Framework for Sentiment Classification for Morphologically Rich Languages: A Case Study for Sinhala

Nishantha Medagoda

PhD

2017

Framework for Sentiment Classification for Morphologically Rich Languages: A Case Study for Sinhala

By

Nishantha Medagoda

A Thesis Submitted to Auckland University of Technology

in fulfillment of the requirement

for the degree of Doctor of Philosophy

Supervisors:

Assoc. Prof. Russel Pears

Assoc. Prof. Jacqueline Whalley

Prof. Philp Sallis

2017

School of Engineering, Computer and Mathematical Sciences

Abstract

This thesis presents a framework for sentiment analysis for morphologically rich languages. Sentiment analysis is the domain of analysing and extracting people's emotions, feelings, expressions, attitudes and experiences expressed in texts especially, in the digital media, such as web blogs, customer reviews. The primary issue of applying the contemporary sentiment classification techniques for morphologically rich languages is the unavailability of lexical resources. That is these techniques are highly resourced intensive, and the required lexical resources are not freely available for such languages. In addition, the methods are weak in adapting to the linguistic complexities that are shown in morphologically rich languages.

The thesis and the related publications represent the first ever attempt of sentiment analysis for the Sinhala language, which is said to be a highly morphologically rich language. The thesis proposed novel approaches for generating the lexical resources for sentiment classification using limited resources. The first approach examined the cross-linguistic sentiment lexicon generation by considering a sentiment lexicon for English and basic dictionary of the target morphological rich language. In the subsequent task, a sentiment lexicon was generated using the novel approach incorporating morphological features. These morphological features include affixes; prefixes and suffixes. Thirdly, a graph based method was proposed to compile a lexical resource for sentiment classification with polarity scores. The researcher investigated the classical text classification techniques for Sinhala. The thesis identified the best classification algorithm for Sinhala with dominant linguistic features. Finally, an extensive set of experiments that demonstrated the exploration of language-specific classification features for Sinhala. These language-specific features include part of speech, negation, intensifiers and shifters. We introduce and discuss rule-based approaches to incorporate negations and intensifiers. The research contributes to sentiment classification for morphologically rich languages by proposing the framework that uses limited resources to build the lexical resources and efficient algorithms to classify opinions. The achievements confirm, concerning classification accuracies, the feasibility of sentiment classification for morphologically rich languages such as Sinhala. In addition, the achieved accuracies would be benchmarks for sentiment classification for Sinhala as well as other morphologically rich languages. Based on the promising outcomes and the simplicity, the proposed framework can be applied to any morphologically rich language.

Table of Contents

List of Figures

List of Tables

Attestation and Authorship

"I hereby declare that this submission is my own work and that, to the my best of knowledge and belief, it contains no material previously published of written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning."

……..
Signature

Acronyms and Abbreviations

Symbol	Meaning
Adj	Adjective
Adv	Adverb
Gloss	An explanation, interpretation, or paraphrase of a text
idf	inverse document frequency
kNN	k-nearest neighbor
ML	Machine Learning
Morpheme	A meaningful morphological unit of a language that cannot be further
NB	Naïve Bayes
Phoneme	the smallest unit of speech that can be used to make one word different from another word
POS	Part of Speech
PMI	Pointwise Mutual Information
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
tf	Term frequency

Acknowledgement

I would like to express my sincere gratitude to my primary supervisor Associate Professor Russel Pears for his immense interest, knowledge, and moral support. I would also like to thank my second supervisor Associate Professor Jacqueline Whalley for her guidance and patience shown to me throughout my research. My heart feeling thanks go to my third supervisor Professor Philip Sallis for his research attitude and guidance. Without them this thesis never be successful.

I am heartily thankful to Dr. Subana Shanmuganathan for her guidance at the beginning of my PhD and encouraged me to continue my research.

My special thanks to editors and IT manager of "Lankadeepa" newspaper providing me the opinion texts for the research.

I thank my wife Mala Nelly for her assistance and patience and taking care of my kids Sirithi and Akitha. I really appreciate the sacrifices you have made during my period of doctoral study.

I would express my deepest gratitude to my father M. R. Ratnayake and mother Nandawathie Ratnayake. I dedicate this thesis to them.

Finally, I would like to express my gratitude to my PhD colleagues Vijay and Michle and the PhD program administrator Karishma Bhat, for their moral support, love and understanding.

Chapter 1: Introduction

1.1 Sentiment Analysis

Sentiment analysis also referred to as opinion mining, can be described as a topic of investigation within the research and practice domain of text mining. This topic domain is founded in the disciplines of computer science (more specifically artificial intelligence and text processing), and computational linguistics. It is mainly concerned with interactions between computers and human (natural) languages, such as English, French, Japanese, and Hindi, to name a few (Manning & Schütze, 1999). Understanding (insofar as that term is used technically within the natural language processing research domain) and the generation of machines of language in human discourse not only involves spoken language but also the use of written scripts.

As a topic for investigation in its own right, Opinion Mining refers to the extraction of sentiments (ideas, concepts, views, and propositions) from usually brief and often informal expressions written in a natural language (the language of the writer and the reading audience). There are instances where the subject of an opinion text expresses the semantic and interpretable intention of the writer sufficiently in these expressions.

Investigating the dynamics of Opinion Mining computability, the methods, processes and analytical reliability characteristics is the work described in this thesis. Foremost is the quest for reliable representations of the semantic intention of the opinion writer. The thesis develops a description and analysis of this endeavour. A case study using the Sinhala Language is explored, and this is described in this thesis.

Opinion mining essentially consists of two methods, namely subjectivity and sentiment analysis, that are run in a sequence (Liu, 2010). Identifying the subjectivity of an opinionated sentence or clause within a sentence and then classifying the opinionated text as a positive or negative opinion are the two main steps of opinion mining. The former is known as subjectivity identification, and the latter is termed sentiment classification. Contemporary research in this area is largely concerned with analysing the opinions (either positive or negative) contained in texts, and there are major language related issues concerning processing texts for identifying and classifying the opinions embedded within the texts.

Sentiment analysis research began around year 2000 was initially focused on the English Language alone. Over the last decade and a half, there have been significant advances made resulting from the vast amount of research conducted in sentiment analysis in the English Language. As a result, there are many more accurate and advanced techniques available to analyse opinions in English than for any other language (Chandrasekaran & Vinodhini, 2012). Such techniques include fine-grained analysis (Niepert, Stuckenschmidt, & Strube, 2011) and contextual modelling (Vanzo, Croce, & Basili, 2014). More critically, thus so far for "morphologically rich" languages fewer attempts have been made to develop techniques for sentiment analysis (Khawaldeh, 2015, Franky & Veselovská, 2015). The term "morphologically rich" refers to languages that contain substantial grammatical information in words and sentences. This information includes semantic cues in words and the arrangement of words into syntactic units (Tsarfaty, et al., 2010). This research focuses on sentiment identification and classification in morphologically rich languages such as Sinhala.

1.2 Motivation

Most of the contemporary sentiment analysis techniques for English are based on statistical methods. These statistics based methods primarily concentrate on the statistical measures of words in given texts or sometimes on the co-occurrence of the words in the texts. The weakness of these statistics-based methods (in most cases, the frequency of words) is that they do not reflect the real meaning of an opinion in the classification (Jang & Shin, 2011). Additionally, the morphological impact of the words is ignored in these techniques. Applying such statistics-based methods for morphologically rich languages by incorporating pre-processing techniques, such as stemming, is unlikely to be successful as valuable information contained in words is likely to be lost in the stemming process. Therefore, the application of such techniques to morphologically rich languages would not attain good results (Jang & Shin, 2011).

Above all, there is one central reason for needing better techniques for opinion mining in non-English Languages. The introduction of Unicode encoding for non-English characters has led to the availability of documents typed electronically in non-English languages, such as Sinhala, Hindi, and Chinese. This encoding has, in turn, led to a rapid increase in web contents written in these languages. However, the software tools and techniques for analysing the web contents in these Unicode based languages for measuring the sentiment of the opinions are limited as are the research experiments conducted in this problem domain (Sharma, Nigam, & Jain, 2014). In fact, the needs for such techniques that can handle non-English Languages has never been

so acute. As the challenges in developing software tools for opinion mining especially, for non-English languages are significant the need for sentiment analysis tools is yet be met.

1.3 Scope

SINHALA is one of the morphologically rich languages that belong to Indo-Aryan family (Jain & Cardona, 2014). The language is an official language in Sri Lanka with more than 15 million speakers. Linguists' believe the Sinhala language developed independently in isolation from other languages in the Indo-Aryan family. Due to this isolation, the language shows distinct differences when compared with the other languages in the same family; Hindi, Marathi, Urdu, etc. Among the 18 vowels in Sinhala, two vowels are unique and not found in the so-called Indo-Aryan or Dravidian languages. For example, two vowels; ඇ (æ) and ඈ (æ:) are long and are only available in Sinhala as an identical phoneme. Though English has an æ sound, there is no separate character to represent the phoneme. These long vowels are used to emphasize some utterances. Moreover, morphologically, postpositions are used in Sinhala rather than prepositions. As an example, in English "be good" would be represented by "ඉහොඳට" where the root word "ඉහොඳ" (good) is inflected by the morpheme "ට". The word "ඉහොඳ" (good) is an adjective and "ඉහොඳට" is mostly an adverb. The sentiment of both words is positive but occurs in different grammatical contexts.

The Sinhala writing system is often based on a "Subject-Object-Verb" word order. However, in the spoken form of the language, this order would be modified or in some cases neglected to denote the pragmatic consideration such as emphasis. Due to this free ordering characteristic, the difference in spoken and literal Sinhala is highly significant. When expressing their views on a particular topic, Sinhala writers prefer to write the views in spoken Sinhala rather than the literal form.

In addition to the challenges explained above, the success of an approach to opinion mining or sentiment analysis primarily depends on the availability of lexical resources. The reason for applying statistical based techniques than resources based methods in most sentiment analysis work is the unavailability of lexical resources, such as WordNet[1], Subjective lexicons with polarity or valance[2] of words, and lists of positive/negative words. These lexical resources are

[1]WordNet is a thesaurus or a linguistic knowledge rich lexical resource.

[2] Negative or Positive attitude

required for sentiment analysis. Languages that do not have such resources can be categorised as “less-resourced languages” in the context of language processing. The following statistics illustrates the availability of lexical resources for world languages: of the 7,097 living languages in the World (Ethnologue, 2016) 2,296 (32.4%) are spoken in Asia, 2,139 (30.1%) in Africa and only 287 (4%) in Europe. Additionally, 60% of the world’s natural language speakers are in Asia, 26% in Europe and 13% in Africa. On the other hand, less than 2% of languages have WordNet type tools for language processing (Wordnets in the World, 2014). However, most of these other languages have a dictionary. There are 7,500 online dictionaries and glossaries in the world (Dictionaries Translation and Language Resources, 2000). Manual compilation of subjective lexicon type resources is the only way for enabling sentiment analysis for these less-resourced languages. Manual compilation of such resources for a given less-resourced language is a challenging task as it always consumes an enormous amount of time and labor (Chen & Skina, 2014). The literature survey conducted for this research reveals that most researchers use the WordNet type resource to construct a subjective lexicon which is an essential repository for any sentiment analysis (Velikovich, Goldensohn, Hannan, & McDonald, 2010) for both English and non-English languages.

Developing lexical resources using efficient and effective techniques for the languages, which lack lexical resources is challenging though such resources are useful not only for opinion mining but also for running any other text processing tasks in many of the non-English languages. Indeed, in this electronic era, the survival of languages that do not have a complete set of resources for electronic communication is seen as a critical issue (Peters & Picchi, 2014). Some of these languages are morphologically richer than the English Language and hence the sentiment classification algorithms applied to English are not applicable even if such resources existed.

With these major setbacks, this research aims to build a generic framework for Sentiment Analysis, which includes efficient and effective methods for the following tasks in morphologically rich languages:

i. Automatically generating subjective lexicons
ii. Sentiment classification

1.4 Research Questions

The objective of this research is to build a framework for sentiment analysis of text written in morphologically rich languages. With this aim, the study concentrates on answering the following research questions.

1. Automatically generating lexical resources using already existing dictionaries for morphologically rich languages.
    a. How can effective and efficient lexical resources be automatically generated?
    b. Can contemporary lexicon building techniques be adapted to morphologically rich languages?
    c. How can newly generated lexical resources in sentiment classification be evaluated?
2. Adaptation of Bayesian classification algorithms for sentiment classification.
    a. How can Bayesian sentiment classification algorithms for morphologically rich languages be evaluated?
    b. How can word level morphological features be applied to Bayesian sentiment classification in the context of morphologically rich languages?

1.5 Overview of Research Direction

The research strategy for this study was summarized as follows. Three lexical resources for sentiment classification were developed using different approaches. The resource compiling was initiated by applying a cross-linguistic method and then a second lexicon consisting of a positive-negative word list was generated using a novel technique. Thirdly, another lexical resource was generated using a graph-based approach.

As this study is the first attempt of sentiment classification on Sinhala, adaptability of contemporary text mining methods to Sinhala is investigated. In this investigation, the bag-of-word vector representation was applied while investigating feature selection methods.

Adjective and the adverbs are the main Part Of Speeches (POS) for English sentiment classifications (Benamara, Cesarano, & Reforgiato, 2007). The hypothesis of adjective and adverb are better lexical categories for Sinhala sentiment classification was tested with an identification of the linguistic features that impact on the sentiment classification such as intensifiers and flow shifters is carried out. The overall research plan depicted in Figure 1.5.

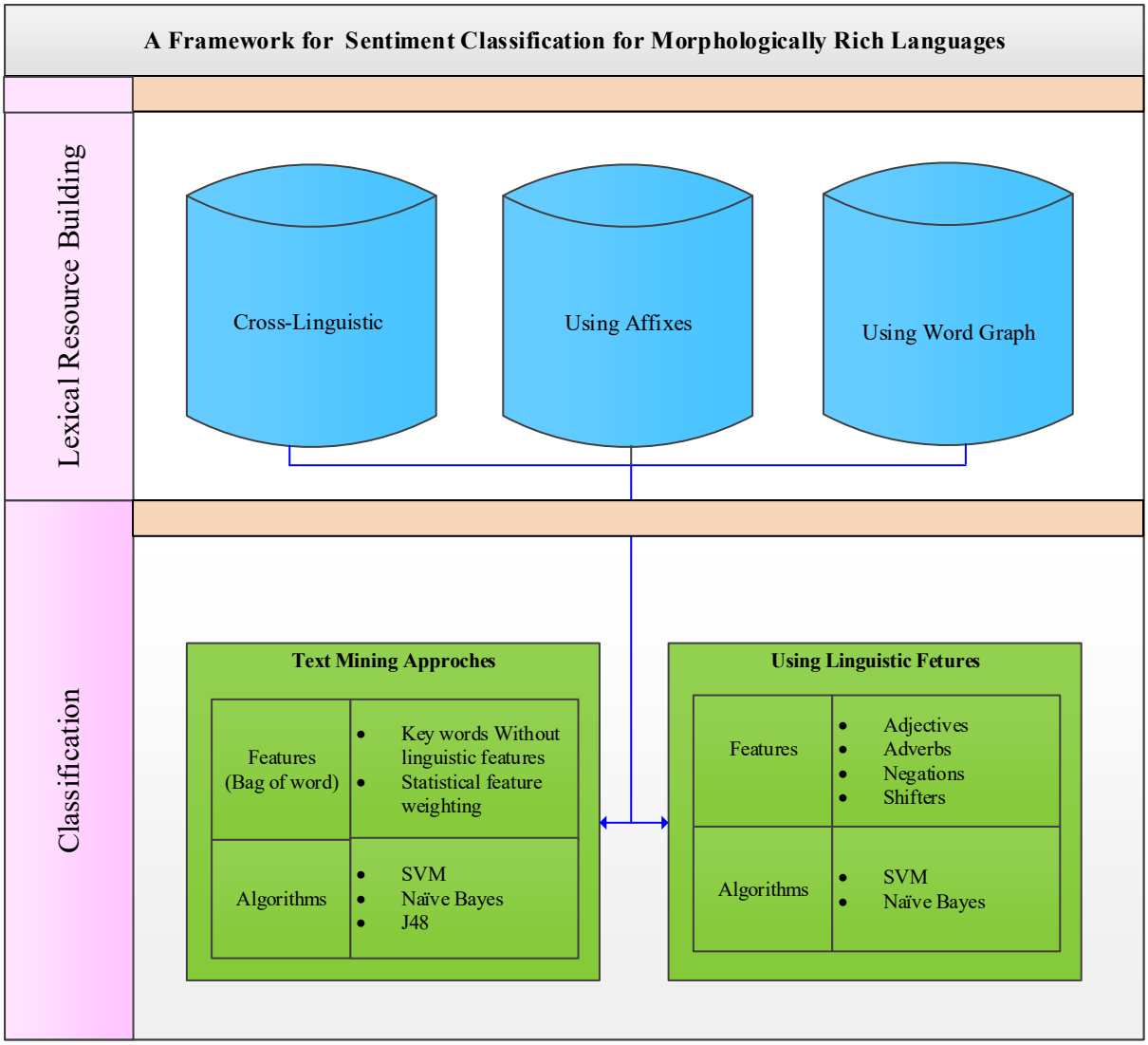


Figure 1.1: Research Plan

1.6 Publications

- Nishantha Medagoda, Sentiment Analysis on Morphologically Rich Languages: An Artificial Neural Network (ANN) Approach, Springer, February 2016
- Medagoda N, Shanmuganathan S, Whalley J, "Sentiment Lexicon Construction Using SentiWordNet 3.0", ICNC'15: 11th International Conference on Natural Computation, Zhangjiajie, China, August, 2015.
- Medagoda N, Shanmuganathan S, "Keywords Based Temporal Sentiment Analysis", FSKD'15: 12th International Conference on Fuzzy Systems and Knowledge Discovery, Zhangjiajie, China, August, 2015.
- Medagoda N, Shanmuganathan S, Whalley J, "A Comparative Analysis of Opinion Mining and Sentiment Classification in non-English Languages", 14th International Conference on Advances in ICT for Emerging Regions December 2013, Colombo.

1.7 The Thesis Organization

The following gives an outline of the chapters of this thesis:

Chapter 1 introduces the thesis and explains the motivation, scope, research questions, and research strategy and thesis organization.

Chapter 2 gives an overview and synthesis of the literature in the area of resources for sentiment analysis and is divided into several subparts. Chapter 2 begins by defining the terms related to the domain followed by contemporary methods for lexicon construction. The challenges in developing opinion mining resources for non-English languages is examined in detail.

Chapter 3 explanations of feature engineering, classification methods, and techniques for evaluation found in the literature.

Chapter 4 focuses on developing a framework for automatic sentiment analysis for this research. The chapter provides a discussion of the methods used for the automatic construction of lexical resources, and the theoretical concepts of sentiment classification and evaluation. An overview of the linguistic features of Sinhala are presented, and the possibility of using parts of speech (POS) is discussed.

Chapter 5 provides a detailed explanation and discussion of the construction of a Sinhala subjective lexicon using dictionaries. This discussion reflects on the major challenges related to the construction of a lexicon and provides details of its implementation (the algorithm used) and the results of the subsequent evaluation of subjective lexicons.

Chapter 6 gives domain independent and dependent aspect-based sentiment classification of Sinhala reviews at the document and sentence level using contemporary text mining approaches. An extensive explanation of feature selection is presented in the chapter followed by machine learning methods for classification.

Chapter 7 presents the adaptability of Sinhala linguistic features in sentiment analysis. It involves an exploration of linguistic features for sentiment classification with special attention on shifter features in addition to the adjectives and adverbs.

Chapter 8 provides a detailed explanation of the results achieved through this research. Conclusions are drawn up from this research and analysis, and possible future directions are detailed at the end.

Chapter 2: Resources for Sentiment Analysis

2.1 Introduction

In this chapter, a detailed study of theories and technologies related to building lexical resources from the literature reviewed for the research is presented. The chapter is divided into four sections. Section 2.2 explains the definitions of important terms relating to sentiment analysis. Literature findings on different lexicon resource[3] building techniques for some languages presented in section 2.3. Finally, the chapter summarised the methods of resource compilation in section 2.4.

2.2 Definitions: Sentiment Analysis

Most of the terms relating to this research area first appeared in the early 90's, for example, the phrase "Subjective Characters" first appeared in 1990 in a publication titled "Identifying Subjective Characters in Narrative" by Wiebe (1990). A subjective sentence was defined in the paper, as a sentence that presented the consciousness of the experiencing character within the topic being talked about. The words "sentiments" and "opinions" used in the vocabulary of this research domain were introduced in the early part of 21st century. These words appear in publications by several many authors. Das and Chen (2001) and Tong (2001), are among the researchers who first introduced the words in relation to sentiment analysis and opinion mining. Pang, Lee and Vaithyanathan (2002) explained the term "sentiment" as a way of labeling articles into positive and negative categories.

Although there is no specific definition for sentiment analysis, many authors have attempted to define this term in different ways based on the work or domain that they were interested in. Dictionaries define the terms in the context of linguistics theories. In the Oxford dictionary, sentiment analysis is defined as a noun:

"the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, or service is positive, negative, or neutral".

[3] Resources in opinion mining and sentiment analysis refer to lexical repositories such as polarity lexicon

While, in Collins dictionary, the term is defined as:

"The computational analysis of internet and social media posts on a given topic to determine whether they approve or disapprove of the topic".

More recently, Liu (2010) provided comprehensive definitions for the terms in the context of opinion mining and sentiment analysis in a chapter titled "Sentiment Analysis and Subjectivity" in the book entitled "Handbook of Natural Language Processing". According to Liu (2010) an opinion can be defined as a quintuple of five attributes, (e_i, a_{ij}, s_{ijkl}, h_k, t_l)

where e_i the object,
a_{ij} feature of the object,
s_{ijkl} sentiment of the opinion,
h_k holder of the opinion and finally
t_l time that the opinion was expressed.

The sentiment s_{ijkl} may be positive, negative or neutral, and expressed with different strength /intensity levels. With this definition, the objective of the process of sentiment analysis is discovering all quintuples.

2.2.1 Type of opinions

An opinion can be classified based on several aspects. In consideration of the complexity of an opinion, it can be classified as either a regular or comparative opinion. An alternative classification considered the granularity of the analysis and based on this aspect divides sentiment analysis into three levels of analysis: Document level, Sentence level, and Aspect level. The following sections of this chapter elaborate on the opinions and how they are classified using these two classification systems.

a. Regular and comparative opinions

A regular or simple opinion can be further divided into direct or indirect opinion (Liu, 2010). A direct opinion expresses the sentiment of the entity explicitly. "It is an excellent film" or "Sam Worthington did a terrible job on the acting" are direct opinions where the experience of the customer is expressed unambiguously. In an indirect opinion, the experience or the feeling of the customer about the entity is expressed implicitly. The interesting feature of the indirect opinion is that the aspect of the entity is compared or dependent on another entity. "That I would like to take my kids to see the show, and it would be a good way to introduce them to

live stage performances" is an example of indirect comments given for the film "Wizard of Oz." This strong positive sentiment or opinion describes the feature — the "stage performance" — indirectly in the comment about the entity — the film.

A comparative opinion is a sentence or comment that expresses a relationship based on similarities or differences of more than one object (Jindal & Liu, 2006). "Windows 10 is better than Windows 8.1" is a comparative opinion which compares the two products on the same feature. It is argued in that a comparative opinion is usually expressed using the comparative or superlative form of an adjective or adverb although this is not always correct (Jindal & Liu, 2006). As an example, the words such as "prefer' and "superior" express a comparison even though they are not comparative or superlative.

b. Level of sentiment analysis

Many researchers focus on Sentiment analysis at different granularities. As mentioned previously, there are three levels of investigation relating to sentiment classification in a given opinion; Document, Sentence, and Aspect level. In Document level classification the whole opinion is considered as a document and classified as either a positive or negative sentiment and in some cases a neutral opinion (Pang, Lee, & Vaithyanathan, 2002). At this level, it is assumed that the document contains opinions describing the sentiment on a single object.

On the other hand, Sentence level opinion classification investigates a sentence and determines whether it is positive, negative or neutral. A sentence can be a subjective or an objective sentence. If the aim of the sentence is to present factual information then it is considered to be an objective sentence, otherwise the sentence is subjective (Wiebe, Bruce, & O'Hara, 1999). Sentence level sentiment classification focuses more on subjective sentences. Therefore, it is also known as subjectivity classification.

Aspect level sentiment classification yields very fine-gained information out of the opinions (Schouten & Frasincar, 2016). Both document and sentence level analysis deals with the polarity alone and so do not target the object of the opinion. Aspect level analysis is deeply concentrated on the opinion and not on the language construct. This level of analysis is more difficult than the other two types, i.e., document and sentence level classification (Liu, 2010).

2.3 Lexicon construction

It is apparent that each word used to express an opinion in a language tends to convey a polarity, either positive, negative or neutral. For example, the word "good" is a positive word, and "bad" can be considered as a negative word as far as the polarity of the words is concerned. An opinion with positive words expresses the respondent's preference whereas the negative type of words are used to show undesirable states. Therefore, the polarities of words are highly influential in sentiment analysis. The polarity encoded for a word is qualitative in nature, and it could be positive, negative or neutral and often qualified with some score indicating the magnitude of the polarity.

This section discusses how to generate a list of words that disclose the sentiments. In the literature, a list of words with the polarity assigned to each word in the list is known as a subjective lexicon. Many methods have been proposed by various researchers as for how to construct subjective lexicons. Among these approaches, manual construction, dictionary based and Corpus based is common in the literature. A manual approach is the simplest form. However, it is a very time consuming and labor intensive process. The accuracy of the words collected can be improved using an automated method on a manually generated lexicon. These automated approaches are primarily based on machine learning techniques such as graph-based analysis using either a dictionary or a corpus as ancillary tools. In the dictionary based method, a compilation process is initiated using a small word list known as the seed list. Normally, this seed list consists of manually constructed adjectives and adverbs with their orientation (polarity). The seed list is then propagated through an online dictionary, such as WordNet[4] (Miller, 1995) to grow the list by adding new terms generated through searching for synonyms and antonyms of the seed list words. The weakness of subjective lexicons constructed in this manner is that they become domain specific in orientation.

The Corpus based method is an alternative approach proposed to overcome the domain specificity. In the dictionary based method, the seed list is searched through the corpus searching for any syntactic or co-occurrence patterns of each and every seed word. An additional adjective (adverb) of the seed word with its orientation is added to the list using a set of constraints or conventions on connectives. The majority of the rules or constraints are designed using the connectives "and", "or", "but", "either-or" and "neither-nor"

[4] A thesaurus or a linguistic knowledge rich lexical resource.

(Hatzivassiloglou & McKeown, 1997). These linguistic rules are called sentiment consistency. One of the limitations of this method is that not supporting to build a list that represents all the sentiment words in a language.

The following paragraphs present some of the work that has produced success in different languages. The section begins the discussion of the work in the English language. Before moving on to discuss the complexities related to the construction of subjective lexicons for morphologically rich languages — the primary interest of this thesis.

2.3.1 Subjective lexicon building for English using the dictionary based approach

Dictionary based methods begin with collecting a small set of known polarity words defined as a seed set, manually. This seed list consists of both positive and negative words. Then a task of search or propagation is conducted through an online dictionary on a synonym or antonym path, and the newly found words are then added to the seed list. Scores expressing the polarity of unknown words are calculated using counts or frequencies of incidence on the node. The advantage of this method is the extremely low cost of assigning polarity (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). The accuracy and validity of the dictionary based approach heavily depend on the comprehensiveness of the dictionary. Some seminal studies, from literature that use this method, are summarized as follows.

Hu and Liu (2004) used the dictionary based approach using a set of adjectives as the seed list. The initial seed list that consisted of 30 common adjectives was expanded by adding all the adjectives in the opinion word list and predicting the orientation as derived from WordNet. The adjectives that WordNet (Miller, 1995) could not recognize or could not predict the orientation of were discarded from the list. In this work, the aim was to predict the orientation of words as either positive or negative rather than assigning a polarity score to each word.

With an assumption that synonyms of positive words are mostly positive, and antonyms are mostly negative, Kim & Hovy, (2004) tested the approach of constructing subjective lexicon using WordNet. For adjectives, the respective list of words was generated using both synonyms and antonyms, but for adverbs, the authors considered only the antonyms. To assign the strength of sentiment, they utilised a Bayesian-based probabilistic method.

Esuli and Sebastiani, (2005) investigated a method for determining the orientation of a term based on the classification of its gloss. Given two seed lists, one positive and the other negative, a semi-supervised method was used to expand the two lists by traversing through the WordNet.

For this search, the lexical relations for synonyms and antonyms used to traverse the WordNet. Then a textual representation of a newly added term was generated by collating all the glosses (as found in a machine readable dictionary) of the term. A binary classification model was trained on this textual representation after converting the terms into a vector form by standard test indexing. The authors reported that this method outperformed all published methods. This work was subsequently extended to classify term objectivity in addition to subjectivity (Esuli & Sebastiani, 2006).

Williams and Anand, (2009) deployed a semantic distance calculation technique to establish the polarity of a word with the aid of WordNet and a reference list. In this study, scores were calculated only for adjectives using an adjective graph. The adjective graph was built recursively by querying the lexical relations defined in WordNet for the set of seed words and adding edges between the words resulting from reference words and query words. The semantic distance of a given adjective was calculated as the relative distance from two reference terms: "good" and "bad". The distance between any two nodes in the graph was defined by five different methods from a simple path length of chains of synonyms to complex antonym relations. Their evaluation showed that the proposed method extended coverage and achieved good accuracy using the lexical relations and similar words in addition to standard synonyms when tested using a holdout data set.

A semi-supervised polarity detection method was proposed by Rao and Ravichandran (2009) using three graph based algorithms: Minicut, Random Minicut and label propagation in WordNet. The authors reported that their algorithm was applicable to any language for which a WordNet type resource is available. They also suggested that a thesaurus with synonyms could be used in the absence of a WordNet. In the evaluation of the methods, it was found that label propagation produced significantly better results than baseline[5] and other semi-supervised learning methods such as Minicut and random Minicut.

Makki, Brooks, and Miles (2004) presented an aspect based automatic lexicon creation approach. They used an initial seed list with known polarities. In this approach, the user is engaged in the polarity assignment process. A new visualization framework was introduced that allows the user to assign polarities. In addition to the new term added to the list by an iteration process on the review corpus, the aspect of the term is also added. Noun and noun

[5] A minimum or starting point used for comparisons.

phrases are considered in order to extract the aspect of the sentiment word. The polarity of the word is predicted based on the evidence observed in the context. The context is the other sentiment word that modifies the same aspect in the same opinion.

In summary, the majority of the dictionary based methods rely on searching through an online dictionary, usually on WordNet. A manual cleaning process is essential for increasing the accuracy of the lexicon even though it is a labor intensive and time consuming task. Another significant feature of this method is that it is domain independent. Most of the lexicons constructed using this method are context independent, and this may affect the application of sentiment classification using the sentiment lexicon because sentiment words are often context dependent.

2.3.2 Corpus based approaches for English lexicon construction

Large collections of a structured or unstructured set of texts that are stored electronically are known as a corpus. Corpus linguistics is the study area of linguistics that uses corpuses. Sentiment analysis and opinion mining are also highly influenced by corpus linguistics, in which most of the corpuses contain opinions, reviews, and human feelings expressed in texts. To overcome the context independence of lexicon constructed using dictionaries, researchers have moved to building corpus based lexicons. An examination of past work reveals two main approaches to sentiment lexicon building using corpuses. Starting with a seed set and discovering additional sentiment words in a corpus is one method. This method is comprised of similar to steps to that of the dictionary based method. In the second approach, a general purpose lexicon is combined to build a new lexicon using a domain specific corpus. The following section outlines several key corpus-based studies identified in the literature.

An adjective sentiment lexicon constructed in early studies by Hatzivassiloglou and McKeown (1997) used a corpus to find additional sentiment words. The proposed method relied on the orientation of the words. The authors assumed the conjunctions between adjectives were the key indicator of semantic orientation. With this assumption, they initially extracted the conjunctions of adjectives from a 20-million-word corpus with some morphological relationships of different sematic orientations. The connectives AND, OR, EITHER-OR and NEITHER-NOR were used to extract the adjectives. Then a log-linear regression model was used to determine if each of two conjoined adjectives were of the same or a different orientation in a graph. Using clustering techniques, the adjectives were then separated into two subsets of different orientation. The set of higher frequencies was labeled as positive, and the remaining

set was labeled as negative. They reported that they achieved more than 96% agreement with human identification/classification of the adjectives orientation.

WordNet independent and web-based lexicon construction techniques for English was introduced by Velikovich et al. (2010). Related language units such as part of speech misspellings or multi-word expressions do not affect this lexicon. A graph propagation algorithm implemented in the method finally retrieved both positive and negative sentiment scores for each node. The phrase graph constructed for this study was generated using n-grams up to 10 extracted from 4 billion web pages. The polarity scores were calculated as the sum of the maximum weighted path from every seed word (either positive or negative) to the node.

A domain specific lexicon was constructed by Yang et al. (2014) using Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). The terms were classified as positive, negative or no sentiment comparing three posterior distribution of the word for a given sentimental topic. The paper concluded stating that the proposed algorithm was capable of learning new sentiment words and discovering the subtle sentimental meaning of the word.

Severyn and Moschitti (2015) proposed a distant supervision automatic lexicon generation model to construct a subjective lexicon using an unlabelled tweet corpus. Their process began by labeling the corpus entries as either positive or negative using the emoticon symbol assigned by the readers of each tweet. Then the unigram and bigram features were extracted to construct a feature vector. A support vector machine was trained in order to obtain the weighted components of the training examples either unigram or bigram. The weight of the model was used as the sentiment association score of the unigram.

2.3.3 Sentiment Lexicon construction for non-English languages

Even though opinion mining and sentiment analysis research was first initiated for the English language in the late 90's, today in 2016 more researchers are investigating sentiment classification for other languages. The main reason behind this emerging trend in research for non-English languages is the availability of electronic text for these languages. With the introduction of "Unicode"[6] representation, the amount information available in non-English languages has increased rapidly. This section describes research carried out to develop sentiment lexicons for some non-English languages. The author of this thesis paid more

[6] Standard for representing the characters of all the languages of the world, including Chinese, Japanese, and Korean

attention to the research conducted for morphologically rich languages with the intention of developing a more efficient approach for a morphologically rich language i.e. Sinhala, in his investigation.

Joshi et al. (2010) developed a lexical resource called Hindi-SentiWordNet (H-SWN) by manipulating two lexical resources, namely SentiWordNet (SWN) (Esuli & Sebastiani, 2006) and the English-Hindi WordNet. The difference between H-SWN and other sentiment lexicons is there is a polarity score attached to the senses of the word instead of the word in the sentiment lexicon. In this approach, the synset that corresponds to the English in Senti WordNet (SWN) was projected to a corresponding synset in English-Hindi WordNet to build the senti wordnet (H-SWN). Using this approach, the authors managed to create the H-SWN of 16,253 synsets which included adjectives, adverbs, nouns, and verbs. The assumption of the sentiment of a synset retained across English and Hindi is critical to the accuracy of the method. However, this approach, using the H-SWN of senses achieved a classification accuracy of only 60%.

Bakliwal, Arora & Vrma (2012) constructed another Hindi sentiment lexicon using a seed list of 45 adjectives and 75 adverbs. The adjective list included 15 positives, 15 negative and 15 objective adjectives. Similarly, the positive, negative and objective of 25 each adverb were also included in the adverb seed list. A breadth first search was performed to expand the seed list on a graph based WordNet where words were connected to each other to indicate their synonym and antonym relationships. A new word was appended to the list assigning the polarity of the word using an assumption that a synonym carries the same polarity and antonym shows the opposite polarity of the root word (the seed list word). Using this method, the authors managed to build a Hindi subjective lexicon with 8,048 adjectives and 888 adverbs. The new subjective lexicon was evaluated using two methods: human judgment and a simple classification on a pre-annotated product review data set. In the classification method, the adjectives and adverbs were first identified using a shallow parser, and then the weighting of the review dataset was calculated using a unigram to determine the positive, negative and objective polarities. The maximum count was used as the final score. The authors noted that this method achieved poor agreement and that this was probably due to the ambiguity of Hindi words. However, approximately an 80% accuracy rate was achieved using the same classification method and stemming the words that were found in the review data set but not available in the generated subjective lexicon.

