 }




 #Import Positive & Negative Words
 positive_words <- scan("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\Positive_Tweak.txt", what='character', comment.char=';')


 negative_words <- scan("C:\\Users\\asanka_09441\\Desktop\\MSC
Project\\sentimental_source\\Negative_Tweak.txt", what='character', comment.char=';')




 #Add additional words to dictionaries
 #negative_words = c(negative_words,'no')


 #Score based on POS words
 result2 <- score.pos( data$review)


 #Score based on Positive & Negative words

 result1 <- score.sentimentGenerator( data$review, positive_words, negative_words)

 #result1 <- rata$googlerating


 #Calculate Lexicon Score

 data$Lexicon = result1$score


 #Calculate Total Score for Approach 01

 data$TotalScore = (data$Lexicon)




 #Assign Predicted Rating for Approach 01

 data$PredictedRating = ifelse(data$TotalScore < 0, 'Negative',
(ifelse(data$TotalScore==0,'Neutral','Positive')))




 #Calculate Performance Matrix


 #For Positive Reviews

 TP1 = sum(ifelse(data$PredictedRating == 'Positive' & data$Rating == 'Positive', 1, 0))

 FP1 = sum(ifelse(data$PredictedRating == 'Positive' & data$Rating != 'Positive', 1, 0))

 FN1 = sum(ifelse(data$PredictedRating != 'Positive' & data$Rating == 'Positive', 1, 0))

 TN1 = sum(ifelse(data$PredictedRating != 'Positive' & data$Rating != 'Positive', 1, 0))


 matrix <- matrix(ncol=4,nrow=3,byrow=TRUE)

 rownames(matrix) <- c("Positive","Negative","Neutral")

 colnames(matrix) <- c("Accuracy","Precision","Recall","F1 Measure")
```

```
#Performance Metrics
matrix[1,1] = (TP1+TN1)/(TP1+FP1+FN1+TN1)
matrix[1,2] = TP1/(TP1+FP1)
matrix[1,3] = TP1/(TP1+FN1)
matrix[1,4]= (2*matrix[1,2]*matrix[1,3])/(matrix[1,2]+matrix[1,3])



#For Negative Reviews
TP2 = sum(ifelse(data$PredictedRating == 'Negative' & data$Rating == 'Negative', 1, 0))
FP2 = sum(ifelse(data$PredictedRating == 'Negative' & data$Rating != 'Negative', 1, 0))
FN2 = sum(ifelse(data$PredictedRating != 'Negative' & data$Rating == 'Negative', 1, 0))
TN2 = sum(ifelse(data$PredictedRating != 'Negative' & data$Rating != 'Negative', 1, 0))



#Performance Metrics
matrix[2,1] = (TP2+TN2)/(TP2+FP2+FN2+TN2)
matrix[2,2] = TP2/(TP2+FP2)
matrix[2,3] = TP2/(TP2+FN2)
matrix[2,4]= (2*matrix[2,2]*matrix[2,3])/(matrix[2,2]+matrix[2,3])



#For Neutral Reviews
TP3 = sum(ifelse(data$PredictedRating == 'Neutral' & data$Rating == 'Neutral', 1, 0))
FP3 = sum(ifelse(data$PredictedRating == 'Neutral' & data$Rating != 'Neutral', 1, 0))
```

```
  FN3 = sum(ifelse(data$PredictedRating != 'Neutral' & data$Rating == 'Neutral', 1, 0))
  TN3 = sum(ifelse(data$PredictedRating != 'Neutral' & data$Rating != 'Neutral', 1, 0))

  #Performance Metrics
  matrix[3,1] = (TP3+TN3)/(TP3+FP3+FN3+TN3)
  matrix[3,2] = TP3/(TP3+FP3)
  matrix[3,3] = TP3/(TP3+FN3)
  matrix[3,4]= (2*matrix[3,2]*matrix[3,3])/(matrix[3,2]+matrix[3,3])

  finalResults[[paste0(elementName, "")]] <- matrix
}

print("Sentimental Results")
finalResults
```