Huang, Niu & Shi (2014) utilized chunk dependency knowledge to extract a domain-specific sentiment lexicon based on constrained label propagation. They divided the whole strategy into six steps. They first detected and extracted candidate domain-specific sentiment terms by combining the chunk dependency parsing knowledge and prior generic sentiment lexicon. To refine the sentiment terms some filtering and pruning operations were carried out. Then domain independent sentiment seeds were selected from the semi-structured domain reviews which had been designated manually or directly borrowed from other domains. As the third step, the semantic associations were calculated between sentiment terms based on their distribution contexts in the domain corpus. For this calculation, the point-wise mutual information[7] (PMI) was utilized which is commonly used in semantic linkage in information theory. Then some pairwise contextual and morphological constraints between sentiment terms were defined and extracted in order to enhance the associations. The conjunctions like "and" and "as well as" were considered to be direct contextual constraints whereas "but" was referred to as a reverse contextual constraint. The above constraints were propagated throughout the entire collection of candidate sentiment terms. Finally, the propagated constraints were incorporated into a label propagation for the construction of domain-specific sentiment lexicons. The proposed approach gave an accuracy increment of approximately 3% over the baseline methods such as chi-square based polarity determination and the PMI-IR (pointwise mutual information and information retrieval) based polarity determination.

Xu, Meng & Wang (2010) built a Chinese emotion lexicon using the graph based method and multiple resources. The graph algorithm ranked the words according to the seed words. Multiple resources were incorporated in calculating a ranking where a similarity matrix was calculated to rank the unlabelled words. The resources included were an unlabeled corpus, a synonym dictionary, and a semantic dictionary.

A graph-based application for constructing a sentiment lexicon has been explored for the Norwegian language. Two strategies were investigated to build a sentiment lexicon for Norwegian language (Hammer, Bai, Yazidi, & Engelstad, 2014). A directed graph known as word graph was built using three thesaurus defining words as nodes and synonym and antonym relationships as edges. Then the lexicon was generated using label propagation (Kim & Hovy, 2006) giving each positive and negative seed word a score of +1 or -1 respectively; all other

[7] Point mutual information, is a measure of association used in information theory and statistics.

nodes were given a score of 0. The algorithm was propagated to each of the nodes updating the weighted average score of neighboring nodes and the value of the sentiment score of the word. The lexicon was evaluated using the classification of manually annotated reviews. The authors found that lexicon based machine translation performed better than the graph based lexicon. Moreover, they commented that the linguistic resources in English could be easily adapted to Norwegian without losing significant value.

Badaro et al. (2014) compiled a large scale Arabic sentiment lexicon using existing resources, namely, English SentiWordNet, Arabic WordNet, Arabic morphological analyser and English WordNet. The two-step approach used in this lexicon construction consisted mainly of mapping from Arabic WordNet to English SentiWordNet and Arabic morphological analyser to Arabic WordNet. Then a linking of Arabic WordNet and English SentiWordNet using synset for the first mapping was carried out, and a sentiment score was then assigned to the Arabic words. The purpose of second mapping was to align in the lemmas in Arabic morphological analyser and their corresponding lemmas in WordNet. Evaluation of the lexicon was based on a review corpus, and they achieved an average F1 score (the weighted average of precision and recall) of 64.5% for their trinary (positive, negative, subjective) sentiment classification.

Ṕerez-Rosas et al. (2012) presented a framework to derive a sentiment lexicon for Spanish using manually and automatically annotated data and available electronic resources of resource rich languages, such as English. For the manual annotation they used the Opinion Finder lexicon (Wiebe & Riloff, 2005) and transferred the annotation into English WordNet by enforcing SentiWordNet based constraints. Finally, using the synset and Spanish WordNet which is aligned to English WordNet, located the corresponding translation of the word base on the sense key. In the evaluation of the lexicon generated using machine learning techniques the proposed method achieved 72% of the F1 score.

Chetviorkin and Loukachevitch (2012) proposed a set of statistical features and algorithms that combined could discriminate sentiment words in a specific domain to create a sentiment lexicon for the Russian language. The method was initiated by manually labeling the words in movie review data which had a frequency of greater than three. The authors trained the supervised machine learning algorithm; Logistic Regression, LogitBoost and Random Forest with two classes; sentiment and neutral words. Finally, they obtained a word list ordered based on predicted probability of their opinion orientation. An average of 81.5% precision was achieved for 1000 words by the classification. The authors evaluated this approach further steps by

applying the model to four other domains movies, books, mobile phones and digital cameras, and achieved precision measures greater than 62% for all domains.

Kaji and Kitsuregawa (2007) built a Japanese sentiment lexicon by collecting the polarity words in a corpus. Polar sentences in HTML documents were extracted by filtering the structural clues known as lexico-syntactic patterns to build a corpus. Two types of structural layouts were utilized in this study. Firstly, the itemization with headers that were indicative of pros or cons type of polar phrases and secondly a Table type; type A, if there were cue words in the left most column otherwise, it was type B table. The method was applied to one billion HTML documents after parsing the dependency trees for the sentences. The quality of the corpus was then evaluated against human classification. Two judges agreed with almost 93% of the polarity in 500 samples. Finally, phrases with adjectives were extracted to build the lexicon.

An attempt to build lexicons for all major languages was undertaken by Chen and Skiena (2014) using available resources a knowledge graph. The knowledge graph was constructed using seven million high frequency words in 136 languages. Each language contributed a hundred thousand most frequent words collected by the Polyglot project (Al-Rfou, Perozzi, & Skiena, 2013). Nodes were connected by semantic relations across the language by integrating several resources including machine translation, transliteration links, WordNet and Wiktionary (Wiktionary, n.d.). Sentiment propagation through the graph started with the English sentiment lexicon, and the sentiment polarities were extended to adjacent neighbors. Both label propagation and graph propagation were experimented with in the study. The authors evaluated the lexicons generated using available lexicons as the standard for a given language based on both accuracy and coverage. Most of the major languages achieved high accuracies, but coverage was not found to be acceptable.

In conclusion, it appears that most of the non-English sentiment lexicon generation reported in the literature has been based on cross-linguistic approaches. In essence, many of the studies have used a lexicon generated for the English language or translated the target language entries to English and then assigned the polarity scores of English terms regardless of the target language. Typically, a graph based algorithm is used when researchers used their own thesaurus or corpus of the target language. Very few researches have been conducted to use language specific features, such as morphology or syntactic information to build lexicons even though many non-English languages are linguistically rich.

2.4 Chapter Summary

The literature review presented in this chapter can be summarised as follows:

- Domain-independent sentiment lexicon construction relies on the availability of lexical resources. WordNet is the most commonly used resource for English. The collection of polarized lexicon entries is typically performed by propagating through a lexical resource using lexical relations. This technique is commonly known as a graph-based method.
- The probabilistic models, with some linguistic knowledge, have proved to be successful when building domain dependent subjective lexicons using corpuse.
- Cross-linguistic approaches are widely used in the construction of a subjective lexicon for non-English languages. The methods reported in the literature for cross-linguistics approaches are not limited to just syntactic parsing and graph-based methods but also included machine learning methods.

This chapter has presented a review of lexical resources, their construction and the role they play in sentiment classification. The next chapter presents the findings from a literature review of the methods used in sentiment classification

Chapter 3: Sentiment Classification

3.1 Introduction

The approaches presented in the literature for sentiment classification are discussed in this chapter. These approaches include a selection of features, sentiment classification methods and evaluation techniques. Features used in different studies and their effectiveness are elaborated on in section 3.2. Section 3.3 investigates sentiment classification algorithms and then gives a comparison of the most widely used methods. A detailed explanation of evaluation methodologies utilized in opinion mining and sentiment analysis is presented in 3.4. A summary is given in section 3.5.

3.2 Feature Engineering

The representation of data (i.e., record, document or sentence) in a certain format is essential for machine learning based classification. The representation is a vector with fixed dimensions. Each dimension of the vector represents a feature of the document or sentence. The accuracy of the classification relies on the mapping function from the document (sentence) to the vector. That is, an adequate as well as representative feature vector will achieve higher accuracies. In sentiment analysis, feature vectors may be constructed from simple bag-of-words methods to more complex linguistic features. This section presents literature related to feature vector construction for sentiment classification at the document and the sentence level.

3.2.1 Lexical features

Lexical features are word based features which explain the surface level characteristics of the document or sentence. A word itself and the different derivatives of the word; stem, prefix, infix, and suffix are lexical features extracted by shallow analysis. The stem is also known as the ‘root word’ which is the simplest form of a word such that the word cannot be broken down further. Lemmatization or stemming is the process of extracting the root from the derived word (Porter, 1997). A bag-of-words is a collection of words selected from a given corpus using a feature selection method which captures the lexical semantic. Simple selection approaches include methods such as highest frequent word list, a list of words by personnel judgments or advanced feature selection methods using correlation, information gain or mutual information. These methods are further discussed in chapter 5.

The n-gram method is one of the most common methods used for sentiment analysis in English. Moreover, it is often used as a benchmark method for evaluating novel methods.

There are many papers in the literature which explore the n-gram features for sentiment classification. A unigram feature represents a document character with a single word. However, in the case of text analysis, the word collocation that is a sequence of words is used as a feature and is defined as an n-gram. N-gram features have the advantage of capturing the compositional semantics. In addition, an n-gram is often used to predict the next word in language models. Consider the comment "I like this camera" the bigrams for this comment are; {I like}, {like this}, and {this camera} and the trigrams are {I like this}, and {like this camera}.

Arguably the most influential early piece of work for English using n-grams is that of Pang et al. (2002) in which the effects of unigrams and bigrams in sentiment classification were examined. They concluded that there was no significant improvement in using bigram features over unigram features. In a later seminal experiment involving a polarity classification of movie review data, an 87.2% accuracy was achieved using unigrams and a 79.5% accuracy using bigrams (Ng, Dasgupta, & Arifin, 2006). Ng et al. (2006) are also found that when all bigrams are used (not dropping any feature or not applying any feature selection algorithm) an 83.6% accuracy was obtained which is still a poorer accuracy than was achieved for unigrams. The authors suggested that the poor classification performance obtained when using bigrams was due to the sparseness of the data used. Ng et al. (2006) also reported that trigram classification produced worse results than that of bigrams. Contrary to Ng et al.'s (2006) study Dave et al. (2003) found that trigrams gave improved classification performance unigrams and that bigram classification was moderately better than unigram classification. Furthermore, Dave et al. (2003) noted that there was some degradation in the performance when lower order n-grams were included. In a study of subjectivity classification using a shallow approach by Raaijmakers and Kraaij (2008) the use of a character n-gram of a substring instead of a word n-gram was tested and was found to result in the better accuracy using super word character n-gram rather than a sub word n-gram (Raaijmakers & Kraaij, 2008).

Research to date on sentiment analysis using n-gram features in non-English languages has been largely limited to unigrams. Bakliwal et al. (2012) classified sentiments in Hindi, a morphologically rich language, using a unigram as the feature representation in the classification vector and obtained an overall accuracy of 77%. To date, the results obtained for

non-English language classification using n-grams has not given as good a results as those reported for the English language.

3.2.2 Knowledge Based Features

In sentiment analysis, the term "knowledge based features" refers to primarily the prior knowledge available in the lexical resources such as lexicons, dictionaries or thesaurus. Sentiment Lexicons, WordNet, and Positive and Negative word lists are important lexical resources that include knowledge base features. Liu (2010) integrated a sentiment lexicon to tag the sentiment word feature as "POS" for positive words and "NEG" for negative words along with parts of speech. Then, different polarized tags in a document are computed as the final features. However, no significant improvement in classification has been shown by the lexicon usage compared to baseline feature extraction method using simple bag-of-words. A similar approach tested by Ng et al. (2006) but tagging bigrams as positive or negative depending on whether or not the bigram included a positive or negative adjective respectively. Movie review data was then classified using a combination of the newly created features with bigrams. The results showed a significant improvement in classification using the polarity of adjectives. In a study of word level polarity features by Wieganda and Klakow (2009), the prior polarities of a word positive, negative and neutral were extracted from a lexicon with the level of strong and weak. The authors reported a 3.4% increase in accuracy over the simple bag-of–words method. In subsequent testing, they achieved a 77.5% accuracy by integrating other linguistic features; part of speech, main predicate, and the main clause.

WordNet is the lexical database (Miller, 1995) which is used in sentiment analysis widely. WordNet provides more general linguistic information; synonymy, antonym, etc. Earlier work by Dave et al. (2003) reported that the adding additional features, such as synonyms, through the use of parts of speech to the classification vector does not improve the accuracy. The authors stated that the inability to provide word sense disambiguation by WordNet as several meanings and many synsets for given the word was the reason for the lack of improvement. Furthermore, they pointed out that using WordNet causes feature sets to grow to an unmanageable size. However, in work by Wieganda and Klakow (2009) an improvement was shown when hypernyms were employed in a sentence level binary classification. Balamrali et al. (2011) undertook a comparison of a word based representation of documents with a sense based representation where WordNet senses of the words were used as features. The sense based method outperformed the word based classification, but it was found that manual sense annotation was better than WordNet sense annotation. Among the similarity metrics used in the

experiment, the best performance was given by the Lesk similarity metric (Benerjee & Pedersen, 2003) where each concept in WordNet was defined through gloss[8].

3.2.3 Linguistic features

Linguistic knowledge enhances sentiment classification accuracy. Knowledge of linguistic theories is important in sentiment polarity determination in word and sentence level. Stemming or Lemmatization, Part of speech tagging and morphological passing are processes of word level linguistic knowledge extraction. Dependency parsing and syntax extract the sentence level linguistic knowledge. The following paragraphs describe some linguistic approaches that are commonly used in natural language processing techniques.

Stemming and Lemmatization are the processes of extracting the root word from the derived or inflected word. The idea behind the application of stemming/ lemmatization is to generalize the words used in feature vectors across the opinion corpus. In lemmatization aiming to remove inflectional endings and return root form using morphological analysis of the word. On the other hand, stemming refers of chopping off the affixes of the derived words. Gamon (2004) proved the application of surface based features that included lemma unigrams, lemma bigram, and lemma trigrams gave better classification accuracies than linguistic features such as, part of speech tagging. However, Dave et al. (2003) obtained poor classification results when applying stemming using porter's stemmer algorithm. They concluded that the stemming over generalized the word forms. The word form is generally highly sensitive to certain linguistic features, for example, negative comments more frequently consist of past tenses that would be removed when stemming. Thus, stemming over generalised the word form to the extent that the sentiment was lost.

Part of Speech Tagging is widely used as linguistic knowledge in sentiment classification when developing machine learning methods. Words annotated with verb, noun, adjective, and adverb are extracted from the text before classification. Benamara et al. (2007) argued that combined use of adjectives and adverbs is better than using adjectives alone. Based on the three scoring algorithms; variable scoring, adjective priority scoring and adverb first scoring proposed approach gave higher precision and recall compared to existing methods that use only adjectives. Chesley et al. (2006) employed verbs and adjectives to classify blog sentiments automatically. Determining the polarity of the adjectives using their method an accuracy of

[8] An explanation about a word or phrase

90.9% was achieved and with verb classes 89.3% and 91.2% in defined class of approving, praising, doubting, or arguing. The POS (part of speech) tags of noun and adjectives were used in a model proposed by Yi et al. (2003) where the sentiment of given topic was extracted using natural language processing techniques. The authors reported a 93% accuracy was obtained in the topic of pharmaceutically related reviews while for other topics they achieved 90% or more. Na et al. (2004), to measure the effectiveness of linguistic processing in sentiment classification conducted a detailed comparison study. In the experiment, feature vectors generated by unigrams, POS tagging and selected words for the verb, adjective, and adverb were compared. The results showed that the highest accuracy was obtained by the classification that used selected words. The effect of POS tagging features was seen as not promising, but it was found to be better than simple unigrams.

The syntax is a deeper linguistic analysis which involves syntactic incorporation of a feature set. The main purpose of applying syntactic relationships in sentiment analysis is to capture the compositional semantics. The argument in favour of using a syntactic feature set presented by Ng et al. (2006) was that frequency based approaches such as n-gram models, deal the local dependencies, but the syntactic treatments help to identify the global dependencies. In this approach, complex linguistic constructs are described in order to capture the sentiment of a sentence based on the syntactic structure used. Kudo and Matsumoto (2004) applied a subtree based boosting algorithm based on a dependency tree to generate two sentence level classification tasks; sentiment polarity classification and modality identification ("opinion", "assertion" or "description"). The word based dependency tree approach outperformed the bag-of-words method, but the difference between the two methods was stated as insignificance of n-gram. Gamon (2004) compared the surface features with complex features extracted using a parser whose outputs included; part of speech trigrams, a constituent structure in the form of context free phrase structure patterns for each constituent in a parse tree, transitivity of a predicate, tense information. The evaluation of the proposed method using 200 feedback (opinions) from four categories revealed that the influence of abstract linguistic features in sentiment classification was minimal. In another study by Arora et al. (2010), the authors used a subgraph mining algorithm to automatically derive features as frequent subgraphs from the annotation graph for sentiment classification. The results indicated that there was no significant improvement over the unigram approach.

Shifter features are important factors in investigating the contextual structure of a sentence. Negation and intensifiers change the polarity of the sentence. Negation shifters invert the

polarity of the sentence and intensifiers either increase or decrease the polarity. Some past research efforts revealed that capturing the effect of shifter features can increase classification accuracies.

The most widely investigated shifter feature is that of negation. Das and Chen (2001) initiated the investigation of negation in a sentence in their pre-processing stage. In this phase, they tagged all the subsequent words with a negation marker (suffixed) for all the words in the sentence after the common negation words, such as not, never and no, were detected with the help of a dictionary. The authors have not stated the effectiveness of negation in a classification other than the detecting the negation word by proposed negation tagging. However, Pang et al. (2002) argued that the scope of negation cannot be properly modeled by this method. Also, they proved that the improvement in classification obtained by adding this artificial shifter feature over a simple bag-of-word method without negation is negligible. However, the advantage of using this negation shifter based feature method is that a plain occurrence and the negated occurrence of a word are clearly distinguished.

Polanyi and Zaenen (2004) modeled negation using a knowledge of polarity expressions such that a positive score is assigned to a positive polar expression and vice versa. In other words, combining positively velancy words with a negation such as “not” flips the positive velance to a negative velance. However, this model was not evaluated. Hence, cannot comment on the effectiveness of the approach. Despite this, Kennedy and Inkpen (2005) evaluated a similar model for document level polarity classification. The authors claimed that the positive effective of adding valence shifters for classification is statistically significant as 1.7% accuracy increased. No conclusions were made about the effectiveness of using the negation only approach.

Wilson et al. (2005) have proposed a more advanced model. In this model, a feature check was made to see whether a negation expression occurs in a fixed window of four words preceding the polar expression. In addition to the direct negative features, they added two types of other negative related features; shifter features and polarity modification features. Shifter features were added for checking the different types of polarity word, such as “little.” The polar expression of a particular type modifies the processing polarity expression, and it is defined as polarity modification features. Their results showed that actual negation features are more effective than the other two; shifter and polarity modification. A new modeling method known as “Scope modeling” is another prominent method of detecting the negation of a sentence. Jia

et al. (2009) used three parameters to detect the negation; static delimiters, dynamic delimiters and heuristic rules focused on polar expressions. Words such as “because” and “unless” present at the beginning of the next phrase are static delimiters. Dynamic delimiters are “like” and “for”, in classification, these are required disambiguation rules using contextual information. In the sentence level examination, the proposed model is compared with a simple negation in a fixed window size of text span of the sentence until the first occurrence of a polar expression following the negation word and in the entire sentence. The evaluation of the model found that linguistic insights for the negation modeling are effective.

Different, yet important, is the negation impact detecting methods applied in non-English languages. Despite the fact that there may be significant structural differences among different languages, the effectiveness of sentiment negation is crucial in polarity detection. The usual way of handling the negation, by reversing the words in a given fixed window forward and backward, was also adopted in a study that investigated the negations in Hindi. Mittal et al. (2013) carried out the study and their approach incorporated three rules. In the first rule, the words before the special negate word reversed the polarity of the sentiment word that followed the negated word. The forward negation is applied if the conjunction and the negation word appear in the sentence given that the index of the conjunction is more than the index of negated word. However, there was no mention of the meaning of the index. If the negated word appears multiple times in sub sentence separated by commas, then negation was applied in the forward direction until a delimiter was encountered in the third rule. The performance decreased for positive sentiments but significantly improved for negative sentiments.

Two types of negation words, natural and functional, were identified in Chinese sentiment analysis (Wu & Oard, 2009). In the first approach named “1-word dependency,” the word immediately following natural negation word was negated. In the second approach referred to as “2-word dependency”, the two words immediately following the functional negation word were negated. The final approach used employed syntactic dependency in order to establish the scope of negation. Even though no evidence of the effectiveness of the negation approaches in sentiment classification was discussed in the publication, the authors concluded that they had achieved a modest overall improvement over the best reported results in the literature. In sentiment analysis research of the Macedonian language, Jovanoski et al. (2015) used predefined a set of negative phrases and words to signal the negation. The special token was annotated to the words in the sentence until a clause level punctuation mark was encountered.

Experimental results showed that adding negative tokens gave a high negative impact when compared to the baseline model using bag-of-words after filtering stop words.

3.3 Sentiment Classification

In this section, the current directions being considered for this research to classify sentiments within the text in English and other languages are elaborated. The objective of the task is to investigate the recent classification techniques and evaluation methods in detail. The definition of an opinion given by Liu (2010), is a quintuple of five attributes (Chapter 2, section 2.2). With this definition the objective of sentiment analysis consists of six tasks as follows:

a. Extracting all entity expressions and group them into clusters. Each cluster describes a unique entity.
b. Similar to task a, but the aim is to extract all aspect related expressions and cluster them. Each cluster represents a unique aspect.
c. Extracting all opinion holders and categorizing them.
d. Investigating the time frame when the opinion was expressed and standardizing the time frame.
e. Aspect sentiment classification that aims to determine the opinion on aspect is positive, negative or neutral. Assigning a rating for the aspect is a part of this task. This is the significant task of this research as it permits the use of special features in the morphologically rich language.
f. Producing all five attributes expressed in an opinion based on the above tasks. According to Liu (2010), this step is a simple task.

A key step in the aspect based sentiment analysis is identifying the sentence or expression that contains the aspect. The sentence that explains the aspect is known as a subjective sentence, while an objective sentence may provide some factual information about the entity. A subjective sentence expresses some emotion, which might be personal feelings, beliefs, views, and thoughts. Emotions are closely related to sentiments, and the intensity of emotion is correlated to the strength of the sentiment (Liu, 2010).

The problem of subjective sentiment analysis is divided into two types depending on the classification objective. If the aim is to categorize the subjective sentence into either positive, negative or neutral, then it is a classification problem. Most of the research reported in the literature deals only two class (binary) categorization; positive and negative and ignores the

neutral class. If the objective is assigned a numeric value or ordinal value within a given range, then it is a problem of regression modeling. Both formulations on document level classification can be carried out by supervised and unsupervised learning. Initial attempts to use a supervised approach is presented in Pang et al. (2002), and Turney (2002) reports the initial use of an unsupervised method. Most of the supervised classification methodologies use machine learning techniques.

Traditionally, sentiment classification is considered as text classification. Text mining is "an interdisciplinary field bringing together techniques from data mining, linguistics, machine learning, information retrieval, pattern recognition, statistics, databases, and visualization to address the issue of quickly extracting information from large databases" (Zanasi, 2007) . The understanding of a given language relates not only to the spoken language but to written scripts as well. Text mining is more suited to the written text of documents including the textual information about, facts and opinions. Therefore, it is essential to conduct language specific pre-processing task prior to classification. Text cleaning, normalization, stop word removal, lemmatization, and morphological analysis are some pre-processing tasks generally employed in sentiment classification. These tasks are explained in detail in following chapter 4.

3.3.1 Supervised Sentiment classification

The supervised classification algorithm is one of the learning algorithms most frequently used in text classification systems (Kobayasi, Inui, & Matsumoto, 2007). In supervised classification, three sets of opinions are required namely, training, validation and testing data sets. The training data set is used to train the classifier to learn the variation of the characteristics of the sentence or document and the test data is used to measure the performance of the classification algorithm. Since sentiment analysis is a classification problem, researchers more often tend to apply supervised learning when the training data is made available. Among the supervised techniques, Naïve Bayes and Support Vector Machines (SVM) are widely used by the current research communities, and they have been proven the most successful in sentiment classification (Vinodhini & Chandrasekaran, 2012). Naïve Bayes algorithm is the most widely used, and it is a simple but effective supervised classification method (Xia, Zond, & Li, 2011). The basic idea of the method is to estimate the probabilities of sentiment (either positive or negative) for the given opinion using the joint probabilities of a set of words in a given category. The method is dependent on the naïve assumption of word independence. SVM machine has been reported to be the best binary classification method (Xia, Zond, & Li, 2011). SVM is a non-probabilistic classification technique that looks for a hyperplane with the maximum margin

between positive and negative examples of the training opinions. An alternative approach is k-Nearest Neighbor classification (kNN) a method that is based on the assumption that the classification of an instance is most similar to the classification of other instances that are nearby in the vector space. In comparison to the other text classification methods, such as Naïve Bayes, kNN does not rely on prior probabilities and is computationally efficient (Liao & Vemuri, 2002). However, Naïve Bayes is more efficient than other supervised classification techniques as it can be trained in a single pass through the training data.

In supervised learning, the input to the algorithm is a vector of some features. These features are characteristic of the opinion or document. In text classification, generally a feature can be a single word known as a unigram or set of words named as n-grams. Also, a feature will be a language specific character, such as a part of speech. In the classification vector, a feature is represented either by quantitative or qualitative measurements.

Pang et al. (2002) were the first to apply machine learning techniques to sentiment classification. Naïve Bayes, Maximum Entropy, and SVM algorithms were tested on a movie review data set in order to classify a sentiment with positive or negative polarity. No language specific pre-processing such as lemmatization or stop word removal was carried out. However, the negation was handled by adding the word "Not" for the words between the negation word and the first punctuation mark (Das & Chen, 2001). Unigrams and bigrams were used as features, and a standard bag-of-word method applied with unigram frequency greater than four and bigram frequency greater than seven were selected for the feature vector. Feature frequency that is a count of unigrams and bigrams in the document was calculated in the first run of the experiment. In addition to the frequencies of the features, the presence of the feature in a document was considered in the vector. SVM was reported to outperform the Naive Bayes algorithm, and the presence of unigrams was among the most effective features.

The work undertaken by the Dave et al. (2003) was similar to the work by Pang et al. (2002) but for a product review data set, and Dave et al. (2003) used different weighting calculations for the unigrams and bigrams. In addition to the frequencies, they calculated tf-idf (term frequency-inverse document frequency) weights for the features with smoothing scores. The linguistic modification of stemming and negation tagging was tested, and authors claimed that there were significance improvements. In conclusion, they found that both unigram and bigram gave better performance for both Naïve Bayes and SVM classification than the simple bag-of-words feature modeling approach.

A supervised learning method using the semantic orientations calculated by pointwise mutual information (PMI) for phrases was undertaken by Mullen and Nigel (2004). These phrases known as value phrases were extracted from review data using a combination of part of speech as defined by Turney (2002). The semantic orientation was used as a real number measure of the positive or negative sentiment expressed by a word or phrase. In addition, two more feature types called semantic orientation features were derived one based on the WordNet and the other on the emotive content of the text. Classification using SVM and a combination of semantic orientation and WordNet features gave the highest accuracy.

Document level supervised sentiment classification algorithm is another type. In this type initially, a differentiate polarity shifting sentence was proposed by Li et al. (2010). The sentence with the top ranked sentiment words is classified as being polarity non-shifted while sentences taking opposite polarity when compared with those sentences containing trigger words are deemed as polarity shifted. Using the above classification criteria, the algorithm automatically generates the two training data sets; polarity shifted and polarity non-shifted. Then they trained the two classifiers for each training data set in an attempt to detect the polarity of the sentence. In addition, a third classifier was derived by combining the two datasets. The combination mechanism used voting and stacking rules. SVM was used for all classifiers, and it was shown that there was an overall improvement on accuracy using the polarity shifting approach over the baseline which applies SVM with all unigrams and bigrams.

A deep neural network based sentiment classification approach using higher order phrases (n-gram) was adopted by Bespalov et al. (2011). The latent semantic indexing (LSI) in n-gram phrases was used as features that selected from a term document matrix[9]. The document represented by the supervised n-gram embedding was fed into a multi layer perceptron classifier to learn a function mapping towards sentiment labels. In addition to the supervised latent n-gram analysis by the neural network, an SVM classifier was trained to compare the performance of both classifications. In classification error based evaluation, the latent n-gram modeling gave the lowest error using data from Amazon. However, for the classification by only perceptions, the 2-gram bag-of-words produced the lowest error model.

[9]Term-document is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.

Fernandez et al. (2014) employed a novel feature selection technique for supervised sentiment classification to classify Twitter (Twitter is a social network) data. The authors used so-called "skipgrams" for the terms obtained after pre-processing and tokenizing the Twitters. The skipgrams are n-grams with skipping of terms for a given window size. A skipgram is described by two parameters n and k. Where n determines the maximum number of terms and k is the number of words skipped. The features in the SVM classification algorithm are the skipgrams weighted by relative frequencies of the term to a total number of terms and skipgrams. The evaluation results indicated that the proposed approach using skipgrams slightly improved on the accuracies achieved using unigrams. The highest accuracy was found when there was no restriction imposed by the skip parameter (k).

Joshi et al. (2010) used a SVM classifier to determine the polarity of an opinion in the first approach which trained on annotated Hindi sentiment corpus. In their second approach, which they called in-language sentiment analysis a Google translation, a machine translation (MT) based method, was used to translate the corpus in Hindi to English. The translated corpus was then inputted into a classifier. In the third approach, a resource based method, the synset corresponding to the English in Senti Word Net (SWN) was projected to the corresponding synset in Hindi to build the set wordnet (H-SWN) for Hindi. Classification under the resource based method was conducted using different structural features, such as changing the n-grams, with stemming, and without stemming. It was stated in the paper that the poor performance of the MT based approach caused by translation errors. In addition to this limitation, the research was based on two key and somewhat flawed assumptions. The first assumption was that the sentiment of a synset is retained across English and Hindi, and obviously, this was significantly critical to the accuracy of the method. The second assumption was that the sentiment of a document was preserved in the translation process this preservation of document sentiment is also crucial to the success of the algorithm. In conclusion, it was highlighted that an annotated corpus was an essential resource for sentiment analysis in languages, such as Hindi.

Yussupova and Bogdanova (2012) used a machine learning approach in their study of sentiment analysis in Russian text. The goal of their research was to discover how lemmatization affected the accuracy of sentiment classification. In this research, the "Bagging algorithm" was integrated into a Naïve Bayes classifier to improve the accuracy of the classification. The training and evaluation of the developed algorithm was carried out using reviews of Russian bank loans. One of the drawbacks of the study was the unbalanced sample used. The authors

analysed only 304 positive reviews but 850 negative reviews. Moreover, valuable information may have been lost in the lemmatization of the keywords.

In another Russian study of texts in Russian, the classification of opinions was attempted with two, three and five classes. The main aim of the study was to find a language independent approach to classifying opinions (Pak & Paroubek, 2011). They used an SVM classifier, which was totally dependent on feature based attributes such as, n-gram, pos-tags, and dependency parsing. In addition to n-grams, the authors proposed a new feature, which was similar to n-gram called d-grams. "d-grams" are constructed from a dependency parser tree, where words are linked by syntactic relations. In order to avoid domain adoption in the classification, the proposed system was tested on all combined reviews, i.e., Books, Movies, and Cameras as a product. The results revealed that the proposed system was the most accurate out of all combinations of options when classifying all the reviews together. The developed algorithm was also run on unseen data in different tracks. Tracks are defined by varying mode (number of classes), features, weights, and training sets. A 2-class track consists of 6 systems of binary class with different d-grams and weights. 3rd and 4th tracks are multiclass and will have different training sets. Based on the performances achieved, in the 3-class track, the experiment with movie reviews showed the highest accuracy while in 4th class track reviews on cameras achieved the best accuracy.

Interestingly, in "A Morpheme based Method to Chinese Sentence Level Sentiment Classification" by Wang et al. (2010) the morphological variations of a set of sentiment bearing words were integrated into the classification algorithm by extracting the morpheme and then inferring the semantic orientation of the words. These morphemes were of two types namely, positive and negative. According to the authors, the Chinese sentiment words can be categorized into static and dynamic sentiment words. These static and dynamic sentiment words contain a key morpheme that determines their emotional tendency. The morphological productivity of positive and negative morphemes contained in words in the Chinese lexicon used was calculated before determining the polarity of a review. Then the opinionated sentence was first segmented into four types of sentiment phrases. Using the morpheme productivity score, the average polar intensity of the review was estimated to decide the semantic orientation. A set of predefined thresholds was used to determine the semantic orientation i.e. whether the given opinion was positive, negative or neutral. Rules were included in the method to establish these thresholds, but there was no justification for the rules or the subsequent threshold values given. The proposed system was tested at different levels of linguistic

granularity namely at morpheme level, word level, and phrase level. The results presented in the paper showed that the phrase level, classification outperformed the other levels according to the F-values. The authors also compared the proposed system with some other morpheme based systems for Chinese languages and concluded that their proposed system outperformed the others even though their classification methods F-score was slightly worse. In Wang et al.'s method, the complexity of classification was very high when compared with the other methods, which explain the slightly lower F-score.

The study entitled "Chinese Sentence–Level Sentiment Classification Based on Fuzzy Sets" by Fu and Wang (2010) was aimed at comparing the Chinese sentiment analysis studies. In this study, as in the previous paper, the sentiment morphemes were extracted from a sentiment lexicon and then an opinion score was calculated using chi-square techniques. The word and phrase level polarities were then identified using a set of rules for each level. The word level polarity was determined by a key morpheme contained in either static or dynamic polar words. Then the final sentiment intensities of an opinionated sentence were calculated by summing the opinion scores of all phrases within the sentence. To handle the intrinsic fuzziness in sentiment polarity such as "positive", "neutral" and "negative," the authors applied a fuzzy set theory to sentiment classification. The fuzzy sets for each category of positive, neutral and negative sentiments were defined by three different membership functions based on semi-trapezoid distribution. The upward rise in the semi-trapezoid distribution for the three cases with different parameters was used to determine the polarity by maximizing the membership. The proposed method was carried out in three modules namely, lexicon analysis module, subjectivity detection, and sentiment classification. A sentiment density based Naïve Bayesian classifier was also embedded into the second module to perform the opinion detection in the sentence. The opinions saved in a standard Chinese opinion corpus were tested in the experiment. Eight hundred and forty-three documents with 62% of opinion sentences were included in the test data set. The phrases analysis outperformed the analysis at the other two levels of granularity studied, i.e., morpheme and word. In the comparison of the best system for Chinese opinions, the proposed system gave a higher F-score, and it was concluded that the fuzzy based system was the best model.

3.3.2 Unsupervised Sentiment classification

The objective of the unsupervised sentiment classification is grouping the opinions into clusters without providing any training sample data. Turney (2002) introduced unsupervised sentiment classification in his work by extracting syntactic patterns that expressed the sentiment in an

opinion. The syntactic patterns consisted of some predefined combination of part of speeches mainly adjectives, adverbs, noun phrases, and verbs. Then the semantic orientation of extracted phrase was calculated using the PMI-IR algorithm. The algorithm calculated the semantic orientation using pointwise mutual information for given two words (Church & Hanks, 1990) and this is a measure of statistical dependence of the two words. In this study, the pointwise mutual information was calculated for the phrases using two reference words, such as "excellent" and "poor." These words were selected with the aim of considering the semantic orientation of the phrase, for instance, a phrase was considered to be positive if it was more associated with "excellent" and considered to be negative if it was more associated with "poor." In the final step the average semantic score of all phrases in a review was calculated, and if the average was positive, then the algorithm classified the review as "recommended" otherwise "not recommended." The evaluation of the proposed method with different types of data sets revealed an average of 74% accuracy in most of the domains except for movie reviews data.

Hu and Liu (2004) applied an unsupervised sentiment classification method by initially identifying some product features on product reviews. The product features were identified using noun phrases after POS tagging. The adjectives identified by the POS tagging were considered as the opinion words. Then the semantic orientation of the opinion words was calculated using WordNet utilizing synonyms and antonyms of the adjectives. The opinion sentences identified were those containing one or more product feature/s and one or more opinion words. Finally predicting the orientation of an opinion sentence determined by the dominant orientation of the opinion words. That is, if more positive dominant words are present in the sentence, then it is regarded as positive. In the case of the same number of positive and negative opinion words, it is predicted by an average orientation of the effective opinions. The evaluation results showed that the method as effective with an average of 84% accuracy obtained for five product domains.

Thelwall et al. (2010) devised a new unsupervised algorithm named "SentiStrength" to detect the sentiment strength in short informal texts. The informal texts are informal messages posted in social media, and the algorithm was tested on the messages from Myspace. The algorithm used was based on several de-facto grammars and spelling styles of cyberspace. The core of the algorithm was the sentiment word strength list, which contains a positive and negative word list with sentiment scores based on a scale of 2 to 5. In addition to the above list, a booster word list, a negating word list and an emoticon list were used to detect the sentiment strength of a message. The algorithm was capable of running spelling correction and removing repetitions

and punctuations. The evaluation results revealed that the proposed algorithm performed well in positive sentiment detection than the negative sentiments. The accuracy for the positive detection was better than the baseline accuracies achieved by the supervised methods.

Unsupervised dependency parsing sentiment analysis work carried out by Gavilanes et al. (2014) used a sentiment lexicon. The lemmatized, and POS tagged Twitter comments were fed into a parser that outputted a full parser tree for a message. Then the tree was converted into the dependencies and functions of the phrases identified. The polarity of a message or tweet was determined by a real number which was calculated by the polarities of the lexical entries in the sentiment lexicon and their dependencies. Special factors such as negation, intensification, and polarity conflict were also taken into account when calculating the polarity value. The proposed method achieved an overall 59% in F-measure and positive message classification was found to outperform negative and neutral opinion classification.

An unsupervised model using common sense and context information was developed by Agarwal et al. (2015) to predict domain specific features of review documents. They used an ontology based ConceptNet to construct a domain specific ontology, especially for product reviews. To increase the coverage of the ontology for the product features the WordNet was combined with the ConceptNet. Then the sentiment phrases were extracted from the product reviews using a dependency parser, and the orientation of the phrases corresponding to the entries extracted from ConceptNet was calculated using a sentiment lexicon. The final orientation of the document was determined by aggregating the score for the phrases. The results gave an overall improvement over the baseline accuracy shown in all four different experiments, and they are with domain specific ontology, in consideration of the importance of the features, with contextual information, and the combination of the last two. In this work, the new approach of ConceptNet was introduced to the sentiment analysis.

3.4 Evaluation Methodologies

The evaluation of classification algorithms especially, in the case of supervised experimental design regarding machine learning approaches is essential. Calculating the performance measures is the way that the solution to the classification problem is evaluated. Several performance measures are used to evaluate sentiment classification. Of these the most frequently reported measures in contemporary studies are; classification accuracy that is a percentage of total correctly classified instances to total predictions and F-measure which is based on the confusion matrix.

In general, text categorization algorithms are evaluated using Precision, Recall, and F-measures in addition to simple classification accuracy. These standard measures have a significantly higher correlation with human judgments (Manning & Schütze, 1999). These are first defined for the simple case where a text categorization system returns the categories.

Precision (P) is the fraction of classified documents that are relevant

$$Precision = \frac{\#(relevent\ items\ retrived)}{\#(\text{classified } items)} = P(relevant|retrived)$$

Recall (R) is the fraction of relevant documents that are retrieved

$$Recall = \frac{\#(relevent\ items\ retrived)}{\#(relevant\ items)} = P(retrived|relevant)$$

These notions can be made clear by examining the following contingency table;

Table 3.1: Precision and Recall Contingency table

	Relevant	Non-relevant
Retrieved	true positives (tp) (t_p)	false positives (f_p)
Not retrieved	false negatives (f_n)	true negatives (t_n)

Then;

$$P = \frac{tp}{(tp + fp)}$$

$$R = \frac{tp}{(tp + fn)}$$

The measures of Precision and Recall concentrate on the evaluation of the return of true positives, giving what percentage of the relevant documents has been classified correctly and how many false positives have also been returned. A single measure that trades off Precision versus Recall is the F-measure, which is the weighted harmonic mean of precision and recall. F-measure is a measure of a test's accuracy. There are different weights that can be calculated for F-measure. The balance F-measure equally weights precision and recall, and it is commonly written as F1

$$F_1 = \frac{2PR}{(P + R)}$$

Then, the F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

Most supervised sentiment classification studies have used the classification accuracy as the performance evaluating measure (Pang et al., 2002, Makki et al. 2014). But the experiments carried out in these studies used different cross validation settings. N-fold cross validation is typically used to provide an estimate for the mean performance of an algorithm on a held out test set. In sentiment classification studies the number of folds in cross validation tends to vary in the range of 3 to 10. On the other hand, the classification accuracies calculated as the performance measure in unsupervised approaches as well. In some cases, the correlation between manually classified cases (either positive or negative) and same returned by the algorithm is examined to prove the success of the method (Turney, 2002). Comparative studies are often carried out in two different domains (i.e. Movie, and product reviews) to evaluate the performance of a proposed classification method. In such cases, it is very difficult to conclude the best choice of a performance measure for sentiment analysis as each study uses different training and testing data, and different features and classification algorithms.

3.5 Chapter Summary

The above literature survey can be summarised as follows:

- Unigram is a proven lexical feature and is often used as a baseline measure. Better performance has been observed in higher order n-grams than for unigrams. Incorporating knowledge based features using a subjective lexicon gives mixed results. Adjectives and Adverbs are seemed to be the dominating linguistic features in sentiment classification. No significant achievement has been obtained in applying complex linguistic features, such as syntactic parsing but the influence of negative words and the scope of negation is important in polarity classification.
- Naïve Bayes and Support Vector Machines are the most promising supervised classification algorithms, and they have been widely tested both for English and non-English languages. Morpheme based approaches that use word morphemes instead of unigrams has been found to improve classification accuracies significantly in morphologically rich languages.

- Even though the objective of unsupervised classification is applying the clustering methods to identify the groups such as positive and negative, some studies have paid attention to calculating a polarity score for the review or opinion in consideration.

The next chapter presents a novel framework for sentiment classification which is based on the resources and methods covered in the previous literature.

Chapter 4: Framework for Automatic Sentiment Analysis

4.1 Introduction

In this chapter, new framework for automatic sentiment analysis is explained in detail. The framework consists of several components integrated together to extract relevant information to classify a collection of reviews in the Sinhala language. The overall process can be divided into four main components; opinion extraction and cleansing, lexical building, feature identification and sentiment classification. The framework is specially designed for non-English languages. The lexical building component is essential if the sentiment classification is carried out using dictionary based techniques. The language Sinhala, in which the study is carried out is described in section 4.2. Section 4.3 presents the methods of opinion extraction and annotation followed by section 4.4 that describes the particular language specific pre-processing steps that are essential for sentiment classification. In Section 4.5 the feature selection procedures are explained in detail. In Section 4.5 special attention is given to the linguistic features of the Sinhala language. Finally, Section 4.6 presents the methods of sentiment classification followed by a summary of the chapter in Section 4.7. Figure 4.1 illustrates the sentiment classification framework with all of its components.

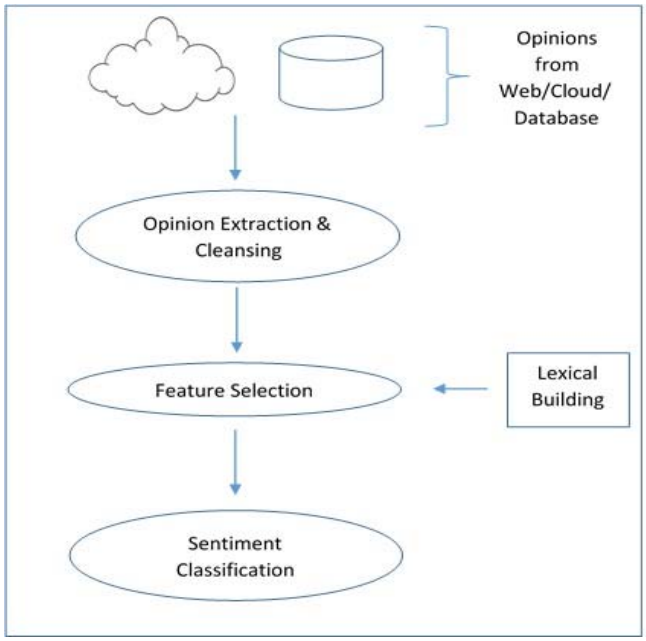


Figure 4.1: Sentiment Analysis process

4.2 The language considered in the study

Primarily, this research was set out to build a framework for author's preferred language; SINHALA/siŋhələ/. Sinhala is one of the several morphologically rich languages for which currently there are no repositories, such as WordNet or Subjective lexicons. It is one of the official languages spoken in Sri Lanka with about 15 million speakers out of the total population of 22 million. Sinhala is spoken in all regions of Sri Lanka except for in the north of the island where Tamil is the spoken language.

Sri Lanka

Figure 4.2: (a) The country in which Sinhala is widely spoken. (b) Language usage in Sri Lanka

(Freeworldmap.net, n.d.) (Politics and History of the Indian Subcontinent, 2014)

Sinhala belongs to the Indo-Aryan branch of the Indo-European languages and it is a morphologically[10] rich language as are some other Indic Languages in the family (Welgama, et al., 2011).

Figure 4.3: Language family that Sinhala belongs to

[10]Morphology is the scientific study of forms and structure of words in a language

4.2.1 The character set

The modern Sinhala character set consists of 18 vowels, 41 consonants, and symbols of dependent vowels signs (Figure 4.4). Dependent vowel signs are known as “pili” in Sinhala (Figure 4.5).

අ ආ ඇ ඈ ඉ ඊ උ ඌ ඍ ඎ ඏ ඐ එ ඒ ඓ ඔ ඕ ඖ
Independent vowels

ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඥ ඦ ට ඨ ඩ ඪ ණ ඬ ත
ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ
Consonants

Figure 4.4: Sinhala vowels and consonants

ා ැ ෑ ි ී ු ූ ෘ ෲ ෟ ෳ ෙ ේ ෛ ො ෝ ෞ
ෘෟ ෲෟ

Figure 4.5: Sinhala dependent vowel signs (pili)

4.2.2 Lexicon, Sentence Structure

The lexicon of the Sinhala language is highly influenced by many languages such as Pali, Sanskrit, Tamil, and English. The ancient Sinhala lexicon came from a source known as “Sidath Sagara” and explains Sinhala grammar in the thirteenth century. Sinhala shares both Indo-Aryan and Dravidian (Tamil) morphological features and some distinct morphological variations (James & Lust, 1998). There are some inflectional[11] and derivational[12] morphological participation observed in the Sinhala language. The inflectional forms of nouns in Sinhala are of five types, gender, number, person, case, and article. From a sentiment analysis point of view, the inflection of case is more influential than the base form of a noun, especially in the possessive case. Furthermore, the analysis shows that Sinhala nouns belong to three categories simple, complex and compound. Among these forms, the Nama vibakthi (complex) is the adjective, which is assumed to be a high dominant candidate for sentiment classification. The validity of the assumption for Sinhala sentiment classification will be explored later in the thesis. Sinhala verbs are mainly divided into two categories; Transitive and Intransitive verbs.

[11]Creates new forms of the same word, the core meaning is same.

[12]Creates new words from old ones, the core meaning might change significantly.

These forms are further inflected to form five linguistic categories of voice, mood, tense, number and person.

Sinhala has two varities; literal and spoken forms. Literal Sinhala is the form of written communication while spoken Sinhala is used for oral communication for all levels of formality (Hilpert, 2006). The syntax structure of the Sinhala is Subject + Object + Verb (SOV) and in this aspect is similar to the other Indo-Aryan languages. In Sinhala, the word order can be changed according to the context giving the free order form of the language. The free order construction of a sentence is significant in the spoken form of Sinhala as well. In this research, the researcher assumed that the flexibility of the order would matter in automatic sentiment classification.

With the introduction of Unicode character encoding system, the number of electronically typed documents in Sinhala increased rapidly. In parallel to this development, making comments or expressions of readers' views on news articles started to grow exponentially. The recent unprecedented growth in readers' comments has opened up an opportunity to research sentiment analysis in Sinhala. Sinhala is referred to as a morphologically rich language for electronic language processing (Welgama, et al., 2011).

Morphology studies the word structure and formation using inflectional and derivational forms. Inflection is the use of morphological methods to generate an inflectional form of the word using lexeme. On the other hand, derivation is used to form new words using affixes. Sinhala is rich in morphology, inflection and derivation. As an example, the sentiment word හොඳ (good) can be infected and derived to different forms as in Table 4.1. The word නොහොඳින් (Adverb) contains both infection and derivational forms, indicating the complex morphological construct of the language.

Table 4.1: Different morphological forms of the word හොඳ (good)

හොඳ (good) (Adjective)	Infected forms	හොඳට (adverb), හොඳින් (adverb), හොඳම (adjective)
	Derived forms	නොහොඳ (adjective), නොහොඳට (adverb), නොහොඳින් (adverb)

The term හොඳම (adjective) intensifies the polarity of the word while නොහොඳ (adjective) determines the negation.

However, there are no inflected or derived forms for the word "good" in English. Instead, by combining another word with "good", a negation or intensification of the word can be obtained. The word "not good" denotes the negation and "very good" intensifies the polarity of the word. In both cases, the grammatical category of the word functions is an adjective.

A similar morphological formation of Sinhala is also monitored in other morphologically rich languages. The Nepali word "राम्रो (good) generally a masculine, inanimate adjective. It can be inflected to "राम्री"(good) form a feminine adjective. Additionally, "राम्ररी" (nicely) denote the adverb of "राम्रो" (good).

The Sinhala language also enriched with well-written language resources, such as dictionaries and texts explaining the language structure, for example, "Vyakarana vivaranya" (Analysis of Sinhala grammar) and "kriya vivaranya" (Analysis of Verbs) are popular among Sinhala scholars. However, this research is the first ever attempt to develop a framework for sentiment analysis using the resources available in the local language. Furthermore, the framework can be generalized for use with any other morphologically rich Indic[13] Language. Through this research suitable contemporary opinion mining methods that are in use for the English Language, will be modified for morphemically rich Indic Languages.

4.3 Opinion Extraction and Annotation

Data for this research on sentiment analysis came from opinions collected from various repositories. A repository can be a website or a database. The opinions collected for the sentiment classification are known as opinion corpora. The main sources are product reviews, opinions on news articles or political debates. The extraction of opinions from a website or database can be performed manually or automatically. Currently, more researchers tend to incorporate automatic opinion collection methods as the manual collection methods are time-consuming and highly labor intensive. Today microblogging websites have become popular, and they are rich sources for sentiment analysis. Unlike other text corpuses, the important task of building an opinion corpus for a product or news reviews is the annotation of the opinions. Researchers in sentiment analysis collect the opinions with the annotation. In general, an opinion or review can be annotated as positive, negative or neutral in consideration of the

[13] Languages that are spoken by the Indian subcontinent people

strength of the polarity of the sentence or document (Pak & Paroubek, 2010). An alternative approach is to classify the opinion as subjective or objective (Liu, 2012). Subjective opinions can be further divided into positive or negative categories (Wiebe, Wilson, & Cardie, 2005). Methods of annotation are typically based on either manual or automatic techniques. For example, a textual news documents MPQA (Multi-Perspective Question Answering) annotated corpus has been built for English using text from a range of sources using a manual annotation process to construct a corpus treating the opinions as private states (Wiebe, Wilson, & Cardie, 2005). The annotation process assumed that there are three types of private state expressions in an opinion namely; explicit mentions of private states, speech events expressing private states, and expressive subjective elements. In identifying these three types, the authors annotated the private state (Opinion) with several categories namely; text anchor, source, intensity, and attitude. The attitude type code was attributed as positive or negative. In manual annotations, researchers annotate the opinions to the categories based on intensity or the polarity of the opinion in order to achieve a fined-grained sentiment analysis (D'Andrea, Ferri, Grifoni, & Guzzo, 2015). Stoyanov & Cardie (2008) incorporate six finer grained attributes namely; Opinion Expression, Source, and Polarity, Topic, Topic span, and Target span tributes enhanced annotation of the MPQA corpus. The polarity in this annotation was three-fold; positive, negative and neutral.

Using the Twitter API, Pak & Paroubek (2010) automatically collected a corpus of Twitter posts. In this corpus, the opinions collected were categorized into three classes positive, negative and objective based on the strength of emoticons assigned to the Twitter comments. Rushdi et al. (2011) generated a corpus of Arabic movie reviews from different web pages and blog sites using a simple bash script for crawling. The corpus consisted of 500 opinions of positive and negative comments. All the cleansing steps were carried manually, and the comments were free from Arabic stop words. Another automatic attempt to collect Twitter comments for the Indonesian language was carried out by Wicaksono, Vania, Distiawan, & Adriani (2014). Their corpus included 5.3 million tweets, but initially, only 637 were annotated manually as positive, negative and neutral for the training data set. Then using an opinion lexicon, the rest of the tweets were classified into three polarities. They used a simple method of classification if more positive words were present in the twitter then it was assigned as positive and if more negative words were present, then it was classified as negative. In addition, the polarity of the tweet was reversed if the sentiment word was proceeded by a negation. In a second approach, a clustering method was proposed and tested to generate more annotated

tweets. A multilingual corpus for English, German and Spanish reviews annotated by Schulz et al. (2010) extended the English corpus constructed by Liu, Hu & Chen (2005). In the annotation process, review sentences for German and Spanish were labeled with a polarity score ranging from 0 to 3 considering the features of English sentences annotation.

4.3.1 Opinions and Data Collection for the study

Data for this proposed study are the comments, feedback or blog contents written by readers, users or customers. Such data for morphologically rich languages are currently limited especially, in Sinhala. Therefore, this research is aimed at carrying out on Sinhala news article comments. The reason of the limitation is a lack of other sources of comments, such as product blogs. However, news article comments in Sinhala are abundant and have been collected from online newspapers. For this study, comments come from an online newspaper "Lankadeepa" (http://www.lankadeepa.lk/) a popular newspaper in Sri Lanka.

The comments on the news articles were collected by two methods. More than 75% of the comments were extracted from a database maintained by the "Lankadeepa" newspaper web administrator. These comments in the context of a variety of domains such as politics, criminals, education, health and environment. In the initial investigation, it was found that the number of comments from each category was not equal and in order to have sufficient opinions for each domain, the additional opinions were extracted from http://www.lankadeepa.lk/ using a web scrapper developed by the author. The scrapper was written in Python, and it supports the utf-8 encoding. The full script of the scrapper available in Appendix A. The comments collected through the web pages were cleaned and pre-processed using the steps given in section 4.4 before storing in a corpus. Table 4. summarized the collected opinions with its categories. The comments collected in this research were stored in a repository which was built flowing the corpus building standards and included all details of the comments (Atkins, Clear, & Ostler, 1992). The details include the source of the comment, date of the comment made, heading of the news articles and the annotation of the comment, whether it is Positive, Negative or Neutral. A sample of the opinions presented in Appendix B. The annotation process is described in section 4.3.2.

Table 4.2: Opinion Distribution across Domain Area

Domain	Number of Opinions
Politics	885
Criminals	908
Education	314
Health	90
Environment	210

4.3.2 Opinion Annotation

As for sentiment analysis conducted in other languages, the collected comments in this research were annotated as positive, negative and neutral. The author engaged three native Sinhala language annotators to assist with the annotation of the comments in the corpus. One expert had a background in linguistics the other two were general Sinhala speakers. Five classification schemas; politics, criminals, education, health and environment was given to the annotators to categorised the comments based on their judgment. The annotators were advised to label the comments into these domains by considering the header of the news article and the content of the comment. They then assigned the polarity of each as either positive, negative or neutral. Besides, the polarity of the comments decided by the agreement between the header of the news article and the content of the comment. In the first round, each annotator was given 700 opinions. In the next two passes, the 700 opinions were swapped among the three annotators. In the end, each of the annotators had annotated a total of 2,100 comments. Then the researcher verifies the final annotation for both topic and the polarity of the 2,100 opinions by examining each manually. The final annotation was based on the following criteria. If all three annotators were agreed in both topic and polarity, then the respective opinion was labeled as given by all annotators. In this process, more than 80% of the interrater agreement was noted. Discrepancies between the annotators were reconciled by the researcher in order to reach a final classification. If two annotators agreed, then the comment was assigned the same label as given by those two annotators on the condition that the author also agreed on both the domain and the polarity classes. In the case of total disagreement (4.04%), i.e. where each annotator assigned a different class, then the author determined the label independently following the same guidelines that applied for the annotators. Some comments (0.74%) were removed from the sample as they consisted of just a single word or gibberish. The inter-annotator agreement was measured using Fleiss' kappa (Bhowmick, Mitra, & Basu, 2008) index and returned an overall agreement of

55.94%. The indexes for the positive, negative and neutral class are 77.32%, 83.45%, and 32.28% respectively. The figures indicate that a moderate degree of inter-rater agreement existed between annotators on an overall basis (for all classes), while substantial agreement existed for the restricted case of positive and negative classes.

Table 4.2 provides examples of some annotated opinions. As an example, the opinion (a) was classified by reviewers as being political. The polarity is negative (N) because the opinion is resigned to the lack of change and the situation remaining or not improving. All annotators determine comment (d) to be about Education – clearly, it is about the implications of policies in the national institute of education (NIE) that are resulting is school children having a work load that is too high for them to manage. This is a negative (N) comment which blames the NIE for the negative impacts on children.

Table 4.3: Annotated Opinions

	Opinion	Date	Domain	Polarity
a.	මොන දේවල් කළද සමහරුන්ගේ ජාතිවාදී අදහස් වෙනස් නොවෙනු ඇත. (Whatever actions taken, the attitude of racists never change)	2013-11-10	Politics	N
b.	ඉන්දියාව අපේ හොඳම මිත්‍ර රට, තරහ කරගන්න හොඳ නෑ. (India is our best friend country; it is not good to antagonise them)	2014-01-08	Politics	P
c.	අනාගත වන්ඩි දැන්මම මර්දනය කලයුතුයි. (Future thugs must be suppressed now)	2013-05-07	Criminal	P
d.	ලංකාවේ ජාතික අධ්‍යාපන ආයතනයේ ඉන්න පණ්ඩිතයන්ගේ වැඩ නිසා තමයි පොඩි දරුවන්ට අනවශ්‍ය බර පැටවිල තියෙන්නේ. (Because of pundits who are working in the National Institute of Education, children have a heavy work load)	2013-10-12	Education	N

4.4 Language specific preprocessing in sentiment classification

Pre-processing describes the type of processing carried out on raw data before it is inputted to the main processing procedure. Commonly used pre-processing stages are; data transformation, noise removal, and normalization. In sentiment classification, the first pre-processing task consists of cleaning the reviews. The cleaning, such as removing the punctuation marks, correcting the spelling mistakes and removing gibberish, is to be completed in this stage. Punctuation marks other than emoticons do not carry any meaning in unstructured text such as customer reviews or opinions and should, therefore, be removed. The words without any meaning, gibberish, are removed in the preprocessing stage to complete the cleaning of opinions. Unlike data mining algorithms, sentiment classification requires special linguistic preprocessing before classification. These stages involve removing stop words, stemming or lemmatization and morphological parsing. Following sections explain each of these stages in detail.

4.4.1 Eliminating the Functional/Stop words

Functional (grammatical) words are words which have little meaning but are essential to maintaining the grammatical relationships with other words in a sentence. Functional words also known as stop words include prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles (Porter, 1997). In text analysis, these words are dropped in order to reduce the dimension of the feature vector. Since function words carry less importance to the meaning of a comment, it is reasonable to remove them. One of the advantages of removing these words is it that their removal reduces the dimension of the word vector constructed for the classification. However, in the case of sentiment analysis, it is essential to decide before classification whether or not these stop words (e.g. “not’, ‘no’, ‘don’t, etc.) will be removed or not because the sense of these words affects the polarity of the opinion (Hidayatullah, 2015). For a given language the set of function words is closed and freely available.

In this study, the list of stop words compiled by the Language Technology Research Laboratory at the University of Colombo Sri Lanka (http://www.ucsc.cmb.ac.lk/ltrl/) is employed. The list consists of a set of negative words that determine the polarity of negative opinions, complex phrases, and contextual feature words. Table 4.3 gives a list the words that were removed by (i.e. these words were included in the classification vector) the author from the stop word list

because these words actively affect the contextual level polarity of a sentence. The explanation of the function is explained in section 4.5.2.

Table 4.4: Word removed from standard stop word list

Word	POS	Translation	Word	POS	Translation
එනමුත්	Conjunction	but	නැත්නම්	Particle	or
එහෙත්	Conjunction	but	නැද්ද	Particle	do not
නමුත්	Conjunction	but	නැහැ	Particle	no
නමුදු	Conjunction	while	නෑ	Particle	not
නැතත්	Conjunction	whether	නිසා	Particle	because
නැතහොත්	Conjunction	or	නිසාත්	Particle	because
නැතැයි	Conjunction	no	නිසාම	Particle	because
නැතිනම්	Conjunction	or	නොමැතිව	Particle	without
නැතොත්	Conjunction	unless	නොව	Particle	not
නැත්නම්	Conjunction	or	බැවින්	Particle	because
නොහොත්	Conjunction	or	බැහැ	Particle	can not
සමග	Conjunction	with a	බෑ	Particle	can not
සමඟ	Conjunction	with the	විරහිත	Particle	unconditionally
සහ	Conjunction	and	විරහිතව	Particle	unconditionally
හරි	Conjunction	right	හරි	Particle	right
හා	Conjunction	and	හරිම	Particle	very
හැබැයි	Conjunction	but	හරියට	Particle	Like
හෝ	Conjunction	or a	හරියටම	Particle	Exactly
නැත	Particle	no	නොවේ	Particle	is not

4.4.2 Stemming

Words in an opinion are made up of many morphological forms of stem words. To normalize the words into their respective stems, a process called stemming or lemmatization is required. Hence, all the opinions undergo stemming to remove the inflectional and derivational morphemes of the non-functional words. Morphemes like plurals, continuous, past, etc. are removed in this process. This helps to reduce the vocabulary size and thereby to improve the accuracy of the classification. But this step is considered to be optional, and several authors have successfully carried out classification directly without stemming. For example, Duwairi & Orfali (2013) have experimentally shown that there was no improvement achieved by stemming for sentiment classification in Arabic. In another recent study conducted in the

Indonesian language, it was also proved that there was no significant achievement gained by stemming (Hidayatullah, 2015).

4.5 Features for Sentiment Classification

As explained in the introduction, the sentiment analysis task is considered to be a classification problem. In general, a review or opinion is classified to determine the polarity strength of the opinion. The classification label can be a categorical type; positive, negative or neutral or a numeric value in range. Features are the primary requirement of any classification problem. The taxonomy of the features used in sentiment analysis varies based on the researcher. However, the thesis considers that in sentiment classification the features can be of mainly two kinds based on the feature weighting scheme; statistical and linguistic.

4.5.1 Statistical features for sentiment classification

In sentiment analysis, the primary feature or attribute for classification is a term or collection of terms. A single word is commonly known as a unigram, and a contiguous sequence of n words is defined as an n-gram. These features or n-grams are weighted by numerical value before applying the classification algorithm. In this section, the standard weighting measures that can apply to sentiment classification are discussed. Statistically, features are weighted based on the frequencies of a feature. In information retrieval, the Term Frequency (tf) is used to represent the relative importance of the feature (word) in a sentence or document. In some cases, a term presence is commonly expressed in terms of a binary weighting, 1 if the feature appears, or 0 otherwise. Some sentiment classification studies found binary weighting to be more valuable than the term frequencies (Pang, Lee & Vaithyanathan, 2002, Thelwall et al., 2010). In addition to tf, some researchers have used a combined feature weighting index tfidf (Term Frequency -Inverse Document Frequency), whereby the term frequency for the comments simply refers to the number of times a given term appears in that opinion (Salton & Buckley, 1998). The tfidf value is normalized to avoid bias in long opinions and to give the exact importance of the word, and is calculated using the following equations:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k k, j}$$

where $n_{i,j}$ is the number of times that the term t_i appears in opinion C_j, and the denominator is the sum of all the words in the opinion C_j.

The inverse document frequency (idf) is a measure of the general importance of the term. idf is obtained by dividing the number of opinions by the number of comments that consist of the term. Then the logarithm of the quotient is calculated as,

$$idf_i = log \frac{|C|}{|\{j: t_i \in d_j\}|}$$

where $|C|$ the total number of opinions are considered and $|\{j: t_i \in d_j\}|$ is the number of opinions in which the term t_i appears. Division-by-zero occurs when the term t_i is not present in the opinions. To avoid this, the denominator can be changed to $1 + |\{j: t_i \in d_j\}|$

Then the final weight is calculated using the following equation,

$$(tfidf)_{i,j} = tf_{i,j} \times idf_{i,j}$$

Paltoglou and Thelwall (2010) proved that this variant of the $tfidf$ weighting improves the sentiment classification significantly. They tested different smoothing factors in order to eliminate random variation. They reported that smoothing has not impact on classification accuracy.

A new calculation of $tfidf$ defined as delta tfidf was first introduced by Martineau & Finin (2009). Delta tfidf is calculated using the following equation,

$$\Delta(tfidf) = n_{i,j} \log \frac{N_t}{P_t}$$

where

n_{ij} - the number of times the term t_i appears in the opinion C_j

N_t - number of documents in the negatively labeled training set with term t_i

P_t - number of documents in the positively labeled training set with term t_i

Martineau & Finin (2009) found that the delta tfidf outperformed the flat term frequency and $tfidf$ weights. They argued that $tfidf$ boosted the accuracy for very frequent words in a document that occur in very few of the other documents. The authors also mentioned that many sentiment words are generic and have low $tfidf$ values. Furthermore, they found that the

accuracy of delta tfidf was higher if the word occurred more often in that text, and was comparatively rare in oppositely labeled documents.

It is a common practice among the sentiment analysis researchers to test the n-gram features by weighting using the weighting methods explained above. However, it should be noted that research into the use of n-grams has been inconclusive. Some researchers have reported that in sentiment classification, the effect of n-gram features is not as beneficial as the unigrams (Pang, Lee, & Vaithyanathan, 2002). While in other studies the researchers have reported moderately better classification performance when using bigrams and trigrams rather than unigrams (Ng, Dasgupta, & Arifin, 2006). In this research, it is planned to experiment the effect of n-gram for sentiment classification for a morphologically rich language, Sinhala.

4.5.2 Linguistic Features for Sinhala

Parts of speech (POS) are the main elements for linguistic features that discriminate among sentiments, such as, positive, negative or neutral. In initial studies on the use of POS of adjectives and adverbs, researchers experimented extensively by applying a combination of both or individually and concluded that POS captures effective linguistic features (Benamara, Cesarano & Reforgiato, 2007). Presently researchers are interested in applying contextual intensifiers, contextual shifters, modal affixes, negations, morphological dependency chunk structures and some morpheme based linguistic identifiers in classification experiments.

Adjectives or adverbs connected by some conjunctions are likely to have the same orientation in some languages (Hatzivassiloglou & McKeown, 1997). Authors developed a model to learn semantic orientation of words based on the above concept, but they removed the conjunction "but" from the conjunction list. The application of nouns and verbs as features of sentiment classification is comparatively limited. But Chesley et al. (2006) considered four classes of verbs; approving, praising, doubting, or arguing with some other features to classify blog posts as objective, positive or negative. In addition to the basic POS features of adjective, adverbs, verbs and nouns contextual linguistic features as well are applied in some recent studies.

The main contextual features are intensifiers and shifters. The contextual intensifier is a lexical item that weakens or strengthens the base polarity of the word that followed the intensifier. For example, "very" is an English intensifier that strengthens the valance of the expression in "very beautiful" positively but it further diminishes the valance in "very difficult". The effect of contextual features in sentiment classification has been measured differently in past studies.

Polanyi & Zaenen (2004) calculated the effect by adding/subtracting one unit (+1/-1) to /from the base valance. The polarity score of the expression “very brilliant” was calculated as +3 by adding +1 to base valance of brilliant, +2. Similarly, “although brilliant” scored 0 as the negative intensifier with “although” acting on ‘brilliant”. Benamara et al. (2007) categorized the adjectives as strong intensifiers or weak intensifiers. The adverbs, such as extremely, immensely and so, are defined as strong adverb intensifiers and weakly, slightly, etc. as weak adverb intensifiers. They defined a set of scoring algorithms which used the degree of meaning and assigned sentiment scores differently. In a study by Jang and Shing (2011) for Korean sentiment analysis, each contextual intensifier that strengthened the original polarity of the term was multiplied by two without considering the semantic intensity.

Two prominent linguistic scholars, Professor W.M.Wijeratne from the Department of Linguistics, University of Kelaniya and Professor Tissa Jayawardena were consulted throughout the research to identify the linguistic features for Sinhala and understand the language constructs.

For this study on sentiment analysis in Sinhala, 11 contextual intensifiers are identified that are considered to be influential in polarity determination. These 11 include seven increasing (scale up) and four decreasing (scale down) intensifiers. In the following section, the sentiment function of each intensifier is explained comparing the effects of the similar form of a word, which is morphologically changed. As an example, the discussion will examine the effect of වඩා (more) in assigning the polarity in an opinion compared with the effect of වඩාත් (even more) in the same context.

a. Increasing contextual intensifiers in Sinhala.

The effect of 6 intensifiers; වඩා (more), වඩාත් (similar to even more), ගොඩක් (much more), විශාල(big), ලොකු(huge), ඉතා(very), ඉතාමත් (very very) always effect the next word and increases the sentiment of the word. That is, if the adjacent word is positive then the polarity of the word becomes more positive and vice versa. The collocation of the above 7 intensifiers were examined in the sample opinion data set experiment in this study. It was observed that in only 40% of instances of the word that followed by වඩා (more) were positive. On the other hand, only 19% of instances in the data set were negative words collocated with the intensifier වඩා (more). The results of the collocation experiments reveals that the intensifier වඩා (more) is more likely to appear before a positive word than a negative word. The intensifier වඩාත්

(even more) morphologically inflects the word වඩා (more) giving greater polarity. The polarity of the phrase වඩාත් හොඳයි (even better) is much higher than the phrase වඩා හොඳයි (more better). It is observed that the occurrence of the word වඩාත් on its own lower and it almost always appears with a positive word. The effect of the intensifier ගොඩක් is similar to වඩා (more). Interestingly, 45% of the next word occurrences are positive with negative word occurrence amounting to 12%. A similar pattern occurs for the word විශාල (huge). In this sample, it is used more frequently next to positive words. However, the total occurrences in the sample was comparatively low.

Next, we examined prepositions which are widely used to elaborate sentiment words in Sinhala. The preposition ඉතා (very), before a noun or verb functions by increasing the strength of the sentiment if the subsequent associated word is a noun. The following sentence illustrates the effect of the preposition ඉතා (very) in an opinion.

ඉතා ප්‍රබල සාක්ෂියක් තියෙනවා නම් තව මොනවට ද මහජන සහය පතන්නේ

(if there is very strong evidence why are you looking for public assistance)

In this sentence, the strength of the sentiment of the noun සාක්ෂියක් (evidence) is increased by the positive sentiment word ප්‍රබල (strong) which is further strengthened by the intensifier ඉතා (very). In other words, having intensifier ඉතා (very) carries more sentiment of the expression than without the intensifier. It was also noted that the above sentence is tends more towards the positive rather than the negative sense. It was observed in the collection of opinions that 63% of positive sentiment words are followed by the ඉතා (more) prepositions and only 18% of negative sentiment words are collocated with this intensifier. This characteristic was also investigated in opinions containing verbs in the context of this research Sinhala corpus. The same intensifier is further scaled up in terms of sentiment polarity when added with the morpheme "මත්" and inflected to ඉතාමත් (very very). The two sentences below show the difference in sentiment expressed with the use of these intensifiers.

ඉතා හොඳ තීරණයක් - A very good decision
ඉතාමත් හොඳ තීරණයක් - A very very really good decision

Even though a direct translation of the two sentences could be considered as giving the same meaning, actually the difference in the sentiment of the second sentence is higher than that of the first one. Hence, when calculating the valance score for sentiment classification, it is important to consider this difference in scale. The collocation of these two intensifiers were

examined in the experimental data, and it was noted that 53% of this intensifier ඉතාමත් (very very) appeared adjacent to positive words, while only 23% of negatives followed the intensifier. In the context of opinions, the intensifiers explained above more often tend to function with positive words than with negative words. This conclusion is helpful when classifying an opinion as positive, negative or neutral especially, when using heuristic based classification techniques.

b. Decreasing contextual intensifiers in Sinhala.

This study identified කුඩා (little), සුළු (small), පොඩි (small), පොඩ්ඩක් (a little) as frequently used intensifiers that degrade the sentiment of the word following the intensifier. The words පොඩි (small) and පොඩ්ඩක් (a little) are spoken words that are not used in standard writing in Sinhala. However, these words are frequently used in expressing opinions in text. Unlike increasing intensifiers explained in part (a) the words කුඩා (little) and සුළු (small) have morphologically inflected forms. The inflected form කුඩාවට (smaller) carries the same sentiment as its base form. Therefore, the valence of the both forms are equivalent and no special consideration is needed for sentiment classification. By examining sample opinions, it was observed that the words කුඩා (little) and සුළු (small) are followed by nouns and no effect by the compound (noun + intensifier) on the total polarity of the sentence of no sentiments. Consider the following sentence,

කුඩා වුනත් මාල දීවයින රජය අපිට වඩා නිබර්ය තීරණ ගත්තා
(Even though Maldives is small the government took the brave decisions)

In this opinion, the intensifier කුඩා (small) has no effect on the assignment of the sentiment of the sentence. The sentiment of the opinion can be determined by the positive word නිබර්ය (brave). In most of opinions the effect of කුඩා (small) on the noun and nouns are less deterministic than adjective and adverbs. The functionality of the word සුළු (little) is similar to කුඩා (small) and the effect of both is experimentally negligible. The occurrence of පොඩි (small) is comparatively higher than of කුඩා (small) in the opinions considered in this study. As mentioned in beginning of the section (b) the word පොඩි (small) is a spoken word and in comment blogs, it was noted that in this blog data that writers tended to use spoken language form rather than written form when expressing their opinions. Additionally, the word පොඩි (small) mostly collocates with nouns that do not have any effect on polarity determination.

In conclusion, the effect of the intensifiers increase the positive sentiment of the context rather than that of negative sentiments. Out of the 11 intensifiers in Sinhala language, වඩා (more), වඩාත් (more), ගොඩක් (more), විශාල (big), ලොකු (huge), ඉතා (very), ඉතාමත් (very very) have more influence than the opposite intensifiers කුඩා (small), සුළු (small), පොඩි (little), පොඩ්ඩක් (a little).

c. Sentiment Shifters in the Sinhala Language

Contextual shifters are of two types; Negation shifters and Flow shifters. Negation shifters reverse the polarity of the term from positive to negative and vice versa. In English, terms such as no, not, nobody, and similar words are an example of negation shifters. In the context of sentiment classification, negations are either; functional (syntactic) negators or content negators (Choi & Cardie, 2008). The negators no, not, and so forth are functional word negators that flip the valance of the neighboring word. Negation handling is important as well as being very complex in sentiment analysis. The thesis observed only two base form function negators නෑ (no) and බෑ (cannot) available in Sinhala. These two base forms of the negators bound with other morphemes make inflected negators. In the following table (4.4) a list the functional negators along with their inflected forms is presented. It is essential to study the grammatical function of the Sinhala negators before adopting the negation handling techniques developed for other languages. The following section describes the linguistic functions of these forms in detail.

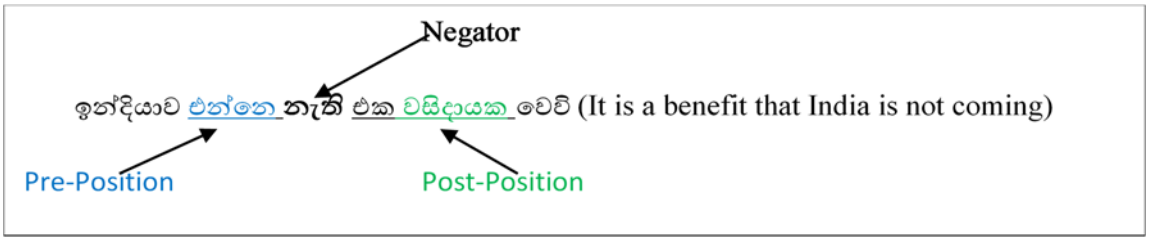
Table 4.5: Inflection of functional negators

Base form	Inflected forms
නැ (no)	නෑ (no), නැහැ (no), නැත (no), නැති (no), නැත්නම් (or not), නැතැයි (not), නැතිනම් (if not), නැතත් (although), නැතොත් (unless), නැතහොත් (or), නැතුව (or not)
බැ (can't)	බෑ (can't), බැහැ (can't), බැරි (can't), බැරිය (can't)

The grammatical function of the above base and inflected forms are complex, and they depend on the different contexts they are used in. Some negators have an effect on the previous words (pre-position) while the others function on the word next to it (post-position).

Definition:

i. Pre-Position: Pre-Position is defined as the word before the negator
ii. Post-Position: Post-Position is defined as the word after the negator



Additionally, the function is dependent on the part of speech. Linguistically, නෑ (no) can function on nouns, verbs and adjectives. On investigation of the sample opinions it was found that 60% of the negators were combined with verbs in opinions. The construct නෑ (no) is an intonation formed of නෑ(no) usually in high pitch of the utterance in order to emphasise with the emotion of the speaker. This negator can also function as නෑ and is predominantly associated with verbs rather than nouns in customer reviews. The table (4.5) summarizes the % of occurrence of pre and post position of negators with the noun, verb, adjective and, adverbs.

Table 4.6: Occurrence of negators after the POS (%)

Negator	Noun	Verb	Adjective
නෑ (no)	27	60	13
නෑ (no)	32	43	8
නැහැ	21	42	6
නැත	58	36	-
නැති	40	22	18

The functionality of නෑ, නෑ and නැහැ affect the pre-position constituents and it was observed that verbs are more affected by these negators. In addition, these negators are in the form of spoken and frequently appeared in comments as follows:

මම දන්නා විදිහට දැනට තිබ්බ සමුළු එකකටවත් ඔය විදිහට නාස්ති කලේ නෑ සල්ලි සහ ධනය
(*To my knowledge, conferences held earlier either did not waste money or wealth)*

Above sentence is in the complete spoken form of Sinhala and a good example of free order. The negator නෑ (no) effected the verb නාස්ති (waste). In order to empathise the, negation present

in spoken Sinhala, in some cases an intonation, is added to the word නෑ (no). The review given below illustrate the work of නෑ (no).

> මේ නිවුස් කිසිම එකක ඇත්තටම වෙච්ච දේ කියල තිබුනේ නෑ
> *(There was not any one that really knew what had happened in this News)*

Even though there is a subtle difference between නෑ (no) and නෑ (no), the significance of the difference in sentiment classification to positive, negative or neutral is negligible (both cases it negates the adjacent word). The most frequent form of negators among these three is නැහැ (no). It is also combined largely with verbs than nouns to decide the polarity of the review. The negator, නැත (no) is a written form used to express the negative polarity of a sentence. Generally, this word නැත (no) appears at the end of the sentence. Therefore, it is a preposition functional negator. In examining the review sample used in this study, it was found that the negator highly affects nouns rather than verbs.

> ආර්ථික අපහසුකම් නිසා ජීවත්වෙන්න මහත් වෙහෙසක් දරන මට මෙහි කිසිම අගයක් හෝ වටිනාකමක් නැත.
> (I have no value or price of this as I do great effort to live because of economic difficulties)

In this sentence, the effect of the negator නැත (no) is on the nouns අගයක් හෝ වටිනා කමක් (value) which refer to මෙහි (this). Therefore, it is justifiable to consider the impact of the negator to its preposition element.

The negator නැති (no) tends to appear in the middle of the sentence when it is functioning on verbs in the sample data examined in this study. Linguistically the negator effects on pre-positions rather than on post-positions. In the first sentence of the following examples the negator acts on the verb එන්නෙ (coming) making the phrase "ඉන්දියාව එන්නෙ නැති (India is not coming)" negative. However, the complete opinion is positive. In the second sentence නැති (no) works on the pre-position noun, backbone (කොන්දක්).

i. ඉන්දියාව එන්නෙ නැති එක ලන්කාවට වසිදායක වෙවි
(It is of benefit to Sri Lanka that India is not coming)

ii. කොන්දක් නැති කෙනෙක්
(having no backbone)

The above described forms of negation in Sinhala are different from the general function of negators; නැත්නම් (or not), නැතැයි (if not), නැතිනම් (unless), නැතත් (whether),

නැතොත්(unless), නැතහොත්(or), නැතුව(without). It is noted that these negation shifters function under a condition i.e., with adjacent, phrases.

> උසස් පෙල කරලා කරකියා ගන්න දෙයක් නැතුව අපේ දරුවො කොච්චර අතරම වෙලා ද දැන්
> *(Our children are stuck at completing the advanced level examination)*

The above sentence can be classified as negative, and it also gives a reason for negativity. The negation is explained in the second part of the sentence. If the aim of sentiment classification is only polarity determination, then the reason is not important. Therefore, in this case, the function of the shifter is negligible in sentiment analysis. To justify the above claim, the author investigated the opinions with a negator that have been classified by annotators. Out of 2083 opinions, only 3% consists of the word නැතුව (without) and of these 51% are negative. With the aim of finding the real influence of the negator, the author manually skimmed all the opinions that consisted of the negator and its measure of the effectiveness. Table 4.6 presents the sentiment distribution (Positive, Negative and Neutral) for each negator.

Table 4.7: The effect of negators

Negator	**Positive**	**Negative**	**Neutral**
නැත්නම්	34	57	9
නැතිනම්	15	54	31
නැතත්	38	50	12
නැතුව	36	51	13

Table 4.6 reveals that more of the opinions containing the negators were classified as negative sentiments rather than as positive or neutral.

The function of the word නැත්නම් (unless) is complex in sentiment classification. In the following sentence two different phrases; one negative and the other one a positive are combined to makes a positive opinion.

> අනිත් රටවල් වලින් එන්නේ නැත්නම් අපිට හොඳයි
> (it is good for us if other countries do not attend)

The phrase "අනිත් රටවල් වලින් එන්නේ නැත්නම් "(if other countries do not attend) is negative and අපිට හොඳයි (good for us) is positive. Overall it is a positive opinion. One can argue that negator has no impact as the final polarity is determined by the second phrase in the sentence.

Linguistically the effect of the negator is on the preposition. The negator නැතිනම් (unless) therefore functions in a similar manner to නැත්නම් (if not) and it is observed that the subsequent phrase after the negator is mostly negative. Consider the following opinion;

> නීත්‍යානු කූලව ළමයි ලබාගන්න සැලස්මක් හදන්න ඕනේ **නැතිනම්** මේ ළමයා අනාරක්ෂිතයි
> (Need legislation to allow adopt the kids. O**therwise,** the child is unsafe)

The negator effect on the first phrase of, නීත්‍යානුකූලව ළමයි ලබාගන්න සැලස්මක් හදන්න ඕනේ (Need legislation to allow adopt the kids) which is neutral. The opinion is negative and the polarity, was assigned in consideration of the second phrase.

d. Flow shifters in Sinhala

Flow shifters control the flow of the sentiment in an opinion. “But”, “however”, and “nevertheless” are examples of flow shifters in English. For the Sinhala language, the author has identified 16 flow shifters which are supposed to control sentiments. The following paragraph explains the flow shifters and their effectiveness in sentiment classification in Sinhala. එහෙත් (but), එනමුත් (but), එනමුදු (however), ඒත් (but), එනයින් (thus), නමුදු (as a reason), හැබැයි (because), නිසා (because), නිසාත් (because), නිසාම (because), බැවින් (because), අනුව (according to the), එවිට (and), හින්දම (by then), හින්දා(solely), නම්(if) are possible flow shifters in Sinhala language. It is observed that only 7 shifters were used by the readers in the opinions extracted from the online newspaper. The shifters එනමුත් (but), එනමුදු (however), නමුදු (of because), නිසාත් (because), and නිසාම (because) are tend to be written forms of the shifters that typically appear in classical writings. This may be the reason for the use of only 7 shifters in the opinions as the comments are in a form that is closer to that of spoken form. These flow shifters and their percentage distribution in the test sample, as identified by manual sentiment classification, is given in table 4.7.

Table 4.8: Sentiment distribution of flow shifters (%)

Flow Shifter	**Positive**	**Negative**	**Neutral**
එහෙත්(but)	67	22	11
ඒත්(but)	38	35	27
හැබැයි(but)	24	45	31
නිසා(because)	22	45	33
බැවින්(because)	29	71	0
අනුව(according to the)	25	42	33
නම්(if)	32	44	22

එහෙත්(but) is a function word that is used to oppose the previous expression. In the following opinion the first expression සොභාවික ආරක්ෂාව ගැන කියාලා තිබුනා (Stated the natural protection) is a positive expression but the second one; එය ක්‍රියාත්මක කිරීමෙන් නිලදාරින්ට කොමිස් ලැබෙන්නේ නැහැ (no commission for the officials by implementing) negates the expression. Nevertheless, the full opinion is negative.

> සොභාවික ආරක්ෂාව ගැන කියාලා තිබුනා **එහෙත්** එය ක්‍රියාත්මක කිරීමෙන් නිලදාරින්ට කොමිස් ලැබෙන්නේ නැහැ. (Stated the natural protection but no commission for the officials by implementing it)

It can be concluded that it is sufficient to consider the second opinion in polarity assignment for such an opinion. The function of the flow shifter ඒත් (but) is more complex than එහෙත්(but). In the case of ඒත් (but) both expressions can be positive or negative. The opinions given below are examples where the (iii) is negative and, the (iv) is positive.

iii. මදු ස්කොලේ ගිහින් නැහැ **ඒත්** යවන්න කාලෙකුත් නැනේ
(Madu hasn't attended school but no time to go to school again)

iv. හරිම ලස්සනයි **ඒත්** මට බය තව ටිකක් කල් යද්දී ගස් කොළන් කපන ආකාරයට මගේ රටේ මේ දේශගුණය රදා පවතියිද කියලයි
(Very beautiful but I wonder how long it will because of cutting down flora and fauna in this manner)

In the (iii) example, both expressions combined by the shifter (ඒත්) are negative hence the complete opinion is negative. On the other hand, the next opinion (iv) was labelled as a positive sentiment by considering the first expression only, and it is observed that the phrase of the shifter is considered to be neutral statement. While the shifters එහෙත් (but) and ඒත් (but) are opposing the expression that is expressed before the shifter, හැබැයි (but), නිසා (because), බැවින් (because), and අනුව (according to the) are supporting the expression. Additionally, it was noted that the word හැබැයි (but) appeared in the beginning of the sentence. In these cases, the opinion implicitly refers to the subject of the news. According to the analysis results in table 4.7 more negative opinions are found in the sample than positive when the sentence consists of the word හැබැයි (but).

> ඉන්දියාව ආවේ නැතුවට කමක් නැහැ **හැබැයි** ඉන්දියාවට ගැති වෙන්න එපා (It is ok that India is not attending but do not surrender to India)

In this sentence, above, the phrases before shifter: හැබැයි (but) is positive. However, this phrase includes a sarcastic expression even though it is positive. On the contrary the second phrase after the flow shifter is negative and that determine the polarity of the opinion. The thesis believed the sentiment of the opinion could be decided by the empathetic phrase and it is the second phrase. In most of the cases, the emphasized content exists after the flow shifter. Therefore, it is sufficient to consider the phrase after the shifter for sentiment classification also reducing the work load.

Another frequently used flow shifter is නිසා (because). The shifter always appeared in the middle of the opinion and it combined the two phrases one of which contained the sentiment bearing content. In the following opinion, sentiment is carried in the second part (In the English translation the second part of the Sinhala opinion is the first phrase of English translation) but the reason of the negative polarity is given in the beginning of the sentence.

> ආර්ථික අපහසුකම් **නිසා** ජීවත් වෙන්න මහත් වෙහෙසක් දරන මට මෙහි කිසිම අගයක් හෝ වටිනාකමක් නැත (I have no value and price of this as I do great effort to live **because** of economic difficulties)

The flow shifter බැවින් (because) occurs in very few opinions and the function of the word has an effect on both the contents before or after the word depending on the comment. The shifter අනුව (according to the) also combines two phrases; one phrase containing the justification and the other one containing the sentiment. In majority of the cases, the sentiment content occurred after the shifter. In the following example, the polarity is defined by the phrase මෙම ක්‍රියාව තරම් අනුවණ ක්‍රියාවක් තවත් හමුවී නොමැත (No action found than this foolish one)

> දැනට ලැබිලා තිබෙන ප්‍රතිචාර අනුව හා මගෙ අදහස **අනුව** මෙම ක්‍රියාව තරම් අනුවණ ක්‍රියාවක් තවත් හමුවී නොමැත. (In my opinion, this is a foolish course of action).

නම් (if) is the most frequent flow shifter found in the sample considered in this study. The function of this shifter is more important than the previously discussed shifters. More than 22% of the opinions manually classified for the study are contained this shifter. The shifter was used in the middle of the opinion joining two phrases that depend on one another. This feature has a much higher influence in polarity determination than the others. The complexity of the නම් (if) shifter is illustrated by the following example.

මේකේ ගමක පෙනුමක් නම් නැහැ තනිකරම ලස්සන කැලෑවක් නේ (If not solely on this village look beautiful forest).

The opinion has a negative word නැහැ (no) that conditionally negates the first part of the sentence මේකේ ගමක පෙනුමක් (look like a village). The emotional expression of the complete opinion is positive even though the first part is negative. This example is difficult to analyse using automatic sentiment classification because understanding the discourse of the sentence is complicated.

4.6 Sentiment Classification Techniques

Text categorization is a very broad and active area of information research. Categorization can be defined as an act of sorting and organizing things into groups, classes, or, as you might expect, categories. Text categorization is the task of automatically building categories, using machine learning techniques. The domain of text categorization can be a set of words, sets of lines, sets of paragraphs or even sets of documents. The particular domain is selected based on the requirements of the classification. In sentiment classification, an opinion is considered to be a document. With the rapid growth of the online information, document categorization has become one of the key techniques for handling and organizing online text data.

Automatic classification of documents is an increasingly important tool for handling millions of documents in World Wide Web. Today millions of documents are accumulated on the internet. Hence, the ability to retrieve a correct document is much more possible though becoming increasingly difficult. Therefore, developing more efficient and effective user-friendly tools for retrieving correct information has great demand in the cyber world.

Classification is a Machine Learning (ML) technique used to predict group membership for data instances. Every instance in any dataset used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical or binary. If instances are given with known labels, then the learning is called supervised in contrast to unsupervised learning, where instances are unlabelled. By applying these unsupervised algorithms, researchers hope to discover unknown, but useful, classes of items. Therefore, the main taxonomy classification techniques are Supervised and Unsupervised classification.

In supervised algorithms, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by a human. In practice, a certain segment of data will be labeled with these classifications. The machine learner's task is to search for patterns and construct mathematical models. These models then are evaluated on the basis of their predictive

capacity in relation to measures of variance in the data itself. Decision Tree and Naïve Bayes are examples of supervised learning techniques.

Unsupervised learners are not provided with classifications. In fact, the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out the similarity between pieces of data to determine whether they can be characterized as forming a group. These groups are termed as clusters.

In unsupervised classification, often known as 'cluster analysis' the machine is not told how the data are grouped. Its task is to arrive at some grouping of the data. In a very common of cluster analysis (K-means), the machine is told in advance how many clusters it should form. This, determination of the number of clusters, is a potentially difficult and arbitrary decision to make.

4.6.1 Supervised Sentiment Classification Methods

The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. Supervised classification is the most common classification technique used in sentiment analysis. There are many methods that have been developed using artificial intelligence and statistics. Logic and Perception based methods are developed in artificial intelligence whereas Bayesian Networks and Instance based techniques have been developed by the statistics community.

a. Support Vector Method

The Support Vector method is primarily defined for a two-class classification problem. Support Vector Machines (SVM) are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Decision planes are the classifiers either a line or a curve. A simple classifier may use linear decision planes whereas more complex structures are used in complex classifiers. Classification tasks based on drawing separating lines to distinguish between the objects of different class memberships are known as hyperplane classifiers (Vapnik, 1998). SVM is primarily a classifier method that performs classification tasks by constructing hyperplanes, in a multidimensional space, that separates cases of different class labels. SVM

supports both regression and classification tasks and can handle multiple continuous and categorical variables.

To construct an optimal hyperplane, SVM employs an iterative training algorithm; this is used to minimize an error function. According to the form of the error function, SVM models can be classified into distinct groups.

In the simplest SVM, training involves the minimization of the error function,

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

Subject to the constrains

$$y_i(W^T\varphi(x_i) + b) \geq 1 - \xi_i \; and \; \xi_i \geq 0, i = 1, \ldots\ldots, N$$

Where,

C is the capacity constant,

w is the vector of coefficients,

b a constant and

ξ_i are parameters for handling non-separable data (inputs).

The index i is the instance of the N training cases. Note that $y \in \pm 1$ is the class label and x_i is the independent variable. The kernel φ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over-fitting.

The success of SVM in text categorization lies in its automatic capacity tuning by minimizing ||w||, extraction of a small number of support vectors from the training data that are relevant for the classification (Kwok, 1998). SVM in text categorization is a problem of very high dimensionality. Since the document's topics are not mutually exclusive, text categorization is usually analysed as a series of dichotomous classification problems, i.e., whether the document belongs to a particular topic or not. Pang et al. (2002) achieved 82.9% accuracy in classifying movie reviews using unigram features with binary weightings. It was noted that the performance was significantly dropped when the linguistic features such as adjectives were introduced. Khoo & Chan (2003) also observed a similar behaviour of reducing

the accuracy by applying linguistic features in the study. Overall they achieved an 81.7% accuracy using WordNet features weighting each term by its frequency.

b. k-Nearest Neighbour Classification (kNN)

The kNN classifier is based on the assumption that the classification of an instance is most similar to the classification of other instances that are nearby in the vector space. Compared to other text categorization methods, such as Bayesian classifiers, kNN does not rely on prior probabilities and is computationally efficient (Han, Karypis, & Kumar, 2001). The main computation involves the sorting of training documents in order to find the *k* nearest neighbors for the test document.

To classify a class-unknown document X, the kNN classifier algorithm ranks the document's neighbours among the training document vectors and uses the class labels of the *k* most, similar neighbours, to predict the class of the new document. The decision of the kNN can represented as follows;

$$f(X) = \sum_{d_i \in KNN} sim(X, d_i)\, y(d_i, C_i)$$

Where $f(X)$ is the label assigned to the document X. C_i category with respect to X, if d_i belongs to the category c_i, $y(d_i, C_i)$ is equal to 1 otherwise 0.

The classes of these neighbors are weighted using the similarity of each neighbor to X by $sim(X,\ d_i)$, where similarity is measured by Euclidean distance or the cosine value between two document vectors. The cosine similarity is defined as follows:

$$sim(X, D_j) = \frac{\sum_{t_{i \in (X \cap D_j)}} x_i \times d_{ij}}{\|X\|_2 \times \|D_j\|_2}$$

Where,

X is the test document, represented as a vector.

D_j is the j^{th} training document.

t_i is a word shared by D_j and X. x_i is the weight of a word in X; d_{ij} is the weight of word $t_i$in document D_j; $||X||_2$ is the norm of X, and $||D_j||$ is the norm of D_j. The norm is defined as

$$\|X\|_2 = \sqrt{x_1^2 + X_2^2 + X_3^2 + \cdots}$$

A cut-off threshold is needed to assign the new document to a known class.

kNN classification is an instance-based learning algorithm that has proved to be very effective in text classification (Han, Karypis, & Kumar, 2001). The success of this method is due to the availability of effective similarity measures, such as cosine measure. However, Han et al. (2001) claim that the effectiveness of these similarity measures becomes worse as the number of words increases.

In an experiment on Chinese sentiment classification, Tan and Zhang (2008) compared kNN with SVM, Naïve Bayes, and Winnow classifier. kNN were observed to be poor in the performance of the methods compared. They set K value to 13 and explained the significant cost of adjusting the value. A novel sentiment classification algorithm was introduced to improve kNN by using single-pass clustering algorithm (Pin et al., 2013). In the single-pass clustering, documents were clustered sequentially. Then kNN was applied to sentiment classification using these clusters. Classification results indicated that the proposed algorithm outperformed Naïve Bayes and SVM.

c. Naïve Bayes

The Naïve Bayes Classifier technique is based on Bayesian theory and is particularly suited to situations where the dimensionality of the inputs is high.

Let R = $\{r_1, r_2, r_3, \ldots r_n\}$ denote the set of training opinions, where each opinion is labelled with one of the category in C = $\{c_1, c_2, c_3, \ldots c_k\}$. Given some new opinions, the aim is to estimate the probability of each code. Using Bayes rule, in general

$$p(c/r) = \frac{p(r/c)p(c)}{p(r)}$$

Since only interested in the relative order of the codes probabilities (given r) and by definition, p(r) is independent of C; one can focus on;

$$p(c/r) = p(r/c)p(c)$$

If the ordered sequence of unique words that compose the opinion r is denoted by

$r = \{w_1, w_2, w_3 \ldots w_p\}$

Then,

$$p(r/c) = \prod_{i=1}^{p} p(w_i/w_1, w_2, w_3, \ldots\ldots, w_{i-1}, c)$$

However, using the Naïve Bayes assumption, we assume that the probability of each word in an opinion is independent of its context (Murphy, 2006). More formally the following approximation ("bag of words" model) is used

$$p(w_i/w_1, w_2, w_3, \ldots\ldots, w_{i-1}, c) = p(w_i/c)$$

Such that

$$p(r/c) = \prod_{i=1}^{p} p(w_i/c)$$

Thus to estimate, $p(c/r)$ all that is required is to estimate $p(w/c)$ and $p(c)$, for all words and all codes. The following is used to estimate $p(c)$

$$p(c) = \frac{n(r,c)}{\sum_{c \in C} n(r,c)}$$

Where $n(r, c)$ is the number of training opinions in the category c. The conditional probabilities of the words in c is estimated by

$$p(w_i/c) = \frac{n(c,w)}{\sum_{w \in W} n(c,w)}$$

Where $n(c, w)$ is estimated by

$$n(c,w) = \sum_{r \in R} n(r,w)$$

Where $n(r, w)$ is a number of occurrences of the word w in the opinion r which is coded as c. Then,

$$p(c/r) = p(r/c)p(c) = p(c) \prod_{i=1}^{p} p(w_i/c)$$

And classify r into a possible category c using

$$\underset{c_j}{argmax} \left[p(c_j/r) \right]$$

Among sentiment classification researchers and communities, Naïve Bayes is a very popular and commonly used method with proven success. One of the reasons of wide use of Naïve Bayes is that it is a fast and accurate classification method (Narayanan, Arora, & Bhatia, 2013).

In the Boolean Multinomial Naïve Bayes probability (BMNB) for the modal is calculated based on the presence or absence of the features (Agarwal, Mittal, Bansal, & Garg, 2015). While the Multinomial Naïve Bayes with term frequencies (TMNB) is a probability-based learning method that constructs a model by using the term frequency of a feature/word/term to compute the probability.

4.6.2 Unsupervised Sentiment Classification

a. Hierarchical Clustering

A hierarchical clustering is a hierarchy with the usual interpretation that each node stands for a subclass of its mother's node. In hierarchical clustering, the assignment is usually hard. In the hard assignment, each object is assigned to one and only one cluster. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Agglomerative techniques are more commonly used.

In text categorization research, most of the studies have focused on flat classification where the predefined categories are considered for classification, and there is no structure defining the relationships among them (Sun & Lim, 2001). Such categories are also known as flat categories. However, when the number of categories grows to a significantly large number, it becomes much more difficult to cluster and classify the categories.

Hierarchical classification allows us to address a large classification problem using a divide-and-conquer approach. At the root level in the category hierarchy, a document can be first classified into one or more sub-categories using some flat classification method(s). The classification can be repeated on the document in each of the subcategories until the document reaches some leaf categories or cannot be further classified into any sub-categories. A few hierarchical classification methods have been proposed recently. In most of the hierarchical classification methods, the categories are organized in tree like structures. On the whole, we can identify four distinct category structures for text classification. They are Virtual category

tree, Category Tree, Virtual directed acyclic category graph and Directed acyclic category graph.

b. K-Means clustering

The k-means algorithm is one of the most widely used central clustering techniques (Ghwanmeh, 1998). In the algorithm, the data set is divided iteratively into k clusters by minimizing the average squared Euclidean distance between the observation and its cluster canter. The algorithm starts with assigning k observations as initial cluster centroids and assigning all the observations to the nearest cluster. After this new clustering, the centroids are calculated as means of the observations belonging to that cluster. The observations are assigned again to the new clusters, and new cluster centroids are once again calculated. This iteration procedure is continued until the centroids stabilize.

The quality of the document list produced after classification depends on the number of clusters. Indeed, k-means like methods require some a-priori decisions about the number of clusters. It is critical but not so easy to determine the number of clusters even if we have shown that it could be computed effectively according to the requirement. This is the main drawback of this clustering technique.

4.7 Novelty of the proposed framework

Linguistic analysis explained in section 4.5, indicates the significance of the linguistic features in sentiment classification for morphologically rich languages such as Sinhala. Therefore, in addition to POS, the linguistic features of negations, intensifiers and flow shifters are included in the classification model under the proposed framework for sentiment classification. The impact of the above linguistic features was modelled by using a novel approach that considered the neighborhood of the word that was inflected. Moreover, the framework also proposed to combine rule-based reasoning with sentiment classification techniques referred to in section 4.6 for this framework. The impact of including the linguistic constructions in the proposed framework is described in chapter 7.

4.8 Chapter Summary

In this chapter, a framework for sentiment classification especially, for morphologically rich languages was explained in detail. At the beginning of the chapter, the methods for data extraction and pre-processing were presented. The statistic features and other linguistic features, such as shifters, negators, etc. that can be utilized for sentiment classification were described extensively with examples in the target language which is Sinhala. The classification methodologies and their theoretical background were elaborated on with some related work. In chapters following this one, the methodologies and approaches investigated for opinions in Sinhala are presented.

Chapter 5: Automatic Lexicon construction for Sentiment Analysis

5.1 Overview

As mentioned in chapter 2, sentiment lexicons are the primary resource for sentiment classification based on dictionaries. This chapter describes three algorithms for constructing a sentiment lexicon for a morphologically rich language, i.e., Sinhala. Initially, a method based on a cross-linguistic approach is investigated as a baseline method which is then compared with two alternative methods, namely, electronic dictionary (morphemes) based and graph theory based. In the electronic dictionary based approach, the aim is to construct a list of positive and negative words retrieved from an electronic dictionary. The objective of graph based lexicon construction is to assign a sentiment or polarity score using a network model. All three resources are built upon a bilingual dictionary available for the Sinhala language. The constructed lexical resources are then evaluated manually and using supervised classification techniques.

It was concluded in section 2.7 of chapter 2 that the majority of sentiment classification studies employ a lexical resource known as sentiment lexicon. A sentiment lexicon or subjective lexicon is a collection of words with their associated polarities. Developing a sentiment lexicon is a challenging task. Generally, there are two approaches used: dictionary and corpus-based. In this chapter, the dictionary based approach for the Sinhala language is investigated. The chapter begins with a discussion of some of the popular sentiment lexicons available for English and some other languages. The properties of a good subjective lexicon and their effectiveness in sentiment classification are discussed in section 5.2. Section 5.3 examines some readily available sentiment lexicons. The common sentiment lexicon building techniques are detailed in section 5.4. Three novel methods for sentiment lexicon construction for morphologically rich languages are proposed. Section 5.5 presents the dictionary based cross linguistic approach and section 5.6 the method incorporating morphological features. The graph based method is discussed in section 5.7. In each case, the proposed approaches were implemented for Sinhala and the adaptation and challenges faced are presented. A summary of the chapter, the novel methods for Sinhala and the experimental results is provided in Section 5.8.

5.2 Properties of Sentiment Lexicon

Three basic properties of a sentiment lexicon are its coverage, content type, and generation process. Coverage refers to the degree to which the lexicon can be applied to any application domain for sentiment analysis. The poorer the coverage a lexicon will be restricted to a particular subset of opinions. Based on their degree of coverage lexicons can be divided into two types:

- *General purpose* lexicons that are independent of the application domain.
- *Domain specific* lexicons that can be used only for domain specific opinions.

As a result of the literature survey undertaken as part of this research, it was observed that currently more domain specific lexicons are constructed than the general purpose subjective lexicons. It was also noted that most of the general purpose lexicons are constructed by gathering the words from dictionaries. On the other hand, corpora are used for domain specific lexicon construction.

Another distinguishing quality of a sentiment lexicon is the content of the lexicon. Different lexicons contain different word attributes and thus have different dimensions. Most sentiment lexicons contain lexical entries along with an assigned polarity for each word. The polarity of the word in the lexicon can be either quantitative or qualitative. Quantitative polarities have a numerical score which represents the degree of the sentiment numerically. Quantitative polarities can be interpreted as qualitative by setting a threshold and assigning words to be either positive or negative. Thus, qualitative polarities may be represented simply as positive or negative sentiments or by a range of classes based on words which express emotion, e.g., happiness, anger, etc. (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

Sentiment Lexicons are also sometimes differentiated by the methods they use to collect the word list. Both manual and automatic assembling process have been reported in the literature. The majority of the lexicons available are compiled manually. The time and intensive labor required are the main disadvantages of the manual lexicon construction approach.

5.3 Some available lexicons and their properties

Bing Liu's “Opinion Lexicon” is comprised of 2006 positive and 4783 negative English words (Hu & Liu, 2004). In this lexicon, no polarity scores are assigned to the words. The list includes words with spelling errors, not as mistakes but to accommodate for misspelled words that

frequently appear in social media. There is no assessment of the effectiveness of this lexicon available in the literature. Wilson, Wiebe, and Hoffmann (2005) maintain the Multi-Perspective Question Answering (MPQA) subjectivity lexicon. There are over 8000 lexical entries in this lexicon, and each entry is annotated with the attributes; strength (strong or weak subjective), the length of the word, part of speech, whether stemmed or not and polarity (positive or negative). Evaluation of this constructed lexicon showed that the classification improved significantly when polarity features were incorporated in the classification.

SentiWordNet is a comprehensive and popular subjective lexicon constructed by Esuli, and Sebastiani (2006). In addition to the part of speech of the word, it contains information about the positive and negative sentiment score for each term, a synset ID and synset term, and gloss. The significant difference between SentiWordNet and other lexicons is its coverage — more than one hundred thousand word entries are available. Empirical studies revealed that this lexicon has high effectiveness not only for English review classifications but also in the analysis in multilingual domains (Ohana & Tierney, 2009; Denecke, 2008).

The Harvard General Inquirer (HGI), which is freely available for academic research purposes, provides a richer linguistic featured sentiment lexicon and contains syntactic, semantic, and pragmatic information along with the part of speech for each term. HGI consists of 11,788 entries and is described in 184 classes. Linguistic Inquiry and Word Counts (LIWC) is a proprietary lexicon that consists of categorized regular expressions (Pennebaker, Boyd, Jordan, & Blackburn, 2015). It is difficult to compare the above-mentioned lexicons and thus it is difficult to decide which is the best one for sentiment analysis. The only way to compare them is to use word-wise comparison for agreement or disagreement. HGI and LIWC are highly correlated and showed 0.5% disagreement. On the other hand, MPQA and SentiWordNet showed the highest degree disagreement (27%) in word-wise comparison (Potts, 2011).

5.4 Automatic lexicon construction

As mentioned in the literature review in chapter 2, automatic methods for sentiment lexicon construction are either dictionary based or corpus based. In both approaches, the process begins with a set of manually selected words known as seed set which is then propagated in a resource using linguistic relationships —usually semantic similarities. In this section, dictionary based sentiment lexicon construction is explained in detail with the aim of applying the same method

for morphologically rich languages. The language considered for this purpose in this study is Sinhala and the structure of the language given in chapter 3.

Thesauri or linguistic knowledge rich lexical resources are known as WordNets (Miller, 1995) are the most common resources used in the dictionary based sentiment lexicon creation. An examination of the literature revealed that dictionary based methods for non-English languages had taken two main directions; the cross-language approach and mapping of words from one language to other. In the cross-language approach, the whole dictionary is translated into the English. The polarity information from the English sentiment lexicon is applied directly to the translated word. In the mapping method, the word from one language is mapped to the other, in most cases using WordNet relations of both languages (Badaro, Baly, & Hajj, 2014). The words are linked through offsets of the WordNet resources in these mapping. Offsets are unique identifiers for entries of WordNet repositories.

The following paragraphs present a novel approach to building sentiment lexicon for Sinhala Language using a bi-lingual dictionary. This approach is independent of a WordNet type tool and a seed set.

5.5 Dictionary based Sentiment Lexicon construction for Sinhala

To construct a sentiment lexicon for the Sinhala language, a baseline method is proposed using an electronic dictionary and a publicly available sentiment resource in English. This is the first attempt at developing a sentiment lexicon for the Sinhala language, and no published records have been found related to sentiment analysis of the Sinhala language. As part of this research, a sentiment lexicon for the Sinhala Language has been developed with the aid of the English sentiment lexicon (SentiWordNet 3.0) compiled by Esuli and Sebastiani (2006). The English SentiWordNet 3.0 used in this study contains more than 100,000 words, which occur in different contexts, along with their positive and negative scores. A part of speech (POS) tag for each word is also included in the SentiWordNet 3.0. The “gloss” gives a hint of the context in which a word can appear, is also included. Table 5.1 summarizes the structure of SentiWordNet3.0 with examples.

Table 5.1: Sample entries of the SentiWordNet 3.0. Where a = adjective and the PosScore and NegScore are in the range of 0 to 1.0

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	5599	0.5	0.5	unquestioning#2 implicit#2	being without doubt or reserve; "implicit trust"
a	11665	0.125	0.375	too-greedy#1 overgreedy#1	excessively gluttonous
a	15720	0.125	0.5	rife#2 plethoric#1 overabundant#1	excessively abundant
a	16247	0.125	0.5	superabundant#1	most excessively abundant
a	16647	0.125	0.5	verdant#1	characterized by abundance of verdure
a	17688	0.375	0.25	unabused#1	not physically abused; treated properly
a	196560	0.25	0.25	unalarming#1	not alarming; assuaging alarm

The English/Sinhala dictionary contains synonyms for each Sinhala word and an English word as the direct translation for the original Sinhala word. A sample from the English/Sinhala dictionary is given in Table 5.2.

Table 5.2: Structure of the Sinhala Dictionary (Sample)

English Word	Sinhala word	Synonym1	Synonym2	Synonym3
abandoned	අතරමං කළ	අත්හළ	අශිෂ්ට	ජරාවාස
abasement	අවමන් කිරීම	නින්දාවට ලක්වීම	පහත් කිරීම	පරිභව කිරීම
aggressiveness	ආක්‍රමණකාරීත්වය	ආක්‍රමණිකත්වය	වණ්ඩතාව	කලහකාරී බව
blinking	අතිශයින් අන්ත	මෝඩ		
blip	සුළු වරදක්	මඳ වෙනස්වීමක්		
bliss	අතිශය ප්‍රීතිය	ප්‍රමෝදය		
brave	එඩිතර	නිහඬීත	වික්‍රමාන්විත	
bravely	නිහඬීතව			
bravery	ධෛයර්ය	නිහඬීතකම		

This dictionary contains detailed inter language relationships including the many possible Sinhala synonyms, i.e. a maximum of 11 Sinhala synonyms are possible for a single English

word in the dictionary. This Sinhala lexical resource is a basic dictionary, and there is no linguistic information such as part of speech or gloss included. Mapping the words between two resources (the English/Sinhala dictionary and the English SentiWordNet 3.0) without having such linguistic information is a complex task. With this difficulty in mind, one of the objectives this research is to evaluate the generation of a sentiment lexicon from such minimal and limited resources.

The novel lexicon construction mapping method adopted is presented in Figure 5.1. In this experiment, the two lexical resources were bridged using the English word that is common in both dictionaries, as the search key. Each English word in the English/Sinhala dictionary was used to search for a matching English word in SentiWordNet 3.0. This initial mapping resulted in the extraction of 72,049 unique entries for 22,296 English words. Subsequently, all the Sinhala synonyms of the English word were extracted from the English/Sinhala and added to the appropriate SentiWordNet 3.0 entry which included the relevant linguistic information for the English word.

In undertaking this mapping, it was assumed that:

- the sense of the word in the two languages was the same.
- the sentiment score of an English word as calculated for use in English opinions was the same for the matched Sinhala word
- POS in both languages is equivalent.

These exhaustive searches consisted of several matching English words embedded in POS. But for this experiment, only the adjectives and adverbs were added to the lexicon as they are considered to be the most important language units (parts of speech) when analyzing sentiments in a language (Benamara, Cesarano, & Reforgiato, 2007). In assuming that POS in Sinhala words are the same as those in English any complexities relating to POS within a Sinhala, sentence have been avoided. Through this selection process, 10,778 Sinhala adjectives were obtained in a different context (POS) where the corresponding English term occurred. The Sinhala adverb search found 1,364 matches in different POS in English. Hence, the final Sinhala lexicon (adjective and adverb list) with positive and negative sentiment scores the same as the corresponding English word in SentiWordNet 3.0 consisted of 5,973 unique adjectives and 405 unique adverbs.

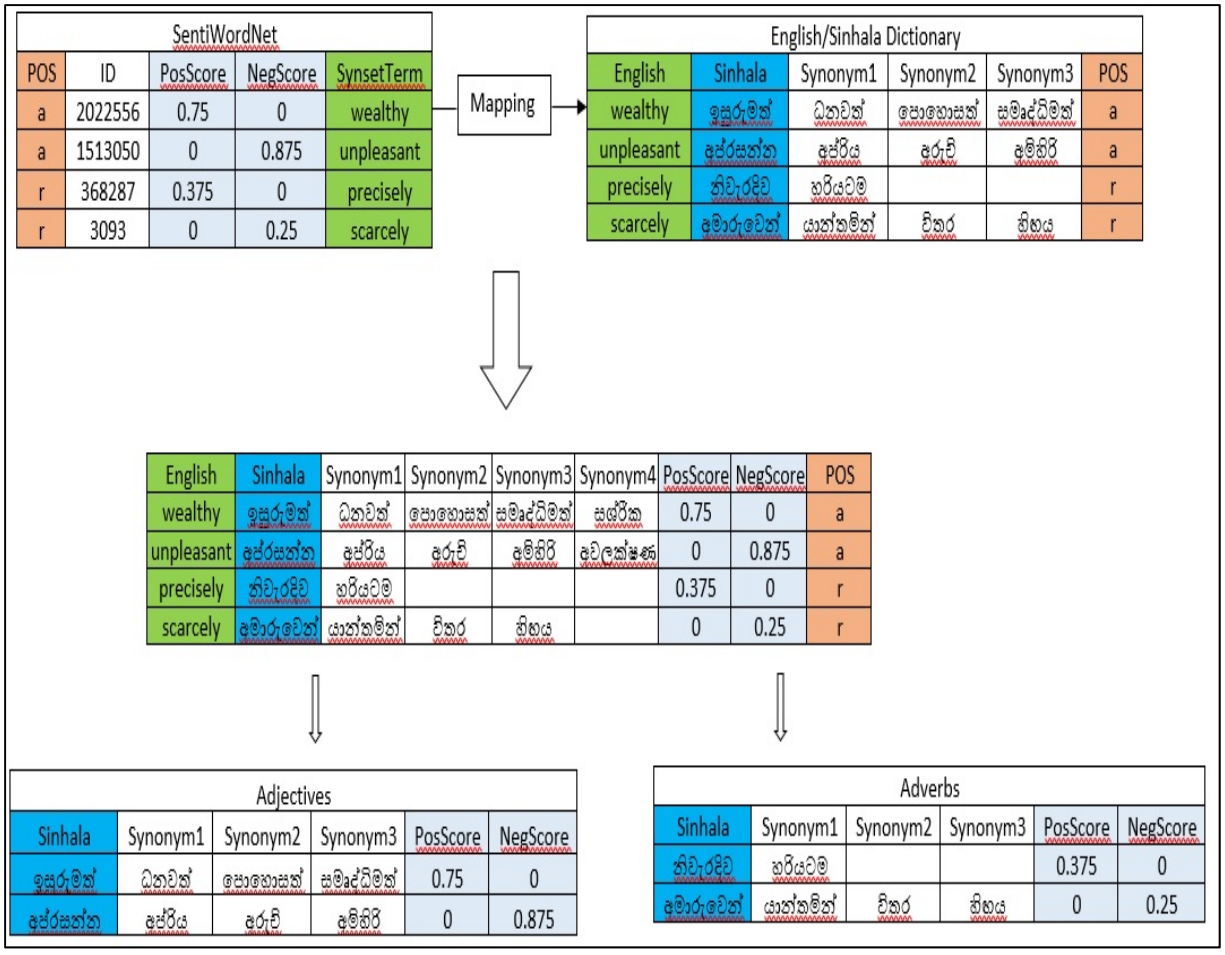


Figure 5.1: Sentiment Lexicon construction by matching two lexical resources

5.5.1 Evaluating the lexicon

The constructed sentiment lexicon was then evaluated using 2,083 manually classified opinions from news article collected from a leading Sinhala newspaper website (www.lankadeepa.lk). The complete explanation of the data (opinions) used for the evaluation is given in chapter 4. The opinions supportive of the article were classified as positive (P) whereas, those criticizing the topic were marked as negative (N) and any unrelated or no sentiment to the topic (neutral) were classified as objective (O).

The newly constructed and novel Sinhala lexicon resource consisting of adjective and adverb scores (a set of positive and negative for both) was used to calculate the scores for the 2,083 opinions already classified by human experts as a positive, negative or neutral opinion. A parser was implemented in Python that traversed through the opinions searching for the adjectives and adverbs in each opinion and then assign positive and negative scores for the lexicon words in that opinion (Appendix C). These total positive and negative scores calculated for all the

adjectives and adverbs in an opinion were used as the input vector for that opinion in the classification analysis. The most commonly used supervised classification algorithms in sentiment analysis are Naïve Bayes and Support Vector Methods. The Naïve Bayes algorithm is the most widely adopted and it is a simple but effective supervised classification method (Medagoda, Shanmuganathan, & Whalley, 2013) and was therefore used as a benchmark method. A Support Vector Machine (SVM) was also tested in this study because it has been reported to be a more efficient algorithm than Naïve Bayes in English sentiment classification (Taboda, Brooke, Tofiloski, Voll, & Stede, 2011). A decision tree method was also selected, namely J48. The J48 algorithm was used in order to extract rules for the classification of the opinions using adjectives and adverbs. All three experiments/methods were undertaken using WEKA a free open source data mining tool (Hall, et al., 2009). The parameters for Naïve Bayes algorithm for WEKA set to default except the "useKernelEstimator" and it was set to true. The kernel estimator is used for numeric attributes rather than a normal distribution. For the SVM classification algorithm, the "KernelType" parameter was set to linear, and the "probabilityEstimates" enable in order to generate probability estimates for the classification. All other parameters set to default. The J48 algorithm kept as default.

The experiments performed 10-fold cross validation with bag-of-word features for classifiers; Naïve Bayes, SVM, and J48. In 10-fold cross validation, the data set is divided into 10 folds and train on 10 sets then test on one set. Finally, the mean accuracy is presented. The advantage of this method is that all occurrences of the data set are equally used in both training and testing.

In the first attempt (approach 1) at classification, Naïve Bayes, SVM, and J48 algorithms were tested for three classes Positive, Negative, and Neutral. The accuracy, precision, recall and F-Measure are given in Table 5.3. The accuracy for all three classification methods was lower than expected, and they were, in fact, less than the benchmark values reported in the literature for English languages and other Asian languages.

Table 5.3: Classification Accuracies for approach 1

	Classification Method		
	Naïve Bayes	**J48**	**SVM**
Accuracy (%)	39	41	39
Precision	0.287	0.335	0.270
Recall	0.391	0.412	0.398
F-Measure	0.295	0.347	0.235

The confusion matrix reveals that the Neutral (O) category has the poorest classification rate of the three. These matrices are presented in Table 5.4.

Table 5.4: Confusion Matrices for Naïve Bayes(a), SVM(b) and J48(c)

(a)			
	P	N	O
P	121	0	622
N	140	693	2
O	79	426	0

(b)			
	P	N	O
P	421	1	321
N	458	376	1
O	291	214	0

(c)			
	P	N	O
P	558	1	184
N	534	300	1
O	390	115	0

A second experiment (approach 2) was carried using a binary classification approach in which the opinions were trained and classified as either positive or negative. As for the first approach Naïve Bayes, J48 and SVM methods were tested using the same 1583 opinions. The results are given in Table 5.5.

Table 5.5: Classification Accuracies approach 2

	Classification Method		
	Naïve Bayes	**J48**	**SVM**
Accuracy (%)	60	58	56
Precision	0.593	0.581	0.541
Recall	0.598	0.577	0.55
F-Measure	0.538	0.578	0.412

Improvements in accuracies were observed across all three algorithms using this second experiment using a binary classification approach. However, the F-Measure is still lower than 50% for classification using the SVM algorithm. A visualization of the J48 decision tree is provided in Figure 5.2.

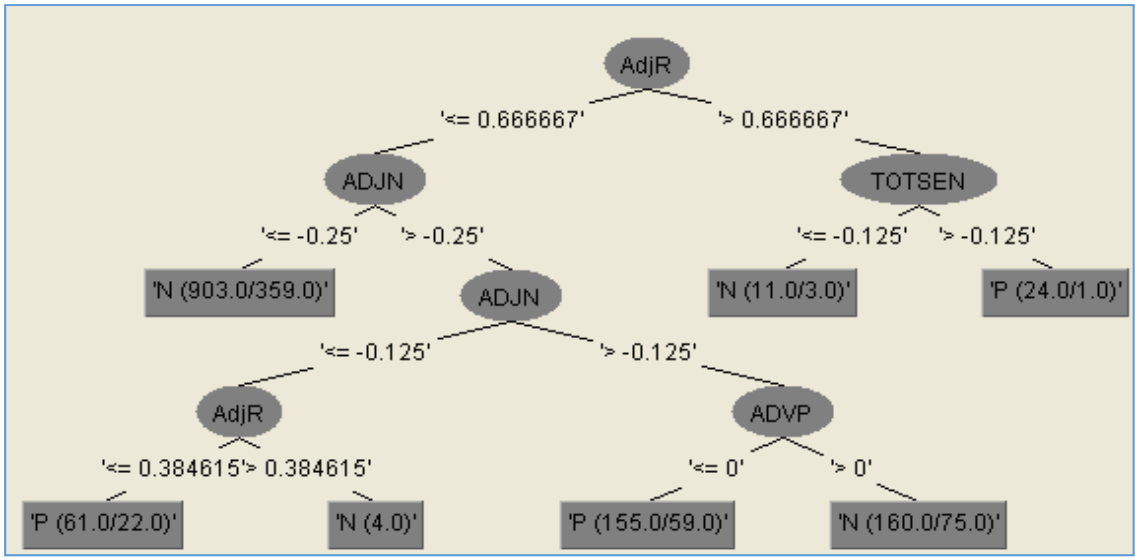


Figure 5.2: Rules generated by J48 for the binary positive (P)/negative (N) classification. AdjR is The ratio of the adjectives to the total words, TOTSEN is Total sentiment score, ADJN is adjective negative score and ADVP is adverb positive score)

Even though the accuracies of the three classification methods investigated were less than expected, the approach can be further optimized to improve the accuracy obtained in this initial investigation. Moreover, the experiments achieved reasonably promising results and can, therefore, be used in this study as the baseline approach for sentiment classification and sentiment lexicon building. The results also indicate that using cross-lingual techniques for the Sinhala language are feasible without applying more sophisticated methodologies, such as a graph or linguistic theories. Additionally, the novel lexicon construction method developed for this research has been found to have good coverage and is capable of retrieving over 72k different entries for different glosses (context) and the main part of speeches. In refining for adjectives, over 10k entries were obtained for various contexts. The coverage for adverbs across different contexts was more than 1,300. In further selection, more than 5,500 adjectives and 405 adverbs were collected in the lexicon. These two lists were compared with a unique adjective and adverb list compiled by language technology the research group at University of Colombo using 10 million word corpus (http://www.ucsc.cmb.ac.lk/ltrl/). It is observed that the collected adjective list covered 80% of the unique adjectives that extracted from 10 million word corpora of Sinhala text. Also noted, more than 60% of unique adverbs are included in the new lexicon that are more frequent in the 10 million word corpus. These figures reveal that the sentiment lexicon constructed using the novel cross-lingual approach developed in this research represents words (adjectives/adverbs) adequately for the Sinhala language. Sentiment score and part of speech are the only information for a word in this lexicon, and this is considered to be

the minimum information that is needed. The researcher believes that the lexicon can be improved the by adding more linguistic knowledge such as gloss. Translation based methods are used extensively in baseline approaches for many non-English languages such as Sinhala in the absence of WordNet type repositories. Therefore, the cross-lingual method deployed for the Sinhala language in this experiment can be justified as it is the first baseline method for this language and its applicability has been demonstrated by the promising classification accuracies.

It should be noted that some contextual and classification features that are not accounted for in the evaluation may effect on the accuracies. As an example, the negation of phrases that contain two or more words with negative meaning has not been considered in this approach. For example, the phrase like "වැරදී නෑ" meaning "not wrong" (i.e. correct), gives a total negative score if the individual sentiment scores of the two terms are assigned to the weight vector. But, as a multiword expression this is actually a positive expression. Handling such cases of negation should improve the classification thereby effectiveness of the lexicon can be further justified. Some inaccuracies present in the generated subjective lexicon scores may also affect classification accuracy. For example, the word "ඉහළ" was mapped to "above" with a negative score of 0.125. It can be argued that this word might have a negative orientation in certain contexts. However, it would be positive in most sentences.

In this first attempt of the sentiment analysis in Sinhala acceptable results (the best being 60% which was achieved for the binary classification approach using Naïve Bayes classification) have been achieved using the available resources such as SentiWordNet. A reported bench mark accuracy of 69% (Ohana & Brendan, 2009) was achieved in similar work for English. It is likely that similar results can be reached for Sinhala by considering a more linguistically complete classification approaches such as negation detection and feature selections. Even though the method developed here is similar to translation methods, in this novel approach direct translation is not used as the underlying mechanism.

5.6 Generating Positive Negative word list using Affixes

The approach for sentiment lexicon construction presented in the previous section is based on several external resources and assumptions. A key assumption was that the sentiment scores for the Sinhala words are taken from the English lexicon but is unlikely that these values truly represent many of the Sinhala sentiments. An alternative mono lingual approach is therefore proposed in which does not depend on any of the external information.

Taking this mono-linguistic approach, a sentiment lexicon was developed using the linguistic features of the Sinhala language. The experiment was conducted in three steps:

1. Identify the affixes to list the positive and negative words
2. Use a dictionary to extract the overtly marked words and their synonyms
3. Evaluate the constructed positive/negative list.

5.6.1 Morphological features of a word

Words are not the basic units of the meaning. Words are comprised of basic units called morphemes (Vikram, 2013). Morphemes are the meaningful morphological units of a language and cannot be further divided. Free morphemes are words that can stand alone, and if bound, can appear as a part of larger polymorphic words. Morphemes cannot be arbitrary joined to form a word. There are definite patterns of combinations of morphemes to form meaningful words. One of the patterns for making a word is known as affixation. An affixation is a process of forming words by adding affixes (Umera-Okeke, 2007). Affixes also referred as bound morphemes and can be a prefix, suffix or infix. The affixation process can be illustrated as follows;

(a)	Verb (V) + *-able*	→	Adjective (A)
	e.g. Predict + *-able*	→	Predictable
(b)	Verb (V) + *-er*	→	Noun (N)
	e.g. sing + *-er*	→	singer
(c)	Un + Adjective (A)	→	Adjective (A):
	eg. un + happy	→	unhappy

In this research, we are interested in the third (c) category where the polarity of the adjective is changed from positive to negative by affixation. In marking theory, overtly marked words are the terms to which the meaning of the word can be changed by adding affixes. The word “unhappy” is the word marked by the prefix “un”. The marked word “unhappy” is deemed negative, and the unmarked word ‘happy” is positive. These patterns of affixation to form positive or negative words led to the definition of a new method of extraction for a list of positive and negative words from a dictionary. In some languages, including English, the words

with prefixes are negative in polarity. For example, the words with the prefixes “dis”, “im”, “in”, “mal”, “mis”, “non”, “un”, “ill”, and “ir”, are all negative and the list generated with suffixes “less” and “ful” also only contains words that are negative (Mohammad, Dunne, & Dorr, 2009). It was suggested that this approach could be generalized to any language in order to build positive and negative word lists as there are many languages with the prefix patterns for negative concept generation (Mohammad, Dunne, & Dorr, 2009). Therefore, for the first time, this approach was adapted and a novel algorithm for generating positive and negative word lists for the Sinhala language was developed. This approach was then further extended to include morphemes, something which to the author’s knowledge has not been previously attempted.

The first step in this algorithm involved extracting all the words with prefixes or suffixes from the dictionary. To undertake this, step a parser was developed in Python (see Appendix D) for the code listing. For this process, it is essential to have a preexisting list of prefixes and suffixes. The extracted list of words by this step/parser can be then be labeled as either negative or positive.

After generating the list of words with affixes, using the first parser, the positive (or negative) list is generated by removing the prefixes and suffixes. The steps for this proposed method are given in figure 5.3.

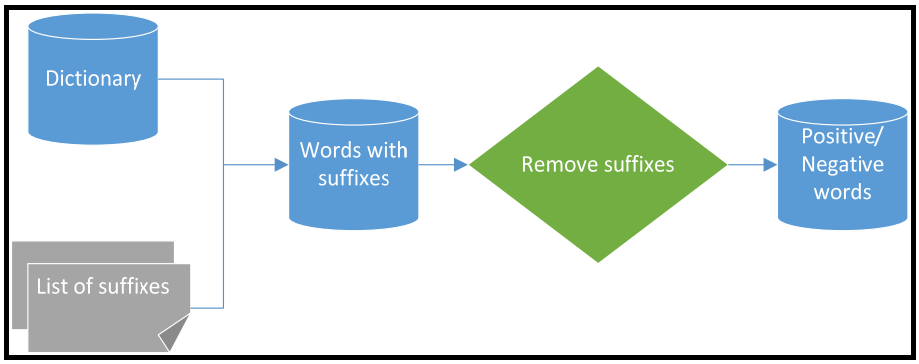


Figure 5.3: Generating Positive/ Negative words

According to the ancient Sinhala manuscripts “Sidatsangara” and “Viyakarana Vivaranaya”, there are 20 prefixes known as “උපසඟර් (upasarga)” in the old version of Sinhala (Dissanayake, 2014). Of these 20 only six prefixes form negative words when added to a word. In the modern Sinhala language, such a list of affixes consists of 17 morphemes (Dissanayake, 2014) which cannot be used by themselves all alone but form a negative word after adding to a positive base form. The identified list of affixes are: “අ”, “අව”,“ අන”, “වි”, “නො”, “කු”, “නිර්”, “නිශ්”, “නිෂ්”,

“නිස්”, “නු”, “දු”, “දුර්”, “දුශ්”, “දුෂ්” , “දුස්” and “නි” (Dissanayake, 2014, Kariyawasam, 2013). A word “සතුට” (happy) when used with the prefix “අ” forms the negation word “අසතුට” (unhappy) — this is an example on how affixation generates a negative word in Sinhala.

The function of some affixes in Sinhala do not always generate the negative word. In some cases, it can revert to either a positive or a negative polarity when forming a new word. Adding the prefix “නි” to some words generates a positive word otherwise is a negative word. For example, නිකැලැල් (Unblemished) is a positive word formed from the negative word කැලැල් (having black mark) by adding the prefix “නි”. While නිගරු (dishonoring) is a negative word derived from the positive term ගරු (respect) in a similar way. Interestingly, the prefixes” නිශ්”,” නිෂ්” (similarly” දුශ්”,” දුෂ්”) make similar utterances but different rules, apply in forming the words. In general, ordinary writers wrongly use both phonemes interchangeably. Therefore, it is necessary to carefully consider both cases when extracting the positive (negative) words. A positive word can be formed with the prefix “නිශ්” but the word formed by the prefix “නිෂ්” is always a negative. An example of both cases is given in table 5.6.

Table 5.6: Polarity changes

Prefix	Base Word	Inflected Word	Polarity
නිශ්	වල (floating)	නිශ්වල (fixed)	Positive
නිෂ්	ප්‍රයෝජන (advantage)	නිෂ්ප්‍රයෝජන (disadvantage)	Negative

After observing the above variations, further investigation was carried out to understand the complexity of the 17 Sinhala affixes (Dissanayake, 2014) when forming polarity inverted words. Via a systematic study, a generic taxonomy was defined for this research in order to categorize the affixes based on polarity changes. Four main phenomena were observed. Initially, the affixes that change the polarity acting as prefix or suffix are identified. Surprisingly, no suffixes were found in the Sinhala language that changed the polarity. There are derivational suffixes that formed the words morphologically, but no change in polarity was identified.

Similar behavior is observed for the English language as well where suffixes do not change the polarity of English words. In English, adding suffixes “-ness”, -less” and “-ful” derive a new form of words, but no polarity change occurs in all cases. The function of “-ness” as suffix always derive a noun, and it retains the polarity. For example, the polarity of “darkness” is

same as “dark”. Although the function of the suffix “-ful” is similar to “-ness” but it always forms an adjective. The polarity of the Words formed by adding suffix “-less” is more complex than for “-ness” and “-ful”. As in the case of adding “-ful” the suffix ‘-less” also produces an adjective. The polarities of “harmless” and “heartless” are opposite even though both are formed by adding suffix “-less”. The polarity of “harmless” is positive and “heartless” is negative. A rule can be established to cope with this issue because if “-less” is added to negative polar stem it will always generate a positive word.

Having established that none of the suffixes in Sinhala can change the polarity of words, next the author studied only prefixes. The Sinhala prefixes were categorized as either “monopolar-prefixes” or “bipolar-prefixes”. Monopolar-prefixes are the prefixes such that when added to a word changed that words polarity to either positive or negative exclusively. This group was then further classified to be either purely positive or purely negative. Table 5.7 shows the classification.

Table 5.7: Monopolar-Prefixes

Prefix	Example	Polarity
අ (a)	**අ**දක්ෂ (incompetent)	-
අව (away)	**අව**ලස්සන (ugly)	-
නිර් (nir)	**නිර්**දෝෂී (innocent)	+
නිශ් (nish)	**නිශ්**චල (clam)	+
නිෂ් (nish)	**නිෂ්**ඵල (in vain)	-
නිස් (nis)	**නිස්**සාර (frivolous)	-
නු (nu)	**නු**සුදුසු (bad)	-
දු (du)	**දු**බල (weak)	-
දුර් (dur)	**දුර්**ගන්ධ (odor)	-
දුශ් (dush)	**දුශ්**චරිත (bad character)	-
දුස් (dus)	**දුස්**සීල (disregard)	-

Prefixes that can generate both positive and negative polarity bearing words were called “bipolar-prefixes” and are listed in Table 5.8. Some bipolar-prefixes were noted to form more positive than negative sentiments, and vice versa.

Table 5.8: Bipolar-prefixes

Prefix	Example	Polarity
වි (vi)	**වි**රාගී (asexual)	+
	විරූපී (ugly)	-
නො (no)	**නොසතුටු** (not happy)	-
	නොවෙනස් (steady)	+
නි (ni)	**නි**වැරදි (correct)	+
	නිරස (boring)	-

A novel rule-based algorithm was developed to extract the morphologically derived words by above affixes from the dictionary. In the first pass of the algorithm, it extracted the term with monopolar-prefixes and grouped them into positive and negative categories according to the polarities denoted in Table 5.7. That is; a list of all positive terms with prefixes (නිර් and නිශ්) were placed into a postive list and rest into a negative list. The words added to the list met a condition that the portion of the word without a prefix should be available in a unique word list when the extracted word is trimed by the prefix. The reason for applying this condition is that removing the prefix does not always result in the extraction of a valid word. As an example "අයතු" is a negative word of meaning "misappropriation" but the word "යතු" is not a meaning full word. In these cases, we keep the negative word in the list but remove the invalid words by searching for the word in a unique list.

A similar extraction process was carried out to collect the words with bipolar-prefixes, but a manual screening was conducted to append them into the positive and negative affixes list. The unique word list considered in this experiment consists of 700k words that are present in a ten million word Sinhala corpus. The pseudo-code for this algorithm is given in figure 5.4.

```
Input: dictionary dic[ ], Prefix list Pre[ ], unique word list uniq[ ]
For each entry x in pre[ ]
        If  x is positive
            For each entry y in dic[ ]
            If (x in y ) and (y-x in uniq[ ]
                y add to positive list
            else discard y
        else y  add to negative list
```

Figure 5.4: Extracting the positive and negative word lists.

After extracting the lists from the dictionary; both positive and negative reverse lists are formed by generating a negative list from a positive list (or vice versa a positive list from the negative list) retrieved by trimming the prefixes. In other words, the negative list is extracted from the dictionary and then the positive words are retrieved by removing the prefixes and vice versa. The basic statistics for the generated list are given in Table 5.9.

Table 5.9: Prefixes and the frequency of words

Prefix	**# Negatives**	**# Positives**
අ (a)	217	215
අව (awa)	15	14
අන (ana)	4	3
වි (vi)	30	30
නො (no)	59	59
නි (ni)	14	12
නිර් (nir)	16	16
නිශ් (nish)	5	5
නිෂ් (nish)	3	3
නිස් (nis)	1	1
නු (nu)	7	7
දු (du)	1	1
දුර් (dur)	2	2
දුශ් (dush)	0	0
දුෂ් (dush)	1	1
දුස් (dus)	2	2
Total	**377**	**371**

By visual inspection, the author found some mismatching cases that do not follow the polarity switching rule explained above. In this mechanism, words extracted by removing the prefix are not always negative or positive. Initially, 250 words with prefix "අ" were extracted from the dictionary. Some of these words with "අ" were not negative words. For example, the word with prefix "අ"; "අගල" is a noun meaning stream, not a negative sense word and by removing the prefix the extracted word, "ගල" is not a positive word either. However, the word "ගල" is a meaningful word (stone) which is available in the unique word list. In this list of 250 words retrieved with prefix "අ "only 33 words were found to be causing confusion. For other prefixes, this mismatch is found only in one or two words. These confusing words can be cleaned manually. For some morphemes such as "කු", neither positive nor negative words are found in the dictionary. But according to Kariyawasam (2013) there should be four cases in the positive and negative lists. It is noted that most of these words are of less frequent usage in everyday language. In addition to above mismatch, in this process, another interesting behavior of the positive and negative word list has been observed. As expected, the words with prefixes always led to negative in the case of building positive/negative word list. However, there are some cases that the rule is reversing and making the opposite polarity of the word. As an example, the word with prefix "නො" the word "නොවෙනස්" meaning "rigid" may be positive but when "නො" is removed, a word of negative polarity is generated. This pattern has been observed in almost all words retrieved by the prefix "නි". The words "නිසරු (unfertilised)", "නිවට (fearful)" and "නිග්රහ (insult)" only makes the negative sense by adding the prefix "නි".

The constructed positive list consists of 371 words of single entries. i.e., no synonyms were associated with each word. To generate the synonyms for each of these words, the thesis again used the dictionary for mapping the corresponding English word, and we retrieved the synonyms for each entry of the list. Similarly, we applied the same procedure to generate synonyms for a negative list. A part of the lexicon generated by this method is listed in Table 5.10.

Table 5.10: Sample of Positive/Negative Wordlist

Negative				Positive			
Word	Synonym1	Synonym 2	Synonym 3	Word	Synonym1	Synonym 2	Synonym 3
අප්රිය	නොරිසි	අසාත	අරුචිය ඇති කරනවා	ප්රිය	පිළිගතමනා	දයාබර	හිතෛෂී
අප්රියජනක	නුරුස්නා			ප්රියජනක			
අවලංගු	අභාවප්රාප්ත	නෂ්ට	න්යෂ්ට	වලංගු			
අසාථර්ක	ලාභයක් නැති	නිෂ්ඵල	වැදගැම්මකට නැති	සාථර්ක	ජයග්රාහී	සපල	
අසම්පූණර්	අසම්පූණර්	නිෂ්ඵල	පළුලු සහිත	සම්පූණර්	අවසන් කළ	නිම කළ	අවසන් කරනවා
වංක	නැමුණු	නීති විරෝධී	හුරුවක් ඇතිව	අවංක	නියම	හොඳ හිතින්	අව්යාජ

For the purpose of score based sentiment analysis, it is necessary to have a lexicon with polarity scores and POS tags. The polarity scores and POS tags were extracted from the SentiWordNet3.0 lexicon using the same English word mapping method described in section 5.5.

5.6.2 Evaluating the generated list

This thesis suggests two methods for evaluating the constructed positive/negative word list. The first method is to compare the list with those already published and the second is to use the word list in supervised machine learning algorithms and compare the classification accuracies.
A. Using Expert knowledge.

The thesis examined the seminal publications by Dissanayake (2014) and Kariyawasam (2013) on the topic. The comparison of a number of positive and negative words reported in each publication with our method is presented in the table 5.11.

Table 5.11: Comparison with available publications

(Our: Proposed method, D: (Dissanayake, 2014), K: (Kariyawasam, 2013))

Prefix	**# Negatives**			**# Positives**		
	Our	**D**	**K**	**Our**	**D**	**K**
අ (**a**)	217	58	91	215	58	91
අව (**awa**)	15	10	12	14	10	12
අන (**ana**)	4	11	17	3	11	17
වි (**vi**)	30	34	14	30	34	14
නො (**no**)	59	74	-	59	74	-
නි (**ni**)	14	16	12	12	16	12
කු (**ku**)	0	4	4	0	4	4
නිර් (**nir**)	16	32	8	16	32	8
නිශ් (**nish**)	5	3	3	5	3	3
නිෂ් (**nish**)	3	3	2	3	3	2
නිස් (**nis**)	1	1	1	1	1	1
නු (**nu**)	7	16	4	7	16	4
දු (**du**)	1	-	5	1		5
දුර් (**dur**)	2	25	8	2	25	8
දුශ් (**dush**)	0	-	1	0	-	1
දුෂ් (**dush**)	1	-	-	1	-	-
දුස් (**dus**)	2	-	2	2	-	2
Total	**377**	**287**	**184**	**371**	**287**	**184**

For all, except one of the prefixes (e.g. අ (a)) the number of positive and negative words extracted using our method is reasonably close to those reported in the two publications. Typically, our list has slightly more words because in our list most of the morphological forms of a given the word are available. For instance, words with close meaning but different grammatical forms “අසංවිදී” (no sense) the noun form and “අසංවිදීනාව” (no sense) the verb forms are included in the list. The prefix “අ” was the prefix which resulted in the least agreement with the published word lists. It is belived that prefix “අ” more tend to generate negative words than the other prefixes. The native speaker can make negative words more comfortably using “අ” prefix.

B. Evaluation of Machine learning methods

A bag of word method is proposed to test the generated positive, negative word list for sentiment classification. Initially, positive and negative words in the opinions sample were extracted using the generated lists. Two thousand and eighty-three opinions from various domains were evaluated in the experiment. The details of the sample used are explained in chapter 4.

The frequency of each positive and negative word is calculated, and their distributions are shown in the figures 5.5 and 5.6. The positive list consists of 221 words with the highest frequency being 216. The highest occurrence of a negative word is 76.

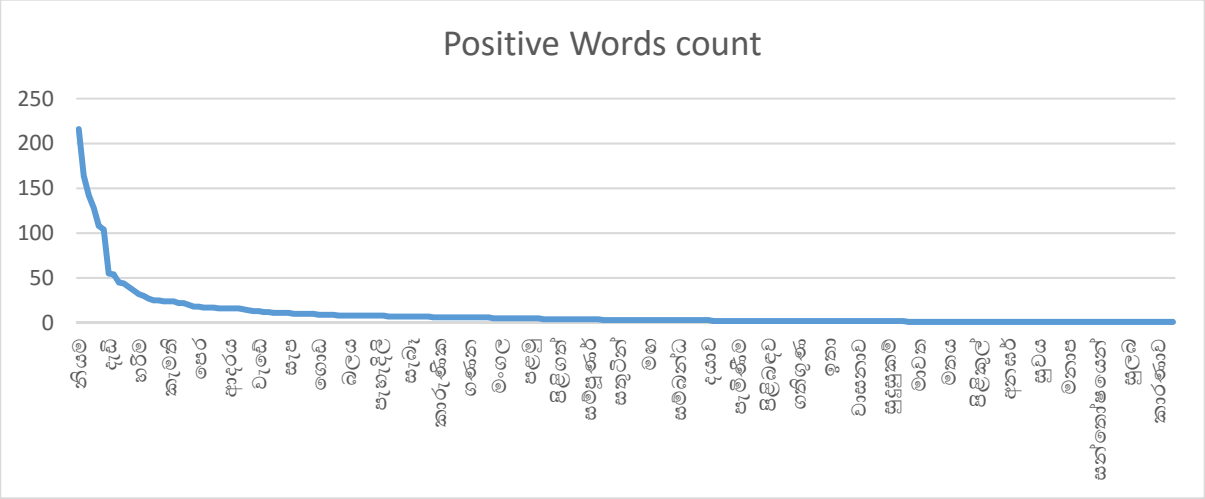


Figure 5.5: Positive word distribution

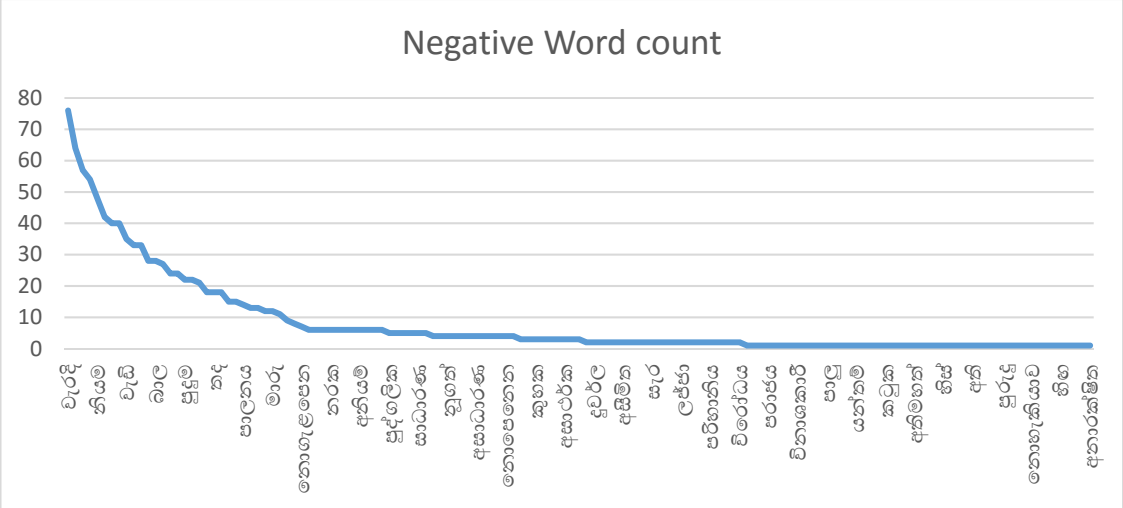


Figure 5.6: Negative word distribution

Both frequency distributions show an approximately power law type pattern, but there is no evidence of the data conforming to the Zipfian Law where the rank of the word is inversely

proportionate to its frequency (Piantadosi, 2014). The next constructed the feature vector. The feature vector for the classification was initially built by using all of the positive and negative words. Then the binary feature vector was classified using several machine learning algorithms, starting with the three class problem: positive (P), negative (N) and neutral (O) classes. The classification accuracies and other evaluation measures are given in the table 5.12.

Table 5.12: Classification accuracies for three classes

	Classification Method	
	Naïve Bayes	**SVM**
Accuracy (%)	45	46
Precision	0.47	0.46
Recall	0.53	0.65
F Value	0.50	0.54

The experiment results show that the constructed positive, negative word list does not significantly improve the classification accuracies when compared with the cross-linguistic approach presented in Section 5.5.1. (Table 5.4). It is also notable that there is no significant difference in classification accuracies between Naïve Bayes and SVM. A possible reason behind the poor accuracies may be due to the influence of the neutral class (O). In the next test, the neutral class was dropped and the opinions classified using the same methods, data, and parameter settings, but this time a binary classification was undertaken: Positive (P), and Negative (N).

Table 5.13: Classification accuracies for two classes

	Classification Method	
	Naïve Bayes	**SVM**
Accuracy (%)	58	59
Precision	0.59	0.60
Recall	0.57	0.69
F Value	0.59	0.65

According to Table 5.13, a significant improvement in accuracy was observed for the two class problem. However, even the binary classification though did not reach the gold standard; the results seem to suggest that in the case of Sinhala a binary classification is a necessity as is

reported to be the case of most languages. To further improve the classification accuracy for a two class problem a number of different experiments were undertaken. These experimental approaches included testing of different feature representations namely, unigram, bigram and trigram, and testing of different weighting measures such as frequency and tfidf. Higher order n-grams were not considered because of the poor results obtained using trigrams and because work reported in the literature has found that bi-grams are normally sufficient. The results of the experiments for unigram and bigram feature representations are provided in table 5.14. No significant improvement was achieved by changing the features from unigram to bigram. However, bigram features gave slightly better classification results than unigram features shown in the case of the Naïve Bayes classifier. Overall SVM gave better or the same performance as Naïve Bayes regardless of feature vector or weighting measure.

Table 5.14: Classification accuracies for different features and weighting measure combinations

Features	**Naïve Bayes**				**SVM**			
	Accuracy	**Precision**	**Recall**	**F-measure**	**Accuracy**	**Precision**	**Recall**	**F-measure**
Unigram+ frequency	56	0.558	0.560	0.558	57	0.565	0.567	0.565
Unigram + tf-idf	55	0.545	0.553	0.518	60	0.607	0.603	0.579
Bigram + binary	58	0.620	0.580	0.496	58	0.612	0.576	0.489
Bigram+ frequency	57	0.654	0.573	0.480	58	0.669	0.583	0.497
Bigram + tf-idf	58	0.576	0.581	0.574	58	0.594	0.583	0.54

When all test cases are considered classification was better for negative opinions than for positive ones. The confusion matrix for the best test case (SVM with unigram features and tf-idf weightings) which, achieved 60.2% classification accuracy and is provided in Table 5.15.

Table 5.15: Confusion matrix

		Predicted		
		N	P	Total
Actual	N	439	101	540
	P	296	163	459
	Total	735	264	999

According to Table 5.14, more than half of the positives were classified as negatives. On the other hand, the error classification for negatives was 19%. This observation reveals that negative words derived by the proposed method are sufficient for sentiment classification. The above classification was carried out proves the appropriateness of the positive and negative words by the proposed method using of basic frequentist techniques. It can be assumed that the classification accuracies can be improved if advanced linguistic features incorporated in the classification method. In conclusion, the generated positive and negative word lists by morphological approaches can be considered as basic lexical resources that can be compiled independently without using any external resources or methods. This list can be used as a seed list for generating lexical resources or for any other sentiment classification purposes for Sinhala.

5.7 Graph based method for Sentiment Lexicon construction

There are several weakness and dependencies when constructing a sentiment lexicon using cross-lingual approaches and building positive/negative words lists as described in the previous sections. In these approaches, the main challenge was to construct lexical resources for a language that have sufficient coverage of terms for the purpose of sentiment classification regardless of domain. In the previous experiments, the sentiment lexicon constructed was dependent on a foreign language (English) and hence the approach was based on three central assumptions which were not necessarily correct. On the other hand, the positive/negative list retrieve using affixes does not cover all words and therefore leads to less than optimal classification accuracies. In this section, a graph based approach to constructing a sentiment lexicon with polarity scores for the positive and negative word list generated is proposed.

The most popular lexical resource used for a graph or network based sentiment lexical construction is WordNet (Miller, 1995). The primary requirement for graph based lexicon

construction is the semantic relatedness between words. WordNet provides a set of relations between two words from simple synonymy to advanced relations such as troponyms. Kamps et al. (2004) built a graph based on WordNet synonyms to construct a polarity lexicon using the shortest path between any given word and a seed positive or negative word. The relative distance of a word referring two seed words is calculated by dividing the difference of distance from given word to seed words by total distance between seed words. In addition to using a synonym relationship, Hu and Liu (2004) used antonyms to construct the lexicon by predicting the polarity using a labeled list of seed words. Instead of using shortest path Hu and Lu applied a bootstrap propagation to collect synonyms for a set of given words with known polarity. If synonym relation found between seed word and other, they labeled as a synonym of the words otherwise checked for antonyms. Kim and Hovy (2004) used WordNet to expand positive and negative word lists. In their approach synonyms of positive words and antonyms of negative words were assigned as positive. The negative list was expanded by considering the synonyms of negative words and antonyms of positive words. In all of these studies, the use of word relatedness is important in the WordNet type lexical resource. However, only 2% of all languages have a suitable WordNet available. But, currently, there are more than 7,500 online dictionaries and glossaries available. Any dictionary contains basic synonym and antonym relations. The following section presents a novel approach which utilises a dictionary and its primary contents to construct a sentiment lexicon for the Sinhala language.

5.7.1 Basic Graph Theory

Graph theory is the study of *graphs*. Graphs are structures made up of objects (nodes) and their relationships (edges). A graph G can be mathematically represented as G = (V, E), where V is a non-empty finite set of elements called vertices (nodes) and E is a finite set of distinct unordered pairs {u, v} distinct elements of V called edges. The edge(s) can be directed or undirected. Edges represented by arrows represent a directed graph.

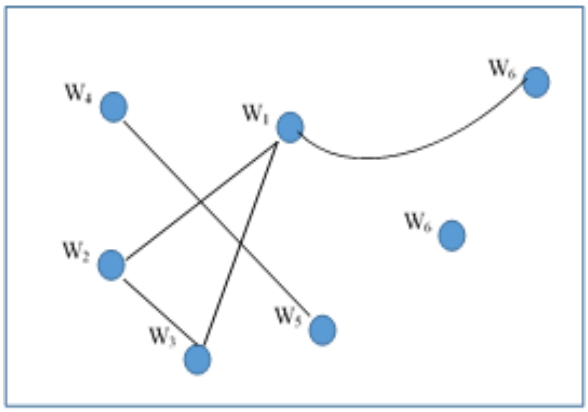


Figure 5.7: Example of an undirected word graph comprised of 6 (w_i) vertices

The path of a graph is defined as an alternating sequence of distinct vertices and connecting edges. The length of a path is the number of edges in the graph. The shortest path between any two vertices is the one with the smallest distance compared when compared with all other paths between the same points.

5.7.2 Building word graph from the dictionary

An undirected word graph of synonyms is constructed where two nodes (words) are linked by the synonym relation. This research defines following terms to explain the novel algorithm developed for the graphical lexicon construction method.

Dictionary Entry: the collection of single or multiword expressions starting from the main entry. The following figure (see figure 5.8) provides as an example of a dictionary entry.

Figure 5.8: An example a single dictionary entry

In this example, the dictionary has four synonyms for the word දක්ෂ (clever). The relationship between each word in this dictionary entry is defined as an *inter-entry*.

Two different dictionary entries are connected by an *outer-entry* relationship as illustrated in figure 5.9.

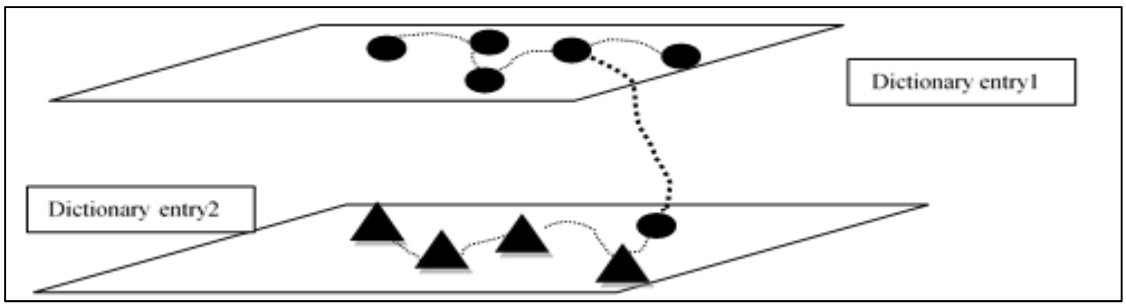


Figure 5.9: Relationship between two different dictionary entries

The structure of the dictionary entries is illustrated in the following diagram (figure 5.10).

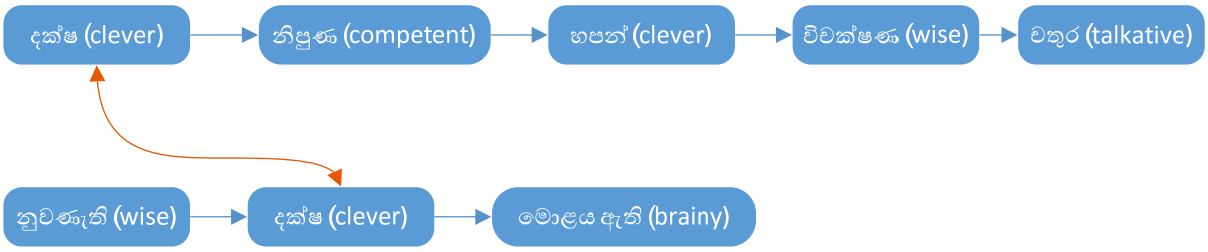


Figure 5.10: Example of dictionary entry relationship

From these described patterns which exist in a dictionary, it is possible to extract a synonym or synonym of the synonyms of a given positive or negative word. Therefore, in this proposed method a word graph is generated using dictionary words and the semantic relatedness between dictionary entries. That is all words are connected by the synonym relationships of the words. The resulting structure is a graph, where the vertices are dictionary entries and the connections between each pair of synonymous words in the dictionary represent the edges.

Then a lexicon was generating from the word graph G using label propagation. Label propagation is semi-supervised learning algorithm that adds a label to unlabelled data using the semantic relationships between large numbers of data points (Covell & Baluja, 2013). An illustration of the proposed novel algorithm is presented in figure 5.11.

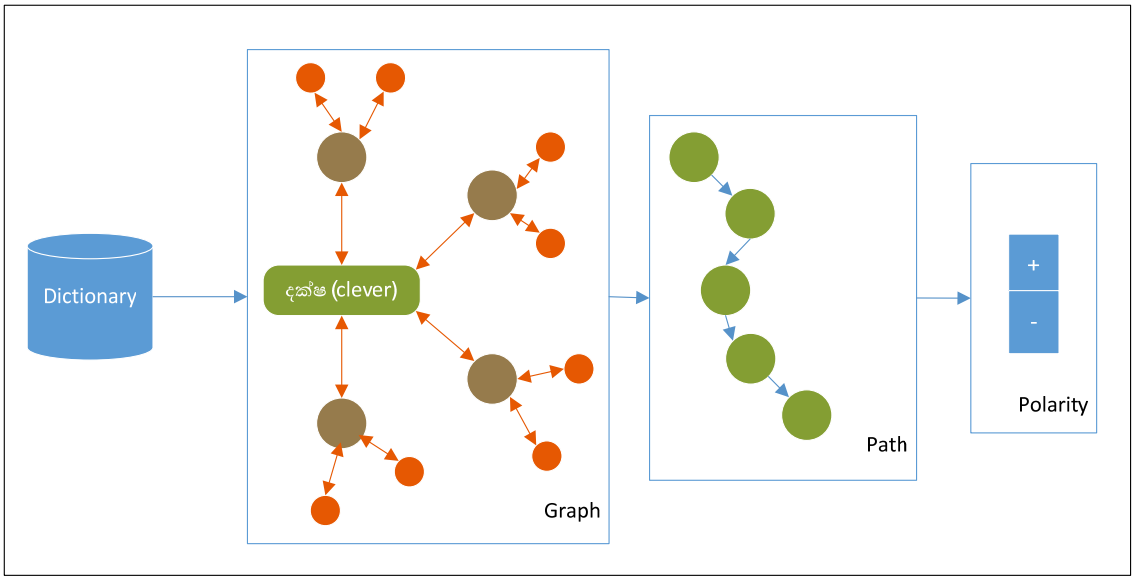


Figure 5.11: Graph based lexicon construction

In the label propagation the seed words of positives were given +1 and negatives were given -1. Starting with an unknown polarity word W_i, all the shortest paths starting from the word are constructed. The path not only included the synonyms of the word but antonyms of the word as well. This process is illustrated in figure 5.12.

Figure 5.12: Path created using the proposed graph based algorithm

In the above example, the shortest path for the word ආදරණීය (dear love) consists of positive words දයාබර (dear), ප්‍රිය (love) and negative words මෝඩ (stupid), and බොළඳ (immature). The words මෝඩ (stupid) and බොළඳ (immature) are not exactly the antonyms of the word but they, can be explained as having the opposite sense of the word ආදරණීය (dear love). This pattern reveals that a word can be related to its antonyms in a dictionary if it is integrated into a graph. In this thesis's research this graph based relationship is used to assign a polarity score for a word by propagating the word through the all possible shortest paths starting from the word itself. The algorithm propagates through each path by updating the two weights W+ and W-. If the $(i+1)^{th}$ word is positive then W+ is updated by +1 and if the $(i+1)^{th}$ word is a negative word then -1 is added to W- , otherwise 0 is added to both the weights. The average weight is calculated by dividing the weight by the length of the path. If the total average (sum of average score for all paths) W+ is greater than total average (sum of average score for all paths) W-

score, the word is classified as positive otherwise it is a negative. The polarity score for the word will be the relevant W score. The algorithm is provided as pseudo code in figure 5.13.

```
Word Polarity and Sentiment Score using Shortest Path
Requirement: A word relatedness graph G
Given a word W in vertex set V
Define shortest path in the graph and weight W_ij between two nodes i and j
for each shortest path k from with maximum number of steps m
        define W+ and W-
        If positive seed word meet then
                add 1 to W+
        else if negative seed meet then
                add -1 to W-
Calculate average W+ and W-
If W+ > W- then
        Classify W as positive
else if
        Classify W as negative
```

Figure 5.13: Polarity classification algorithm using a graph

The above algorithm resulted in both a positive and a negative score for a word. It was noted that the lexicon that was produced using this algorithm contained some words with zero scores for both the positive (W+) and the negative weight (W-). It is clear that neutral words, with no sentiment, should result in a positive and negative weight of zero. All words in the constructed lexicon with a positive and negative weighting of zero, and therefore an overall score of zero, were neutral words (e.g. "අපි" (we)). Using following equation, introduced the objective score (Wo) for each positive and negative word.

Assume that $Positive(word) + Negative(word) + Objective(word) = 1$

Therefore Wo = 1- (|W+| - |W-|) (EQ 1)

The three scores for a sample of Sinhala words that are obviously either positive or negative words are shown in Figure 5.14 and Table 5.16. The differentiation between the absolute positivity (|W+|), negativity (|W-|) and objectivity (|Wo|) weightings of the words are clearly distinguishable in the radar graph. For example, the word අසතුටු (displeases, unhappy) which is a negative word and was correctly classified by the proposed graph method, which resulted in an absolute negative score of one, and a positive score of zero (see figure 5.14, box 1).

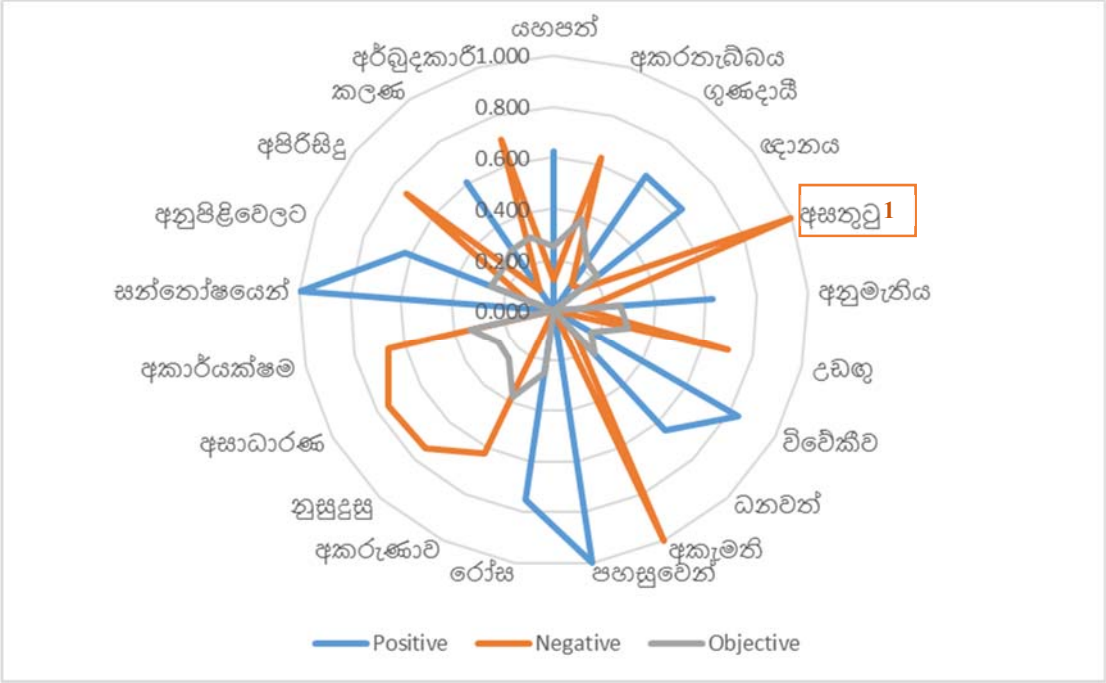


Figure 5.14: Radar chart of a sample of Sinhala word sentiment scores generated using the novel graph method.

On the other hand, word ගුණදායී (healthy) is a clear positive word but the algorithm has assigned 0.643 as positive polarity score and 0.125 as negative and 0.232 objective score (see Table 5.16).

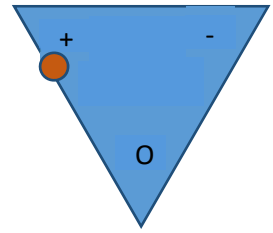
Table 5.16: Sentiment scores for the words

Word	Translation	Positive (|W+|)	Negative (|W-|)	Objective (|Wo|)
යහපත්	good	0.625	0.123	0.252
අකරතැබ්බය	tradagy	0.000	0.625	0.375
ගුණදායී	healthy	0.643	0.125	0.232
ඥානය	knowledge	0.643	0.141	0.217
අසතුටු	displeases	0.000	1.000	0.000
අනුමැතිය	approval	0.625	0.111	0.264
උඩඟු	haughtiness	0.000	0.699	0.301
විවේකීව	leisurely	0.833	0.000	0.167
ධනවත්	wealthy	0.641	0.125	0.234
අකැමති	unwilling	0.000	1.000	0.000
පහසුවෙන්	comfortably	1.000	0.000	0.000
රෝස	pink	0.750	0.000	0.250
අකරුණාව	unkind	0.000	0.625	0.375
නුසුදුසු	unfit	0.000	0.740	0.260
අසාධාරණ	unfairly	0.000	0.748	0.252
අකායර්ක්ෂම	ineffective	0.000	0.667	0.333
සන්තෝෂයෙන්	happiness	1.000	0.000	0.000
අනුපිළිවෙලට	sorted	0.625	0.111	0.264
අපිරිසිදු	dirty	0.000	0.740	0.260
කලණ	friendship	0.610	0.100	0.290
අබර්දකාරී	crisis	0.000	0.700	0.300

It is beneficial to represent the polarity score in three dimensions; positive (+), negative (-) and objective (o) graphically. To represent the polarity scores diagrammatically, a graphical model defined by Esuli and Sebastiani (2006) is used that employs a triangle to visualize sentiment scores across the three dimensions (Figure 5.15).

Figure 5.15: Graphical Representation of sentiment scores taken from Esuli and Sebastiani (2006)

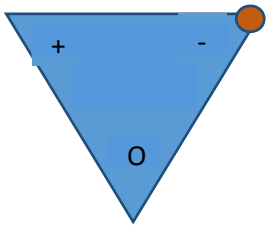
In the subjective-objective dimension (SO-polarity) a word lies somewhere in a continuum of subject to factual. The position of the word in this continuum is used to determine whether the word is subjective or factual. If the word is categorised as subjective, then it can have a positive or a negative polarity (PN polarity). For this research, the objective score (|Wo|) is calculated using the formula given in EQ 1. The points at the corners of the triangle have the maximum score of 1.0 for one dimension and 0.0 for the other two dimensions. The same three dimensional structure was adapted for determining the sentiment scores for Sinhala, and some examples are given in the table 5.16.



P =0.625, N =0.123, O =0.252

යහපත් (good) – Adjective

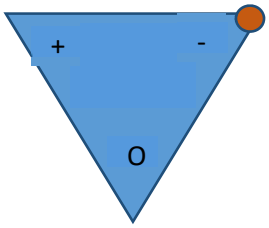
Eg. ලංකාවේ මිනිස්සු හරිම ගුණ **<u>යහපත්</u>** මිනිස්සු කියල්ලා තවත් වරක් ගෞරවයෙන් ලෝකෙටම් පෙන්නුවා. (It is proven that Sri Lankans are good people)



P = 0, N = 1.0, O = 0

අපිළිවෙළට (desultorily) – Adverb

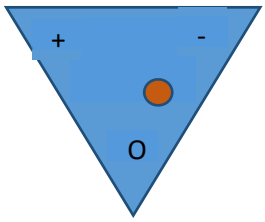
Eg. ඒවා හරියට ***<u>අපිළිවෙළට</u>*** එක පෙළට හැදුණු සෙල්ලම් ගෙවල්, ගවගාල් වගෙයි. Those look like huts that were built in line randomly.



P = 0, N = 1.0, O = 0

අකටයුත්ත (injustice) – Noun

Eg. චන්දිමාල්ට සිදුවූ **<u>අකටයුත්ත</u>** බොහෝ දෙනෙකුට නොපෙනෙන්නේ ඔහු විස්සයි 20 සහ එක්දින තරගවල අසාථර්ක වීම නිසාය. (People did not care about the injustice to Chandimal as he has failed in 20 or so marches)



P = 0, N = 0.623, O = 0.375

අඩුවෙනවා (reduce) – Verb

Eg. මරණ දඩුවම්න් අපරාධ ***<u>අඩුවෙනවා</u>*** කියලා කිසිම තැනක ඔප්පු වී නැහැ. (No evidence of reduction in crimes with capital punishments.)

Figure 5.16: Sentiment scores for the words

The above polarity scores were calculated using the novel graph based approach for and were then used to classify an opinion as either positive, negative or neutral. The polarity scores for each of the unigram features was assigned in the classification vector prior to applying the classification methods. The use of these scores in classification is explained in the following section.

5.7.3 Evaluating the Lexicon generated using the Graph based method

A. Using Expert Knowledge

As for previous experiments, the quality of the lexicon constructed is evaluated using both expert's knowledge and supervised classification methods.

In manual testing, the researcher evaluated how well an expert can categorize a set of words. Initially, a language expert classified randomly selected 200 words which had already been classified using the graph method. Then expert's classification was then compared with the polarity assignment provided by the algorithm using precision and recall values. Table 5.17 illustrate the precision, recall and the F-Measure for Positive(P), Negative(N) and neutral (O) classes.

Table 5.17: Expert and Algorithm comparison

	Recall	Precision	F-Measure
N	0.836	0.4	0.541
P	0.357	0.208	0.263
O	0.495	0.845	0.625

The table indicates that the negative and neutral classes achieved above average F-Measure values compare to positive class. On the other hand, the negative words extracted by the proposed method gave higher recall than others. One of the reason for the low F-Measure by the positive words is it's sample representation. Out of 200 random sample, only 24 positive words were classified by the graph based algorithm.

However, in careful observation of the list generated by the proposed method, it was found that some nouns were categorised as negative. When considering the path generated by the graph algorithm, it was found that this error occurs due to translation errors in the dictionary. Several cases were investigated, the following example illustrates how the word නෑව (nævə, ship) generated a path that leads to misclassification. The dictionary entry for the නෑව is given in figure 5.17.

Figure 5.17: dictionary entry of the word නැව (nævə, ship)

This entry contains the word බඳුන (ba~dunə, vessel) and the word is the translation of English word potty. The dictionary entry of the word "potty" is given in figure 5.18

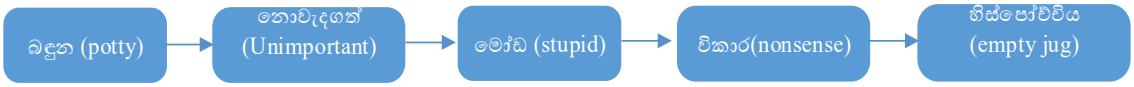


Figure 5.18: Dictionary entry of the word බඳුන (ba~dunə, vessel)

The dictionary entry starting with the word බඳුන (potty) consists of several negative words. In generating the word graph for the word නැව (nævə, ship) it is also connected to the word බඳුන (ba~dunə, vessel) and this results in ship being classified as a negative word.

Having gained knowledge of how general words behave in the proposed method, the next experiment was carried out using the part of speech of the words that were classified by the algorithm. The part of speech or word type mostly determines the sentiment of an opinion as either adjective or adverb or both (Benamara, Cesarano, & Reforgiato, 2007). With this inference, this research was expanded in order to investigate the classification of adjectives and adverbs using the graph based method. An expert manually tagged each random sample of 200 adjectives and adverbs and the annotation was compared with the classification performed by the proposed algorithm. The precision, recall, and F-Measures were used to compare the annotations generated by the two methods. The evaluation measures for adjectives are given in Table 5.18.

Table 5.18: Expert and Algorithm comparison - Adjectives

	Recall	Precision	F-Measure
N	0.943	0.651	0.770
P	0.465	0.833	0.597
O	0.28	0.304	0.291

According to the table, a higher recall for negative class achieved and the F-Measure is greater than other two categories. It also noted that the F-Measure for both positive (P) and negative (N) are greater than 0.5.

A similar experiment was carried out for adverbs and the results of are given in Table 5.19. The both measures; recall and precision show good scores for both categories, positive and negative. As in the adjective classification shown in the table 5.19, the negative class shows grater F-measure compare to the positive class.

Table 5.19: Expert and Algorithm comparison - Adverbs

	Recall	Precision	F-Measure
N	0.797	0.610	0.691
P	0.620	0.754	0.681
O	0.565	0.603	0.583

Above explained manual statistical evaluation method considered only the polarity of the words and not the sentiment score (numerical) allocated by the method. In order to further investigate the effectiveness of the scores assigned by the method the annotated data sets were used as input for rule based and machine learning classification techniques.

B. Heuristic Method

Heuristic or rule based sentiment classification methodologies are popular for the classification of sentiments expressed in morphologically rich languages (Mittal, Agarwal, Chouhan, Bania, & Pareek, 2013). A vector of sentiment scores for the adjectives and adverbs in an opinion was generated using the polarity scores obtained using the graph based method. The tests were conducted on two sets of data; one set consists of opinions drawn from comments on political news items (politics domain specific) and is called the *domain dependant* data set. The second opinion set does not belong to any of the domains, and it is therefore referred to as the *domain independent* data set. In the rule based method the total sentiment score of the opinion was calculated, and if the value was negative, then the opinion was classified as negative otherwise, it is a positive opinion.

The classification accuracies obtained using the rule based approach are presented in Table 5.20. Using this approach, negative opinions were more often correctly classified than positive opinions regardless of whether or not the data was domain dependant. The rule-based method performed better on the domain independent data set than the domain dependant data set.

Table 5.20: Performance of the rule based method using features constructed by the novel graph-based lexicon construction method

Sample	Positive	Negative	Average
Domain Dependent	53%	60%	54%
Domain Independent	56%	63%	59%

C. Machine Learning Method

The vector constructed for the rule based methods was used to classify the opinions using machine learning approach. In this approach, a Naïve Bayes classifier is applied for domain independent data set. The performance measures obtained using this approach are given in Table 5.21.

Table 5.21: Classification performance using a Naïve Bayes classifier

Class	Precision	Recall	F-measure	Accuracy
Positive	0.673	0.455	0.543	-
Negative	0.523	0.730	0.609	-
Overall	0.605	0.579	0.573	58%

Both classification performance (Table 5.20 and Table 5.21) measures reveal that negative opinion detection by the sentiment scores are higher than the positives. The F-measure, which combines both precision and recall, reveals that negative opinion detection is more accurate than positive opinion detection when using the Naïve Bayes method.

The results of all three evaluation methods provide further support for the conclusion that negative polarity scores are more important than positive scores to classify negative comments (see Tables 5.17, 5.18, 5.19 and 5.20). A higher recall value for all the experiments was observed in the negative sentiment words and the negative opinion classification. This evidence leads to the conclusion that the negative polarity scores generated using the graph based approach proposed in this research are much more accurate predictors of sentiment than the positive polarities.

5.8 Chapter Summary

In this chapter novel methods for constructing sentiment lexicons for the Sinhala language were detailed and implemented. The usefulness and adaptability of the constructed lexicons were

critically evaluated using sample opinions. Supervised, as well as unsupervised techniques, were used to assess three lexicons constructed using three different methods. The lexicon constructed by the cross language approach gave promising results in terms of its suitability for classifying positive/negative opinions. The novel approach based on prefixes, uses affixes, to build a positive/ negative word list for Sinhala showed high correlation with expert classifications of synonyms and antonyms. Finally, the graph-based method, primarily used for assigning sentiment scores for the positive/ negative list, gave good scores for negative words and better classifications were achieved for negative than for positive opinions.

This research is the first ever attempt in constructing lexicons for sentiment analysis in the Sinhala language. The constructed lexicons and the findings will make a significant contribution to improving sentiment analysis in Sinhala. In the next chapter attempts to classify Sinhala opinions without dictionaries using frequentist and text classification approaches is presented.

Chapter 6: Sentiment Classification using Text-mining Approaches

6.1 Introduction

In the previous chapter, potential methods for constructing lexical resources and their adaptation for sentiment classification of Sinhala reviews was discussed. The main advantage of using sentiment lexicons is the integration of linguistic knowledge into the classification of opinions, either positive or negative. To achieve the best results, the lexicon should cover the target domain in which the classification will be carried out. In general, as seen in the last chapter, lexicons are based on dictionaries and are therefore limited to the coverage of the dictionary. On the other hand, machine learning based methods provide an opportunity to explore complex patterns and correlations between sentiments and concepts which may exist implicitly in opinions or comments.

This chapter focuses on lexicon independent review classifications using supervised machine learning methods for sentiment classification. The primary objective is to investigate the polarity detection process at both document and sentence level. As this is the first attempt at applying sentiment classification for Sinhala reviews, it is important and essential to experiment at both document and sentence level. Document level sentiment classification has been extensively studied for other languages (Pang, Lee, & Vaithyanathan, 2002, Pang & Lee, 2004). In a recent study Xia et al. (2016) decomposed the document into sub-sentences using different polarity shifters (a lexical item that changes the polarity of a phrase). After that, they trained classification algorithms for each sub-sentence and then used a weighted combination for component classifiers to detect the polarity of the entire document. In this ensemble approach, large weights assigned to the classifier trained on polarity upshifted parts and smaller for the classifier that trained of polarity shifted parts.

Analysing sentiments at the sentence level allow the user to extract more detailed information relating to the sentiments expressed in the review or opinion. Compared to document level classification that includes advanced tasks such as combining sentences that are different sentiments are not required in sentence level classification, which is an advantage. On the other hand, the complexity of making annotated corpora for supervised classification at the sentence level is a drawback. An alternative approach to overcome resource limitations, such as the use of annotated corpora, is applying an unsupervised or semi-supervised approaches. All

experiments are conducted on opinions extracted from an online newspaper site. Complete information on the opinion data is given in chapter 4 section 4.3.

The remainder of this chapter is organized as follows. Section 6.2, provides the problem definition for supervised classification. The experimental method is presented in section 6.3. In section 6.4 the results achieved are explained, and a general discussion is presented in Section 6.5. Finally, Section 6.6 outlines the chapter summary and draws some conclusions.

6.2 Problem Definition

Sentiment classification is a multi-class classification problem that assigns a label to a document from a set of labels. Multi-class classification is based on an assumption that each document is assigned to one and only one label. The problem is formulated by the following definition (Tsoumakas & Katakis, 2008).

Definition 6.1: (Multi-Class Classification) Let $C = \{c_1, c_2, c_3 \ldots . c_n\}$ *be a finite set of n labels and* $X = x_1, x_2, x_3 \ldots . x_m$ *are set of documents that are to be classified. Then multi-class classifier* γ *is a function from X to C,*

$$\gamma: X \longrightarrow C$$

The above type of learning is known as supervised learning as humans define the labels for a set of samples that is used to train the classifier.

The sentiment classification task is a problem of text categorization. The task is performed using various classification methods based on machine learning techniques. The supervised machine learning methods, evaluated in this chapter, are used to classify opinions into either positive, negative or neutral class. The objectives of this chapter can be summarised as follows;

- The initial aim is to examine Bayesian based sentiment classification in morphologically rich languages. In other words, to understand the application of traditional text categorization methods to Sinhala language text.
- Secondly, the best set of features for optimal classification accuracy are investigated. To this end, a comparison of classification accuracy for both statistical and linguistic based features is undertaken.
- The final goal is to compare the results achieved using lexicon based classification (discussed in chapter 5) with supervised sentiment classification. As lexicon based classification requires sufficient and adaptable lexical resources that are costly to

construct, this leads to the question as to whether a lexicon-independent supervised approach can produce significantly better results.

6.3 Experimental Method

The following sections explain the general experiment steps carried out for all experiments in this chapter. Firstly, the thesis discussed the overall training and testing processes for sentiment classification. Each component of the classification process is described extensively.

Traditionally, a supervised classification process includes preprocessing, feature extraction, feature selection, and classifier training using machine learning algorithms. For the training and testing or prediction stage, the steps as given in figure 6.1.

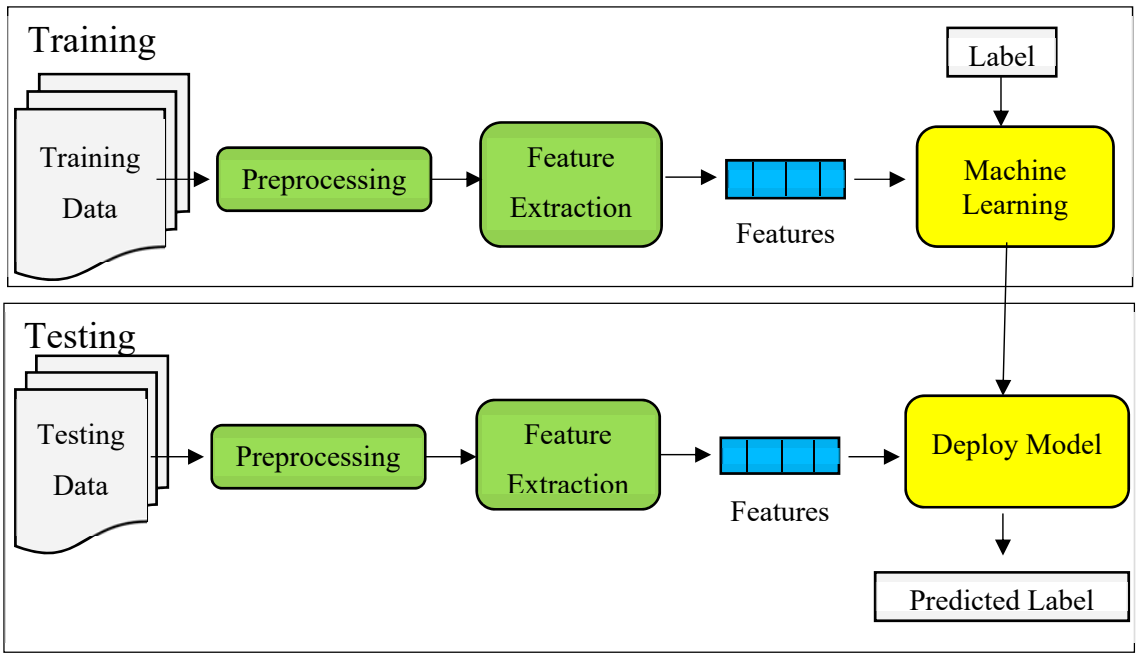


Figure 6.1: Framework used for supervised sentiment classification

6.3.1 Data and Preprocessing

In this chapter, the experiments are performed using a set of Sinhala opinions from newspaper articles. In total, 2,083 comments were extracted from the text data from a leading online newspaper called “lankadeepa” (http://lankadeepa.lk/).

In machine learning, pre-processing describes any type of processing on raw data before it is used as input in the main classification (or clustering) procedure. Commonly used pre-processing methods for text mining are:

i. Data cleaning and noise removal
ii. Stop word removal
iii. Stemming

For these experiments, the critical pre-processing task is that of cleaning the comments. During the cleaning process tasks such as removing all punctuation marks, correcting spelling mistakes, and removing gibberish were undertaken. Punctuation marks have no meaning in unstructured text such as reader comments. Since most of the comments are unstructured texts, the removal of the punctuation marks does not significantly affect the accuracy of the classification. Words without any meaning (gibberish) were removed next, and comments which were written in *transliterated*[14] form were translated into native language.

Stops words in the opinions were removed in the second stage of pre-processing prior to performing feature extraction. The list of stop words incorporated in this step is the same as that described in Section 4.4.1 as explained in Chapter 4. It is assumed that the effects of morphologically inflected words are more important than the root form and thus stemming is not carried out in these experiments.

6.3.2 Feature Extraction

Sentiment classification methods based on vector models always require a vector which represents each opinion or review. The components of the vector are known as features. These features may be a language specific characteristic or statistical information about a feature retrieved from an opinion. In other words, statistical features represent statistical information about features while linguistic information is represented by part of speech and syntactic formation denote the linguistic features. In the document level analysis undertaken in this study, the features are limited to statistical features. These features may be anything from a simple binary feature to complex frequencies. The most widely used approach in document classification is referred to as "bag of words" methods where several terms present in the document combined are considered to be a feature vector (Bharti & Singh, 2014). A relevant set of features always provides useful information that can be used to discriminate between different opinions. On the other hand, irrelevant, redundant or noisy features decrease the accuracy and also increase the computational complexity.

[14] Representing the characters of a given language script by the characters of another language

A word is the basic feature unit for any text classification problem. Therefore, a feature can be a term itself or combined with another term to represent the semantic knowledge of a document. If a single term is used, then it is defined as the unigram feature. Otherwise, there can be a combination of two or more words defined as n-gram. However, unigrams only capture the lexical semantics considering the meaning of the word (Qu, Ifrim, & Weikum, 2010). It is understood that a combination of two or more words changes the sentiment of the opinion (see section 4.5.2). The effect of the combination of words in sentiment classification is known as compositional semantics; the construction of meaning based on syntax. These compositional semantics can be extracted using the combination words known as n-grams. With the aim of understanding the compositional semantics in sentiment classification for Sinhala, experiments were designed to evaluate the effect of n-gram features in polarity classification. To investigate this aspect, a sequence of experiments conducted starting with bigrams and then move on to higher order n-grams.

In the initial investigation, bigrams were extracted after removing the stop words. In this initial set of bigrams, there were bigrams that included greeting words. Some comments for the news articles, which were related to obituary notices of prominent people, tend to contain these greeting words. Because these greeting bigrams contain no sentiment, such bigrams were removed. Subsequently, a classification vector is constructed by selecting a list of bigrams using different thresholds based on bigram frequencies.

6.3.3 Feature Selection Methods for Classification

As explained in the previous section the features are the set of words or words n-gram that are extracted from the opinions. It is essential to include the most relevant and useful features extracted to form classification vectors. The process of selecting such features is known as feature selection or attribute selection. Feature selection methods can be grouped into; filter methods, wrapper methods and embedded methods. Filter methods assign a score to each feature using statistical measures. Features are selected or rejected by the ranking score. Correlation coefficient scores, Chi-squared test and information gain are some examples of filter methods.

Some general feature selection methods have proven to be particularly useful for text classification too. For example, filter method techniques, such as Chi-squared test, information gain, and correlation coefficient scores are very popular and efficient methods for text

categorization (Yang & Pedersen, 1997) . For its simplicity and high interpretability, this research uses filter methods for feature selection.

In additional to above three filter methods, there are many sophisticated feature selection techniques, which have been developed for data mining for general application. For example, CfsSubsetEval selects the feature/ attribute subsets that correlate highly with the class value and which have low correlation with each other (Hall & Holmes, 2003).

Because the focus of this research is to examine the influence of linguistic features on the classification of sentiments in morpheme rich languages rather than on optimising the classification process two of the most commonly used classification algorithms were selected for the experiments – namely Naïve Bayes and SVM. The results of each classification experiment were examined in terms of F-measure and ROC value (Receiver Operating Characteristic area) for each subset of features selected by the feature selection methods.

This investigation focuses mainly on four attribute selection algorithms; CfsSubsetEval (CFS), Pearson's Linear Correlation (CR), Gain-ratio attribute evaluation (GR) and Information-gain attribute ranking (IG). A brief description of each of above attributes selection algorithms is given below.

I. Pearson's Linear Correlation (CR)

In general, a good feature is one that is highly relevant to the class (prediction) and less redundant to other features. Features that are strongly correlated with another set of attributes are known as redundant features. A feature is redundant when it can be derived by another attribute or set of the attributes. The redundancy of features can be detected using a correlation analysis. One of the most well-known measures for correlation detection is the linear correlation coefficient. For two features X and Y the linear correlation coefficient is given by the following formula:

$$Correlation(r) = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

Where n is the number of observations.

II. CfsSubsetEval (CFS)

The Pearson's linear correlation select the best features that are highly correlated with the class variable. However, it does not consider the inter correlation among the features. CR algorithm developed by Hall (2003) is based on a hypothesis that "A good feature subset is one that contains features highly correlated to the class, yet uncorrelated to each other". The objective of CFS is to select subsets that are correlated to the class rather than the individuals. With this objective, CFS evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. In other words, subsets of features that are highly correlated to the class while having low inter-correlations are selected. The equation for CFS is given by:

$$r_{zc} = \frac{K\overline{r_{cf}}}{\sqrt{K + K(K-1)\overline{r_{ff}}}}$$

Where r_{zc} is the correlation between the summed feature subsets z and the class variable c, k is the number of subset features, r_{cf} is the average of the correlations between the subset features and the class variable, and r_{ff} is the average inter-correlation between subset features (Hall, 1999).

III. Information-gain attribute ranking (IG)

Information gain measures the information obtained for category prediction by knowing the presence or absence of a feature. For a feature f and prediction class c, the information gain IG is defined as:

$$IG(f) = -\sum_{i}^{m} P(c_i) log P(c_i) + P(f) \sum_{i}^{m} P(c_i|f) \log(c_i|f)$$

For each feature, information gain is measured and removed from the list if the value is below a predetermined threshold.

IV. Gain Ratio (GR)

IG prefers to select features that have a higher number of values. GR reduces the bias and is the modified version of IG. Using GR is a way of applying normalization to IG

by taking the intrinsic information of a split into account. Intrinsic information is, how much information need to decide which branch an instance belongs to. It reduces bias of multi-valued attributes (Priyadarsini, Valarmathi, & Sivakumari, 2011, Han, Kamber, & Pei, 2012). The gain ratio for the attribute f is given by;

$$Gain\_Ratio(f) = \frac{Gain(f)}{intrinsic\_info(f)}$$

$Gain(f)$ is defined as;

$$Gain(f) = -\sum_{i}^{m} p_i \log_2 p_i + \sum_{1}^{m}\sum_{i}^{m} p_i \log_2 p_i \left(\frac{C_{1i} + C_{1i} + C_{1i} + \cdots + C_{mi}}{C}\right)$$

where, p_i is the probability that an arbitrary sample belongs to class C_i and m is number of classes.

The performance evaluation of these four feature selection techniques was carried out iteratively on subsets of features starting from a cardinality of two. Initially, the techniques were run for all features, and a ranked feature list was obtained. Features were ranked using the information measures and correlations respectively for each attribute selection method.

The feature selection method ran for several iterations henceforth known as a pass. Each pass returned the optimum number of features and then iteratively searched further for the best feature list by incrementing one feature at a time. In each iteration, the F-measure and ROC value were obtained. The ROC curve is a graphical representation of sensitivity (true positive) versus specificity (false positive rate), and it depicts the relative trade-off between sensitivity and specificity. The ROC curve can be employed as a technique for selecting a classifier or as a diagnostic test.

6.3.4 Classification Techniques

As previously mentioned, throughout the chapter the experiments were carried out using Naïve Bayes and SVM classification algorithms. A detailed explanation of these algorithms was presented in Chapter 4, Section 4.6.1.

6.4 Results

The following sections presented the results of the experiments and provided an explanation of the results.

6.4.1 Document Level Sentiment classification for Sinhala opinions

It is important to investigate sentiment analysis for Sinhala at document level to analyse the combined effect of sentences in an opinion. In this section, a complete opinion is treated as a document without splitting it into sentences. In the initial analysis, it was noted that a document contains an average of two sentences a minimum of one sentence and a maximum of five sentences.

In the initial study, all words comprising an opinion were extracted after removing the stop words. These extracted words are known as keywords that are non-grammatical terms in a written document that explain the main concept of the text. To understand the set of documents or comments, it is essential to select the keywords. The frequencies of all the words in comments after removing the stop words were calculated to generate a list of keywords. The minimum frequency of 1 indicates that the word is less important or has little relationship to the concept of the comment being talked about. On the other hand, a word with maximum frequency is highly correlated with the concept. It is also noted that it approximates Zipf's word frequency law that there are more words with less frequency and that only a few words have the highest frequency count (Piantadosi, 2014). Piantadosi (2014) has empirically proved that word distribution of a given language is near-Zipfian. Figure 6.2 shows that the keyword distribution of the Sinhala comments has a trend which is consistent with to Zipf's rule.

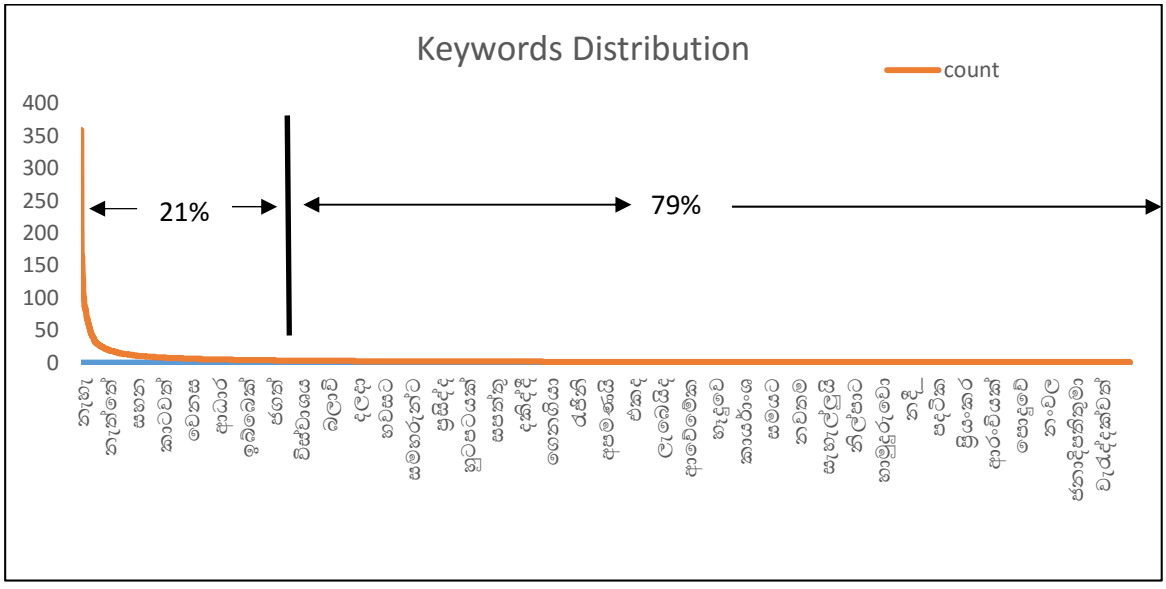


Figure 6.2: Keywords distribution of 2083 Sinhala opinions

In this sample, the word "නැහැ" (no), which in general indicates negation of concepts, has the highest frequency. Moreover, more than 65% of the keywords appear with single frequency (1)

in the sample revealing that there are more unimportant words than concept bearing words. Out of a total of 9,575 keywords, counts with different levels of frequencies were calculated in order to investigate different keyword densities. The distribution presented in Table 6.1.

Table 6.1: Keyword densities

Keyword Frequency	Count	%
Greater than 3	2581	26.95
Greater than 5	1429	14.92
Greater than 10	611	6.37
Greater than 15	363	3.78
Greater than 20	260	2.72
Greater than 50	85	0.89

According to Table 6.1, a significant difference in keyword densities of the size of three and five was observed. The process of determining the feature size as well as the best feature set is known as feature selection. Keyword frequencies provide information which can be used for feature selection in supervised sentiment classification. The number of occurrences (frequency) is the simplest method for feature selection.

A number or experiments were conducted:

a. Without feature selection, and using all 9575 features as the feature set
b. With binary weighting and term frequencies feature selection
c. Using term frequency-inverse document frequency (tf-idf) feature selection method

A comprehensive explanation of feature weighting is presented in chapter 4. Finally, a Naïve Bayes classifier and a SVM classification was undertaken to evaluate the performance of the feature selection approaches.

6.4.2 Classification Accuracies

(a) Experiment with unigram features

When searching for occurrences of the 9,575 keywords in each opinion, it was observed that all opinions contained at least one of the keywords in the set of 2,083 opinions. In the supervised classification, opinions that consisted of at least one keyword were considered. The

classification accuracies with evaluation measures using the binary weights are presented in Table 6.2.

Table 6.2: Classification Accuracies with all features

	Classification Method	
	Naïve Bayes	SVM
Accuracy (%)	46.952	44.215
Precision	0.477	0.440
Recall	0.470	0.442
F Value	0.468	0.441

Classification using Naïve Bayes performs better than SVM, but for both methods (without feature selection) the accuracies are extremely poor. A possible reason for the poor performance could be attributed to the sparseness of the data as the document vector represents a high number of dimensions. With the aim of reducing the dimensionality, the experiment was repeated varying the number of features to examine the effective of feature size. The accuracies of each test case where the features were weighted by one when the keyword was present in the opinion and zero otherwise, is given in Table 6.3.

Table 6.3: Classification performances for different feature sets

		NB				SVM			
# Features	# Opinions	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
2581	2083	0.481	0.477	0.475	**47.671**	0.445	0.446	0.446	44.647
1429	2082	0.479	0.473	0.472	47.336	0.435	0.436	0.436	43.639
611	2071	0.458	0.454	0.454	45.410	0.426	0.429	0.427	42.899
363	2063	0.452	0.447	0.446	44.665	0.416	0.422	0.419	42.241
260	2050	0.454	0.449	0.449	44.949	0.421	0.431	0.425	43.094
85	1951	0.443	0.437	0.439	43.721	0.398	0.441	0.389	44.131

The classification accuracy, of both Naïve Bayes and SVM, was found to decrease as the number of features decreased (Tables 6.2 and 6.3). Even though no significant difference exists between Naïve Bayes and SVM accuracies, the performance of Naïve Bayes was slightly better than that of the SVM. The precision is better than recall for Naïve Bayes but in SVM recall is higher than its precision. It is also noted that performances are comparatively low when compared with what has been reported for other languages (Deng, Luo, & Yu, 2014). However, the benchmark accuracies for other languages are for classification with only two classes: positive and negative whereas in this experiment a neutral class is included. Most sentiment classification experiments are conducted using only positive and negative classes. In order to examine the difference between two-class and three-class classifications the experiments were repeated using two classes of opinion (positive and negative) and the performance is presented in Table 6.4. As in the previous experiment, the features were weighted by 1 and 0 (binary).

Table 6.4: Classification performances by positive and negative classes

		NB				SVM			
# Features	# Opinions	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
9575	1579	0.612	0.612	0.612	61.217	0.594	0.593	0.593	59.252
2581	1578	0.608	0.608	0.608	60.773	0.590	0.589	0.590	58.935
1430	1578	0.603	0.603	0.603	60.329	0.585	0.584	0.585	58.428
611	1567	0.596	0.594	0.594	59.413	0.715	0.586	0.491	58.647
363	1561	0.604	0.602	0.603	60.218	0.641	0.630	0.612	**62.972**
260	1550	0.609	0.606	0.607	60.645	0.599	0.599	0.599	59.935
85	1484	0.569	0.565	0.566	56.536	0.626	0.597	0.545	59.704

The results of these experiments given in Table 6.4 show that two class classification provides significantly improved sentiment classification for Sinhala opinions. The precision and recall values are almost the same in all test cases as feature size increases. In most of the cases, Naive Bayes achieved higher accuracies than the SVM. However, when 363 features (keywords) with a frequency greater than 15 were extracted the highest classification accuracy in SVM learning was obtained giving an indication of the optimum feature size. These tests show promising results for two class classification using positive and negative codes alone. Hence, it was decided to continue the rest of the research using these two categories only omitting neutral opinions/ reviews.

With the assumption that better accuracies can be obtained by weighting features by their relative frequencies, the experiment was repeated with the same set of keywords but assigning a frequency to each keyword as a factor of its weight. The classification accuracies are presented in Table 6.5.

Table 6.5 reveals that the highest classification performance is achieved with the 611 keywords set by the Naïve Bayes algorithm. Therefore, at this stage, the conclusion can be drawn that the initial feature set for the frequent list approach should be based on 611 features. That is keywords whose frequencies are greater than 10 are the set of features for optimum classification performance. To validate the features selected by the binary and relative

frequency weighting, the experiments were repeated this time using tf-idf weights for the two class problem.

Table 6.5: Classification performances by relative frequencies

# Features	#Opinions	NB				SVM			
		Precision	Recall	F	Accuracy	Precision	Recall	F	Accuracy
9575	1575	0.607	0.597	0.595	59.682	0.690	0.542	0.399	54.222
2581	1563	0.598	0.589	0.587	58.861	0.682	0.541	0.396	54.063
1430	1551	0.596	0.588	0.587	58.842	0.682	0.540	0.395	54.019
611	1529	0.600	0.600	0.600	**60.039**	0.701	0.546	0.402	54.611
363	1503	0.578	0.580	0.576	58.017	0.595	0.597	0.592	**59.680**
260	1486	0.586	0.588	0.581	58.816	0.696	0.545	0.400	54.508
85	1503	0.578	0.580	0.576	58.017	0.697	0.546	0.402	54.624

The performance measures with accuracies achieved are presented in Table 6.6. The result further revealed that the highest accuracy by 611 keywords set using Naïve Bayes irrespective to the weighting method. However, SVM gave the best accuracy for 1430 keywords.

Table 6.6: Classification performances by tfidf weights

# Features	#Opinions	NB				SVM			
		Precision	Recall	F	Accuracy	Precision	Recall	F	Accuracy
9575	1578	0.608	0.603	0.603	60.329	0.634	0.627	0.614	62.738
2581	1563	0.582	0.582	0.566	58.157	0.603	0.604	0.603	60.397
1430	1555	0.596	0.588	0.587	58.842	0.635	0.619	0.595	**61.865**
611	1578	0.613	0.605	0.603	**60.456**	0.638	0.613	0.579	61.280
363	1503	0.576	0.578	0.574	57.818	0.644	0.612	0.570	61.211
260	1486	0.588	0.590	0.582	59.950	0.636	0.601	0.551	60.094
85	1362	0.582	0.583	0.560	58.297	0.652	0.593	0.517	59.325

6.4.3 Applying Feature Selection Methods for Classification

The previous experiments use feature extraction approaches that were developed specifically for text mining tasks. In the following section, the thesis tested the feature selection methods explained in 6.3.3 for document level sentiment classification. In the first trial with 9,575 features, 70 features were extracted through the CFS method. The performance for both, Naïve Bayes and SVM improved using the features selected by CFS (Tables 6.7, 6.5 and 6.6).

Table 6.7: Performances by CfsSubsetEval feature selection

	Accuracy	F-Measure
Naïve Bayes	63.117	0.574
SVM	63.308	0.574

Also, it can be noted that this feature selection method results in an increase in accuracy in all most all cases (Table 6.7). With the aim of further reducing the dimensions and increasing the classification accuracy, a test was carried out with the same algorithm but increasing the number of features one at a time until 70 features were reached. It was noted that no significant improvement in performance was observed and the highest accuracy was attained when using all 70 features.

The other feature selection techniques were tested in a similar manner and the F-measure and ROC value were computed in addition to classification accuracy. In these experiments, several passes for each feature selection method were tested. In the first stage, the features of each feature selection method were ranked by applying Naïve Bayes and SVM for all 9,575 attributes. Then the classification performances were observed by incrementing the feature sizes by 1,000 extracted from the ranked list. In the second pass, iteration was performed by increasing 100 features for a subset of the list of highest performance feature list decided in the first pass.

The best number of features that gives the highest classification accuracies in the first and second passes of the experiment using Naïve Bayes and SVM is given in Table 6.8 and 6.9.

Table 6.8: Best number of features for Naïve Bayes

Feature Selection Method	First Pass		Second Pass	
	Number of features	Accuracy	Number of features	Accuracy
CR	2000	63.945	200	68.568
IG	2000	60.963	200	60.837
GR	2000	63.945	200	60.836

Firstly, 2,000 features ranked by the CR gives the highest accuracy of 63.945%. Similarly, 2,000 features were selected by IG selection method, but the accuracy was lower than that achieved by the CR and, GR methods. The F-measure for each iteration plotted against the number of features is shown in figure 6.3.

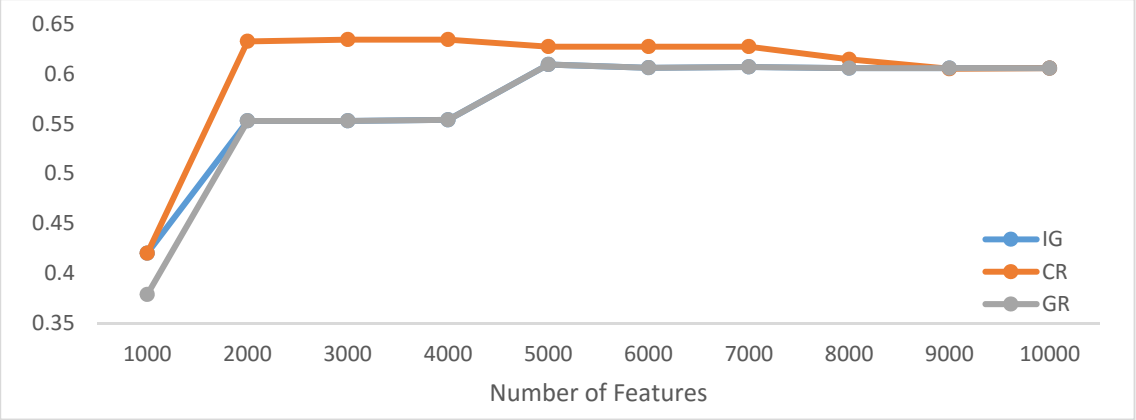


Figure 6.3: F-measures for feature selection first pass – Naïve Bayes

Figure 6.3 shows the change in F-measures with respect to feature selection methods and number of features. The CR method gave the best F-measure of 0.634 with 2,000 features and the F-measure plateaus at 2,000 and remains relatively stable as the number of features increases to 7,000 features. IG and GR also have an F-measure which plateaus at 2,000 features but then increases again at 5,000 features before stabilising. The best performance for IG and GR was observed with 5,000 features which have an F-value of 0.609. The plot of ROC values for the feature selection methods also indicates that the best feature selection algorithm of the three is the CR method.

The F-measure distribution for the second pass stage was plotted against the number of features for each of the iterations and gives a good indication of the best feature selection (F = 0.657) method (CR) and an optimum number of features (200). Using CR the classification accuracy was improved by 7.22% through the reduction of dimensions in the feature vector.

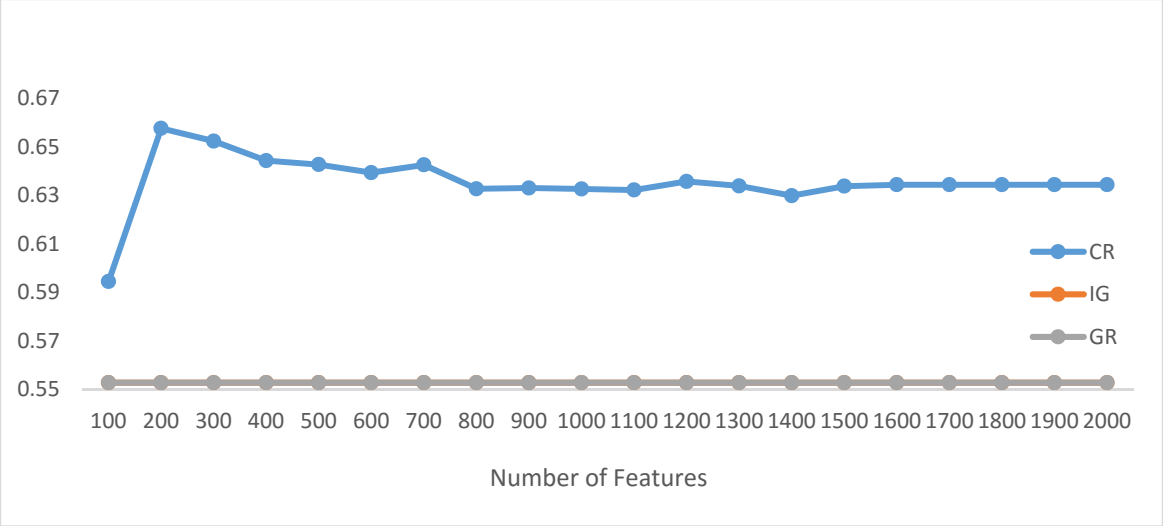


Figure 6.4: F-measures for feature selection second pass – Naïve Bayes

The IG and GR methods showed no change in the classification accuracy in the experimental range of 100 to 2,000 range of features.

The similar experiments undertaken with the Naïve Bayes classifier were repeated for SVM in order to examine the effect of the feature selection method on classification performance. The evaluation measures are presented in Table 6.9.

Table 6.9: Best number of features by SVM

Feature Selection Method	First Pass		Second Pass	
	Number of features	Accuracy	Number of features	Accuracy
CR	2000	81.559	1100	81.876
IG	6000	64.132	600	64.068
GR	3000	62.927	500	64.195

This experiment using SVM gave better performances than for Naive Bayes but again, features selected by the CR method gave the best classification accuracy. The highest accuracy achieved was recorded in this experiment with 2000 features using SVM classification algorithm. Remarkably, the highest accuracy achieved with information gain (64.132%) was achieved using 6000 features – two-thirds of the total number of features. The result confirms that a second pass is required in order to reduce the number of dimensions further. The GR) approach also requires a higher number of dimensions than the CR method. The F-measures of the experiment for all three attribute selection methods are provided in Figure 6.5.

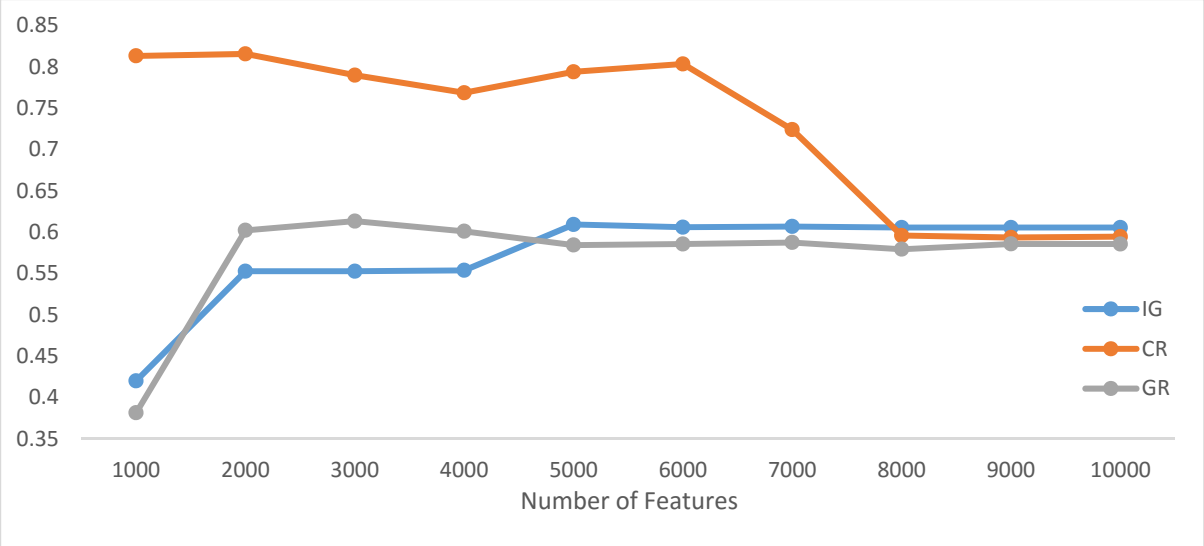


Figure 6.5: F-measures for feature selection first pass - SVM

According to the graph (Figure 6.5), the CR approach is a good feature selection method as it gives the highest F-measure even outperforming CFS. Unlike Naïve Bayes, for SVM the IG and GR F-measure distributions differ from each other.

The second pass of the experiment shows a clear improvement in performance caused by a reduction the number of features. However, the second pass reduction using the CR approach is not significant as it gives an accuracy of 81.876 for 1,100 features, for which the reduction is only 30% in terms of features used. In further observation of accuracies, 300 features gave the next best performance with 75.79% accuracy. In the first pass, the IG and GR selection methods show good classification accuracies, however, the reduction is not significant when compared to CR selection. On the other hand, for the other two methods, the second pass reduced the number of features by a considerable amount with both methods achieving a similar level of classification accuracy. However, when compared to the accuracies achieved with the correlation-based reduction, it is not significant as correlation-based gives 300 best features at 75.79% level of accuracy. The best number of features for IG and the GR was achieved at 600 and 500 features respectively.

Figure 6.6 shows the comparison of F-measures and values for the three attribute selections methods for SVM classification in the second pass.

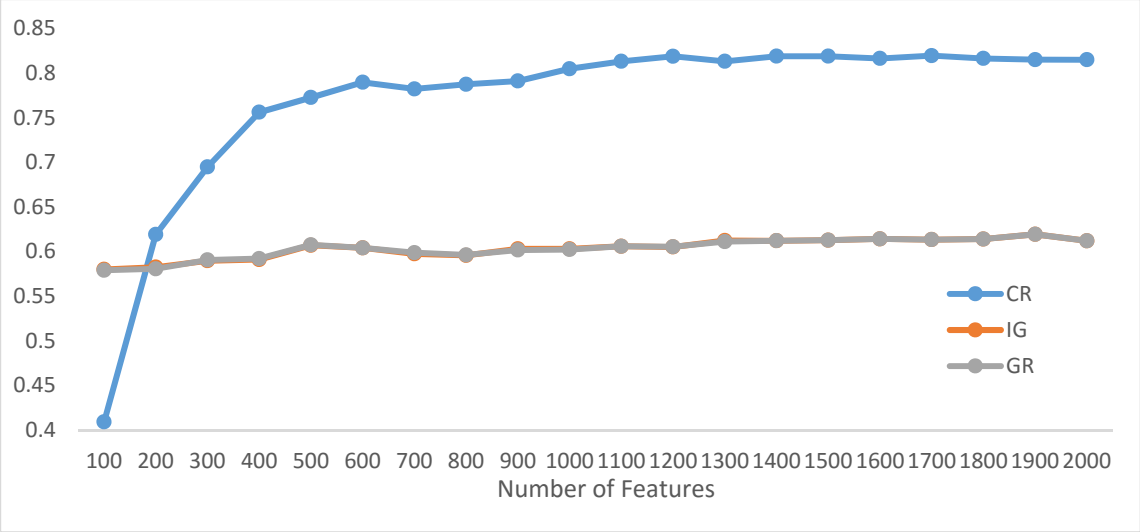


Figure 6.6: F-measures for feature selection second pass - SVM

Figure 6.6 shows the significance of the CR method for SVM classification. IG and GR show almost same performances based on F-measures. The graph also indicates that performances with more than 300 features selected by the CR approach give better classification accuracies.

In conclusion, the best feature selection algorithm for the frequentist based approach for Sinhala comments is the CR method. Even though Naïve Bayes classification achieved lower performance values with all features, the feature selection by CR method showed better results for both Naïve Bayes and SVM. In the final investigation, 1,100 features gave the best accuracy of 81.88%. However, the dimensions of the feature set is comparatively high. Nevertheless, as alternative methods, the results of these experiments suggest starting with 300 features which give 70.79% accuracy and then improving the classification by turning the parameters of the classifier.

(b) Experiments with n-gram features

So far in this study of sentiment analysis for the Sinhala language only a single word feature (unigram) has been considering in the classification experiments. Figure 6.7 presents the bigram distribution for the sample opinions.

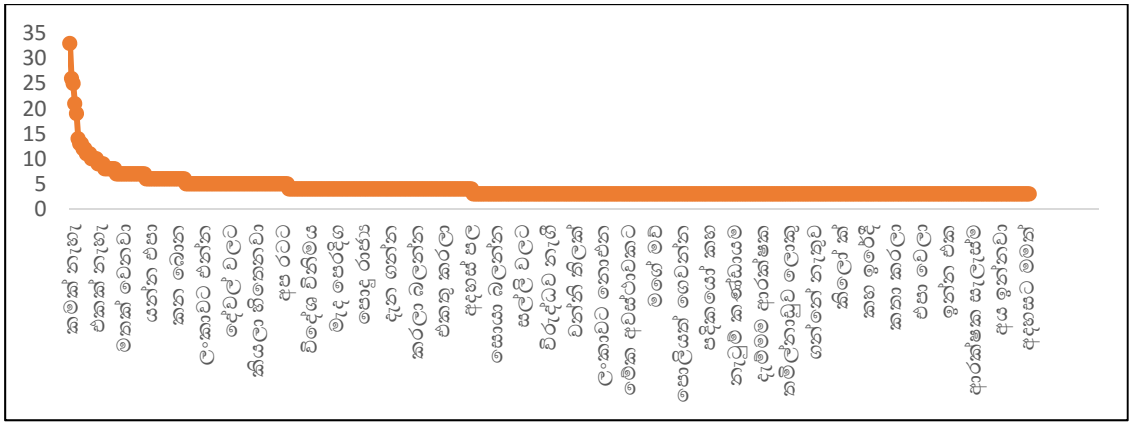


Figure 6.7: Bigram Distribution

More than 92% of the bigrams were observed to be in single counts, and there are 33,069 bigrams formed by the keywords of the sample data. To minimize data sparsity, bigrams whose count were more than three were extracted for further classification in this research. Consequently, 678 bigrams were selected for the classification vector. It is also noted that only 44% of the opinions contained at least one bigram feature within the sample data. This is a disadvantage of using bigrams – the sample representation of the bigrams is considerably lower than for unigrams when comparing the occurrences of each in the sample. Using this bigram set classification was again carried out using Naïve Bayes and SVM by changing the feature weighting as tested in the unigram analysis. The results are given in Table 6.10.

Table 6.10: Baseline accuracies by bigram features

Features	NB				SVM			
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
binary	0.596	0.596	0.596	59.611	0.587	0.591	0.586	59.071
Relative freq.	0.588	0.573	0.502	57.343	0.579	0.583	0.579	58.315
tfidf	0.576	0.581	0.574	58.099	0.614	0.616	0.604	61.555

According to Table 6.10, the highest baseline performance is achieved by experiment setup of tfidf weighting classified by the SVM algorithm. Noticeably binary weighting gives equal performance accuracies for both algorithms, and it is the highest among all features for Naïve Bayes classifications. Also, it was noted that in all cases, negative opinion classification shows better performances in terms of F-measure.

Since this is the first attempt of sentiment analysis for Sinhala, it is interesting to observe how the higher order n-grams classify the opinions into the positive and negative classes. Out of 39,110 trigrams, only 82 (21%) trigrams had a frequency greater than 3. By searching these 82 trirams in the sample of 2,083 opinions, only 96 contained at least one trigram. Not unexpectedly, there is a significantly smaller representation of tri-grams in the sample than bigrams and unigrams. However, the baseline performance by the trigrams significantly boosts the accuracies when compared with that of bigrams and unigrams. The classification accuracies are presented in Table 6.11.

Table 6.11: Baseline accuracies by trigram features

	Accuracy	F-Measure
Naïve Bayes	66.25	0.665
SVM	77.4	0.772

The results in Table 6.11 further prove that the SVM classification algorithm is the most suitable for sentiment classification for Sinhala even with trigram features.

Next, the research focused on looking at the effect of feature reduction for bigrams. It is essential to test for the improvement in the above classification accuracies when applying feature selection techniques as conducted for unigram features. Hence, the feature selection methods applied for the unigram analysis were also tested in a similar experiment setup for bigrams.

Since binary weighting is the best feature weighting for Naïve Bayes when employing bigram features, this set of experiments adheres to the same weighting method for feature selection testing. According to Table 6.10, the binary weighting scheme gives better accuracies for Naïve Bayes classification. Therefore, in the study of feature selection, binary weightings are initially iterated by the number of features, 50 at each iteration for a total of 678 features. Then for the second pass, the increment is performed by increasing with single feature until the optimum number of features is obtained by the first stage. On the other hand, better performance is shown by the SVM when using tfidf weighting as per Table 6.10. For the attribute selection by SVM classification, the tfidf feature weights were applied in a similar iteration mechanism as explained above for the Naïve Bayes. Table 6.13 gives the classification accuracies for both passes.

Table 6.12: Bigram feature selection by Naïve Bayes

Feature Selection Method	First Pass		Second Pass	
	Number of features	Accuracy	Number of features	Accuracy
CFS	56	68.250	56	68.250
CR	150	61.988	134	61.987
IG	100	61.231	70	61.231
GR	100	61.231	70	61.231

Clearly, CFS is the best feature selection method for Naïve Bayes when using binary features (see Table 6.12). The accuracy and number of features are optimal, and CFS outperformed other selection methods. On the other hand, no significant improvements in classification were noted using CR, IG or GR feature selection methods. However, the number of features selected by IG and GR is less than that of the CR method.

Table 6.13: Bigram feature selection by SVM

Feature Selection Method	First Parse		Second Parse	
	Number of features	Accuracy	Number of features	Accuracy
CFS	12	56.911	12	56.911
CR	300	73.974	180	69.762
IG	100	67.494	90	67.819
GR	100	67.279	70	67.927

As Table 6.13 shows, the CR feature selection methods work well in finding the best bigram feature set using SVM classification. Out of 678 attributes, using the first 300 ranked by correlation gives a high accuracy in the first stage and the best accuracy for all classifications. In the second parse, the CR feature selection also gave the best performance when compared with the other feature selection methods.

6.4.4 Conclusions on document level sentiment classification

The following conclusion was drawn based on the sentiment classification experiments conducted so far;

a. Two-class classification provided better results than three class classification. In other words, sentiment classification based on the Positive and Negative categories experimentally proved superior to the analysis with Positive, Negative, and Neutral
b. For unigram classification, the tfidf feature weighting method resulted in better classification for SVM but not for Naïve Bayes.
c. For Naïve Byes, of the weighting methods evaluated binary feature weighting gave the best classification results.
d. The best features were selected by CR for SVM classification. For Naïve Bayes classification, the GR feature selection method gave good classification accuracies.
e. The tfidf weighting approach was proved to be the best feature weighting for bi-gram features and SVM classification. Binary weighting is best for Naïve Bayes classification. However, the performances do not significantly differ.
f. It was demonstrated that better classification accuracies could be obtained by using higher order n-grams such as trigrams
g. CR feature selection was confirmed as the best feature selection method for SVM classification when using bigram features as well as unigram features.

6.4.5 Domain-Specific Sentiment classification of Sinhala opinions

Results of the previous experiments indicate that a more detailed investigation is required into sentiment classification for Sinhala opinions. One of the approaches worth investigating is a finer-grained sentence level analysis to see if it results in improved classification accuracies. Some researchers argue that there is no fundamental difference between a document and sentence-level sentiment classification (Medhat, Hassan, & Korashy, 2014). Despite this viewpoint, it is still worth investigating to see if this is true for morphologically rich languages.

One of the challenges of sentence-level sentiment classification is making annotated corpora for the experiment. A document (a set of opinions) must be further divided into sentences, and each sentence has to be coded for making such a training data set. Another challenge of sentence-level analysis is sparsity in data as some sentences contain fewer words than documents. In the following section, sentiment classification investigated at the sentence level.

The same experimental setup as was used for document-level analysis was adopted to allow comparison of the results. The experiments were limited to binary polarity classification (positive and negative) based on previous experiments 6.3.2 (Table 6.3) which showed that ternary (positive, negative and neutral) classification resulted in poor performance. In this experiment, opinions were limited to a single domain. They were extracted from comments on political articles. Political comments formed the second highest reader response category for the news articles in the sample.

The experiment began by dividing original 608 political opinions into sentences resulting in 944 sentences. These 944 political related opinions were annotated as either positive, negative or neutral by the same annotators as discussed in chapter 4 section 4.3.2. A quantitative overview of the opinions analyzed is given in figure 6.8.

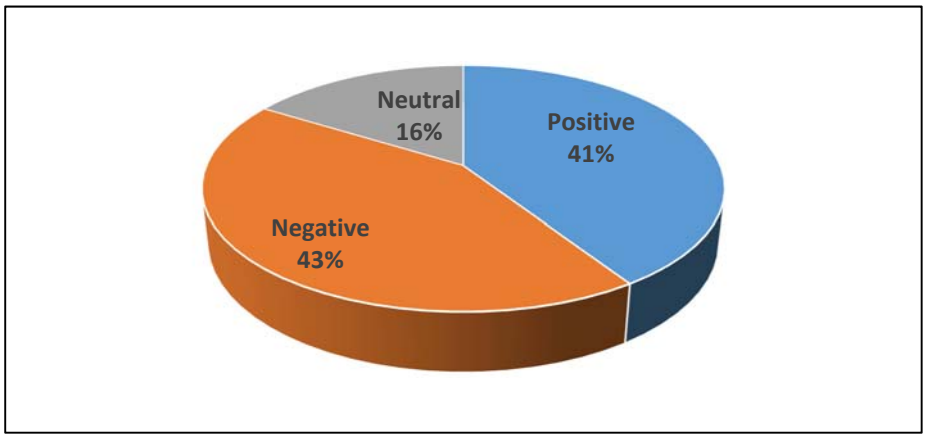


Figure 6.8: Politically related opinions Distribution

The sample selected for the analysis is almost balanced on positive and negative labels as shown in figure 6.8. Therefore, further analysis of this sample was carried out only for these labels to remove the class imbalance problem caused by the small number of neutral opinions. The decision for considering only positive and negative opinions is also supported by the performance achieved by the previous sections and by the literature.

With the aim of understanding the opinions for political news articles, the research investigates keyword distribution after removing the stop words. The distribution is given in figure 6.9. In this domain, 5,061 keywords were extracted among them 32% have a frequency greater than 1. The keyword "නැහැ" (not) is the most frequent keyword as is the case for all opinions regardless of the comment's domain (figure 6.2 section 6.3.1)

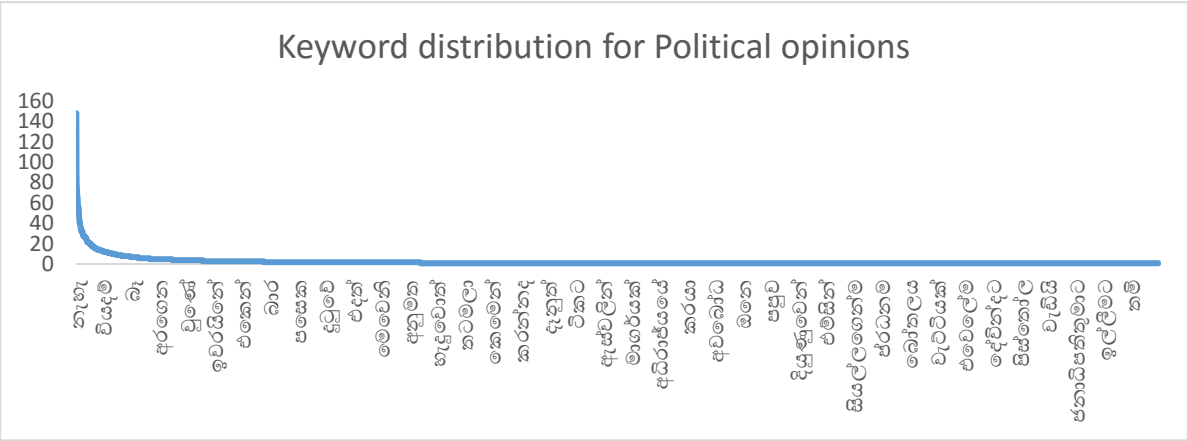


Figure 6.9: Keyword distribution for "Politics" related opinions

6.4.6 Unigram and Bi-gram Analysis

The primary objective of this experiment is to test the effectiveness of keywords in a particular domain in sentiment classification. The set of political opinions were classified using Naïve Bayes and SVM with tfidf feature weighting and binary feature weighting methods as explained in detail the unigram and bigram analysis in section 6.3.3. The experiments were conducted with bigrams and unigrams (keywords) with a frequency greater (877 features) than two to reduce the dimensions – this was an arbitrary choice. The results are given in Table 6.14. No significant improvement was observed in the performance of Naïve Bayes or SVM classification for any of the feature types using domain dependent sentiment classification for Sinhala opinions. However, feature weighting using tfidf proved to be the best weighting method this is in line with earlier results in this research which found that tfidf weighting was the best feature weighting method for Sinhala sentiment classification independent of the domain.

Table 6.14: Domain-dependent base-line accuracies by unigram features

Feature	Weighting	Naïve Bayes		SVM	
		Accuracy	F- Measure	Accuracy	F- Measure
Unigram	binary	60.886	0.609	60.633	0.606
Unigram	tfidf	60.483	0.602	63.659	0.636
Bigram	binary	56.401	0.563	59.516	0.591
Bigram	tfidf	58.478	0.582	58.824	0.588

6.4.7 Feature Selection for domain specific analysis

In this experiment, an attempt was made to mine the best feature set out of the 877 unigrams features. The Table 6.15 shows the best classification performances with the optimum feature size for each case using Naïve Bayes classification.

Table 6.15: Domain dependent unigram feature selection using Naïve Bayes

Feature Selection Method	Number of features	Accuracy
CFS	76	69.113
CR	178	67.975
IG	75	67.594
GR	96	70.253

The optimum number of features and the relevant accuracies obtained using the feature selection method are shown in Table 6.15. An approximately similar accuracy is achieved with 76 features selected using the CFS algorithm, but CFS is better because it resulted in lower feature dimensions - a 20% reduction in the number of features selected by GR.

Table 6.16: Domain dependent unigram feature selection using SVM

Feature Selection Method	Number of features	Accuracy
CFS	30	60.864
CR	339	82.973
IG	115	67.090
GR	725	62.135

As shown in Table 6.16, in the experiment of using SVM the best accuracy was obtained when features were selected with CR which results in a relatively high number of features when compared with CFS and IG selection methods.

For further examination with bigrams for domain dependent sample, the feature selection process test was carried out in a similar manner for both algorithms Naïve Bayes and SVM. The results are presented in Tables 6.17 and 6.18.

Table 6.17: Domain dependent bigram feature selection using Naïve Bayes

Feature Selection Method	Number of features	Accuracy
CFS	20	63.668
CR	62	67.820
IG	22	64.359
GR	23	64.359

CR bigram feature selection was proven to be the most effective approach for domain dependent sentiment classification but the number of features selected was much higher than for the other methods evaluated. Using CFS, IG and GR resulted in poorer classification but reduced the feature space approximately by two-thirds. When compared with the experiments using CR selection 96 unigram features (Table 6.15) gave better classification accuracy than 62 bigram features.

Table 6.18: Domain dependent bigram feature selection using SVM

Feature Selection Method	Number of features	Accuracy
CFS	20	64.706
CR	59	70.242
IG	91	58.823
GR	91	58.823

Table 6.18 further proved that the correlation based (CR) feature selection is good for the SVM based classification using bigrams. The number of features has been reduced by 72% compare to unigram consideration with SVM. While the classification accuracy is dropped by 15%.

6.5 Discussion

The above analysis has demonstrated the feasibility of frequentist-based sentiment classification. The performances of the classification improved from 44% to 83% by applying different experimental strategies. In the following section, the possible causes of poor accuracies are discussed with a view of improving the performance. The discussion is mainly based on the classification accuracies and confusion matrices generated by the classification at each stage. Both classification algorithms, Naïve Bayes and SVM, are taken into account in the following discussion.

This thesis first explained (section 6.3.2) the problem of three class classification over two class classification. In the initial work, it was found that the performance of positive, negative and neutral categories was below average, and performance was increased when only positive and negative labels were used. The classification by binary feature weighting using Naïve Bayes with 2,581 features gave the highest accuracy (47.671) for three-class analysis. The highest individual F-measure was achieved in the negative class (0.541) and the lowest score (0.364) was observed in the neutral class. The confusion matrix indicated that 40% of the neutral class opinions were classified as negative. In the application of SVM for the same feature set with binary weighting, it was revealed that the lowest F-measure was again achieved by the neutral class and majority of the neutral opinions (39%) were classified as negative – this is the same trend as observed when using the Naïve Bayes classifier.

Next, the effectiveness of the feature weighting on classification to positive and negative classes were analysed in this research. The Naïve Bayes gave 60.039 accuracy with 611 unigram features weighted by relative frequencies. Examining the confusion matrix, it was observed that 44% of the positive opinions were incorrectly classified as negative giving the lowest F-measure for the positives. Classification using SVM gave similar results, 45% of the positive opinions were classified as negative, and 56% of positive opinions were identified as negative when using the tfidf weights resulting in 63% accuracy. The Naïve Bayes exhibited the opposite trend, 46% of negative opinions were classified as positive opinions. Classification accuracy of SVM using tfidf was better than that of the Naïve Bayes.

As shown in Table 6.8, the best 200 unigrams were selected by the CR feature selection when using the Naïve Bayes classifier. In this classification, 12% of the negative opinions were classified as positive, and 53% of positives were classified as negative.

The first ten words that were selected by the CR method were then examined. It was noted that in these first ten words, the negation word "නැහැ" ("not") was at the top of the list. Secondly, the adjectives "හරිම" (very) correlated to the labelling significantly. In the Sinhala language context, the adjective හරිම (very) was found to be followed frequently by a positive or negative word. For example; හරිම ලස්සනයි (very beautiful), හරිම රළුයි (very rough) the word is considered to be functioning as an intensifier. Among these words some greeting words were in the top of the list leading to a conclusion that these words are highly dominant in these kinds of reviews (about news articles) and such opinions should be removed prior to classification.

Three hundred unigrams were selected by CR technique and SVM classification with accuracy 75.79% (Figure 6.6). In this selection, the confusion matrix revealed only 21% of positive opinions were misclassified as negative. This is a significant improvement in terms of classification accuracy when compared with that of Naïve Bayes. However, the number of unigram features selected increased to 300. In comparison, 200 unigrams features gave the optimal results when using Naïve Bayes. When comparing the selected feature sets, the 300 SVM unigrams consisted of 21% more negative words and 20% more positive words than the optimal feature set for Naïve Bayes.

In the investigation of bigram features, the misclassification of the negatives (37%) was less than that of the positives (45%) for Naïve Bayes classification where the highest accuracy was achieved using binary weighting. Twenty-three percent of negatives were detected as positives, and 56% of positive opinions were wrongly classed as negative opinions. The misclassification of positive opinions is higher using bigram features than using unigram features.

An error analysis of bigram feature selection was also carried out using a similar approach, and it was noted that selection of the Naïve Bayes optimal feature set using CFS reduced the misclassifications of negative opinions more than for unigram features. On the other hand, positive misclassification increased to 69%. Even though SVM gave the highest classification accuracy using CR feature selection, the number of positive misclassifications was higher. Among these top bigrams, the feature selection algorithms selected both positive and negative bigrams. The negative bigrams such as "ගන්න බැරි" (could not), "දන්නේ නැහැ" (do not know), "වෙන්න එපා" (do not be) and "ගන්නේ නැහැ" (do not take) were found in the best negative features. Positive bigrams including greeting words were observed. It was also noted that the number of positive bigrams was less than negative bigrams. These postive bigrams are not direct positives as the polarity of the bigram depends on the context of the sentence.

In the investigation of domain dependent analysis, the misclassification percentages reduced in both positive and negative opinions significantly when compared to the domain independent classification. The misclassification for negatives opinions decreased to 12% and 25% of positive opinions incorrectly classified.

6.6 Chapter Summary

The primary aim of the research presented in this chapter was to investigate the adaptability of the contemporary text mining approaches in sentiment classification for a morphologically rich language. The study was carried out using Sinhala language opinions. The discussion presents

sentiment analysis purely based on statistical approaches where Sinhala opinions were classified using statistical feature weighting approaches, such as relative frequency, ifidf weights. An investigation of the effect of the classical text mining techniques and lexical features, such as unigram and bigrams was presented and the effect of statistical features, the influence of feature weighting, and feature selection techniques. The results showed that SVM classification is more suited for Sinhala opinion classification than Naïve Bayes. However, the performance was found to dependent on the features and the weighting scheme used. Higher performance was obtained using tfid with SVM, and in the case of Naïve Bayes, it was noted that binary weighting gave the most promising results. Significant improvement was shown when higher order features/n-grams, were used.

In the next chapter, the research presented focuses on the analysis of linguistic features in the Sinhala language and how exploiting these can influence the success of sentiment classification.

Chapter 7: Linguistic features in Sinhala for sentiment classification

7.1 Introduction

Chapter 6 gave details of an investigation into the application of classical text mining techniques in sentiment classification for the Sinhala language. All experiments in chapter 6 were carried out using approaches based on purely statistical measures independent of any language-specific knowledge or rules. Statistical measures may fail to capture linguistic features explicitly. This research is a pioneering attempt to undertake sentiment classification in the Sinhala language. Hence, it is essential to experiment with language specific features of Sinhala in sentiment classification. With this purpose, this chapter focuses on experimenting language specific features to the Sinhala starting from simple parts of speech to complex structural analysis.

Parts of speech (POS)s are the linguistic representations of lexical items in a sentence or phrase. Noun, Verbs, Adjectives, and Adverbs are commonly explored linguistic features in language processing and text mining. In advanced investigations, a sentence is further divided into phrases known as chunks. Structural features such as flow shifters combine these chunks. In this study, sentiment analysis specific features, such as, intensifiers and shifters that are extensively present in Sinhala opinions are examined.

In this chapter section, 7.2 discuss the impact of adjectives and adverbs followed by the role of Sinhala negation in sentiment classification in section 7.3. Sections 7.4, Scope modeling such as contextual features investigates in details. A novel morphological approach for sentiment classification for Sinhala elaborated in section 7.5. Finally, the summary of the chapter presented in section 7.6.

7.2 Impact of Adjectives and Adverbs

Most of the recent sentiment analysis researchers have argued that adjectives and adverbs are the most influential parts speech (POS) in sentiment classification (Benamara, Cesarano, & Reforgiato, 2007) . With this hypothesis, an initial experiment was conducted using adjectives and adverbs with their prior polarities taken from constructed lexicons. In general, a special tool known as part of speech tagger detects any part of speech in a text. As mentioned earlier, Sinhala is regarded to be a less-resourced language (in the context of language processing),

currently, there is no tagger system available for the language. Therefore, a gold standard Sinhala adjective and adverb list were utilized in these experiments to tag the adjectives and adverbs. The adjective and adverb lists were compiled by analyzing 10 million corpora generated by a leading language academic research group affiliated with the University of Colombo in Sri Lanka (Language Technology Research Laboratory, 2011). The adjective list consists of 7,503 items while 671 items comprise the adverb list. These adjectives and adverbs were examined in 2,083 Sinhala opinions (sample set used in this study) by running a parser (Appendix E) in the initial step of this investigation. Of 2,083 opinions, 929 (45%) contained at least one adjective, and only 292 (14%) opinions consisted of one adverb. The sample opinions dataset contained only 4.78% of the 7,503 possible Sinhala adjectives. The most frequent adjective found in the sample is ලොකු (huge). Eleven percent of the adverbs described in the main list are found in the opinion dataset (sample) with දැන් (now) as the most frequent adverb. In the manual classification of the adjectives extracted from the opinions, it was found that the list consisted not only of pure adjectives but also some intensifiers and negation shifters. These intensifiers and shifters explained in section 7.4.

The experiments began by examining the impact of the adjectives and adverbs alone, that is, one at a time as well as in the combination of twos. The main aim was to examine the impact of adjectives and adverbs in polarity classification using machine learning methods. With this expectation, the prior polarities of the pure adjectives (other than the intensifiers and shifters) were assigned using the lexicon constructed for this research which explained in chapter 5.

7.2.1 Impact Adjectives

Adjectives are classifiable into five classes, and they are:

i. Descriptive Adjectives
ii. Possessive Adjectives
iii. Numeral Adjectives
iv. Demonstrative Adjectives
v. Interrogative Adjectives

Of these, Descriptive (e.g., brilliant, awesome, etc.) and Numeral (e.g., many, each, etc.) are relevant to sentiment classification. The descriptive adjectives explain the sort or quality of the noun. The polarity of the noun depends on the descriptive adjective prior to (immediately before) the noun. As an example, the phrase “good students” leads to positive polarity and “bad

students" would be classified as negative. However, there are cases in which an adjective appears in a sentence after a noun. These occur when describing a person or entity.

In Sinhala typically, the descriptive adjective is in immediately before or after the noun. However, there are some situations where the adjective occurs as the final word in the sentence. For example, in the following comment, the adjective "හොඳයි" (good) is functioned at the end of the sentence resulting in a positive comment. In this case, the stem adjective "හොඳ" (good) has been inflected into "හොඳයි" by adding the morpheme "යි". This form of word, the word with "යි" will always be at the end of a sentence in written Sinhala.

"ඉන්දියාවට වැඩිය චීනය ගොඩක් හොඳයි" (China is much better than India)

Numerical adjectives are divided into three classes; definite, indefinite and distributive numeral adjectives. It is the conjecture of the researcher that the impact of the definite adjectives in sentiment classification is insignificant when compared with the other two numerical adjectives. The definite adjectives, such as "ten" and "twelve" in the expressions "ten credits" and "twelve credits," are sentimentally negligible as both are neutral comments depending on the context. On the other hand, the influence of indefinite adjectives is much higher than that of the other two; definite and distributive. For example, the words "much" and "little" are indefinite adjectives that can influence the polarity of an opinion. In addition to these three classes, Sinhala has a special type of adjective defined as a Nominal adjective (Dileep, 2010). This class has no impact on sentiment classification. For example, the nominal adjective "පඩි" (stairs) describes the noun "පෙල" (steps), and is neutral in polarity and may be combined to form a compound noun "පඩි පෙල". Verbal adjectives are another kind of adjective in Sinhala that are also used to form compound nouns and also result in a sentiment polarity of zero.

With the aim of finding the effects of Sinhala adjectives and the type that is most relevant to sentiment classification, a set of experiments were conducted using the sample opinions described in chapter 4 section 4.3. The adjectives and their distribution details within the sample are presented in figure 7.1.

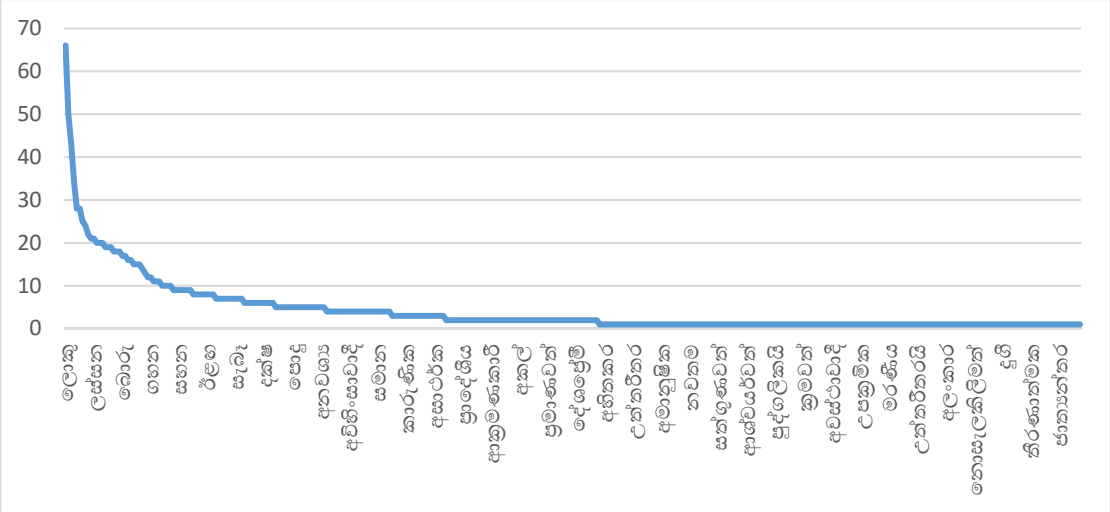


Figure 7.1: Adjective Distribution

From an in depth investigation of the list of 358 of adjectives available in the sample, it was found that majority of them were descriptive adjectives (85%). Of these descriptive adjectives, following are the sentences ending in adjectives which are inflected by morpheme "යි".

Table 7.1: Sentence ending Adjectives

Adjective	Stem	Polarity	Translation
හොඳයි	හොඳ	+	Good
ලස්සනයි	ලස්සන	+	Beautiful
කැතයි	කැත	-	Ugly
අමාරුයි	අමාරු	-	Difficult
පහසුයි	පහසු	+	Easy
නිවැරදියි	නිවැරදි	+	Right
විරුද්ධයි	විරුද්ධ	-	Against
විශාලයි	විශාල	+/-	Great
පැහැදිලියි	පැහැදිලි	+	Clearly
අවශ්යයි	අවශ්ය	+	Need
සැහැල්ලුයි	සැහැල්ලු	+	Light
පුද්ගලිකයි	පුද්ගලික		Is private
අවිවාදිතයි	අවිවාදිත	+	Indisputable
නිශ්ශබ්දයි	නිශ්ශබ්ද	+	Silent
උත්තරීතරයි	උත්තරීතර	+	Supreme
නරකයි	නරක	-	Bad

Very few (8%) of the Sinhala indefinite adjectives were present in the sample. However, these indefinite adjectives were more frequent than descriptive adjectives.

In the initial investigation, the performances of the classification based only on adjectives are shown in Table 7.2. The classification was carried out for both Naïve Bayes and SVM algorithms. The feature set consists of all adjectives extracted from the opinion data.

Table 7.2: Classification by Adjectives

Features and Weighting	Accuracy		F-Measure	
	Naïve Bayes	SVM	Naïve Bayes	SVM
adjectives & binary	55.585	55.186	0.549	0.545
adjectives & *tfidf*	55.377	55.377	0.634	0.696
adjectives & polarity score	52.613	52.787	0.517	0.528

The above results indicate that there is no significant improvement on classification accuracies by adjectives alone compared to keyword based classification (See the Tables 6.2 & 6.3, Chapter 6). However, recall for the negative category by the SVM with tfidf weights gives the highest value achieved so far in this research of 0.974. At the same time, the highest accuracy obtained using the tfidf weighting was with SVM which indicates the suitability of such weighting in the SVM based classification of Sinhala opinions.

It is also interesting to examine the type of adjectives that dominate the classification. As mentioned earlier, descriptive and indefinite adjectives are assumed to be relevant in the sentiment classification. To test the hypothesis, an experiment was conducted only with descriptive adjectives and the classification accuracies were compared. The classification performance measures are presented in Table 7.3.

Table 7.3: Classification by Descriptive Adjectives

Features and Weighting	Accuracy		F-Measure	
	Naïve Bayes	SVM	Naïve Bayes	SVM
Descriptive adjectives & binary	55.556	54.106	0.551	0.534
Descriptive adjectives & tfidf	54.559	55.755	0.528	0.552
Descriptive adjectives & polarity score	55.769	55.962	0.556	0.553

The only improvement observed when dropping the indefinite adjectives were shown in the trial where the descriptive adjectives with prior probabilities were used. The comparison of Tables 7.2 and 7.3 reveal that use of adjectives by themselves for sentiment classification is not a viable approach (See the Table 6.6 Chapter 6).

7.2.2 Impact of Adverbs

Adverbs are another part of speech that may effect on the sentiment of an opinion. Linguists classify adverbs differently depending on the application domain. However, semantically adverbs are sub-grouped into six categories:

i. Adverb of Time
ii. Adverb of Place
iii. Adverb of Manner
iv. Adverb of Frequency
v. Adverb of Probability
vi. Adverb of Degree

Unlike adjectives, most of the above adverb categories were assumed to be relevant to the sentiment classification except for adverbs of time and place. Of these categories, adverbs of degree seemed to be most relevant (Benamara, Cesarano, & Reforgiato, 2007). As in the study of adjectives, the adverbs available in the sample extracted from the opinions are summarized in figure 7.2.

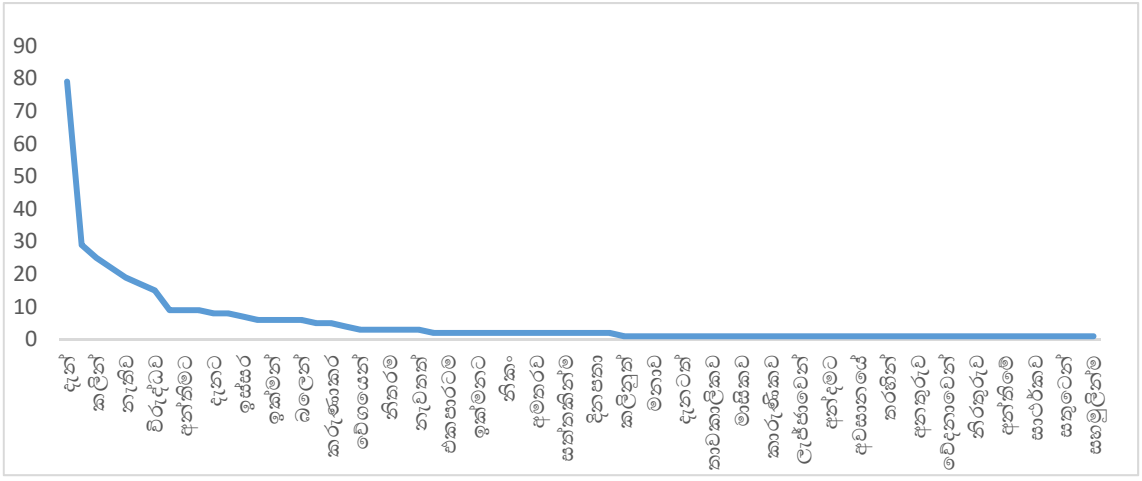


Figure 7.2: Adverb Distribution

Two adverbs of time "දැන්" (now) and "කලින්" (earlier) were the most frequent in the in the sample opinions. The distribution of the 71 adverbs present in the sample based on adverb type is given in table 7.4.

Table 7.4: Adverb Distribution

Type of adverb	# Occurrences in the sample
Degree	10
Manner	41
Time	17
Frequency	1
Location	2

Based on the adverb type distribution in the adverb list, it appears that there are more adverbs of manner available in the Sinhala opinions sample than the other types, namely, degree, time, frequency and location.

The most frequent adverb of manner is "එකට" (together) in the opinion collection. Even though the word "එකට" (together) is classified as an adverb in literal Sinhala, it has been used to denote a particular person or incident. As an example, the word "එකට" (together) in the following sentence refers to an incident explained in the opinion; ඉන්දීය අගමැති තුමා අවේ නැති (Indian prime minister didn't come).

"ඉන්දීය අගමැති තුමා අවේ නැති එකට දුකයි" (It is sad that the Indian prime minister didn't come)

On the other hand, the adverb refers to a noun in the following sentence which is expressed in English transliterated form.

"ගමේ දුප්පත් දරුව රොයල් එකට යනවට අති ඉරිස්සියව" (A jealousy of a poor boy attending Royal College)

It is clear from the above example and explanation that different techniques are needed to incorporate the adverb "එකට" (together) in sentiment classification. As explained in above example the function of the word is complex and in some cases, it functions as an adjective. Due to the complexity of the word "එකට," (together) the analysis of the term in sentiment classification is not explored in this research and is left to future research.

Some researchers in sentiment classification for English believed that adverbs of degree is the most relevant type of adverb for sentiment classification (Benamara, Cesarano, & Reforgiato, 2007). In the sample considered in this research, it was found that adverbs of affirmation, are

common in the sample. The class adverb of affirmation is a subclass of adverb of degree. Adverbs of affirmation included words such as "අනිවායර්යෙන්" (compulsory), "තරයේ" (thoroughly), "සත්තකින්ම" (absolutely), "නිවැරදිව" (exactly), "සම්පූර්ණයෙන්" (fully) and "සහමුලින්ම" (totally). Among these adverbs සත්තකින්ම" (absolutely) and "නිවැරදිව" (exactly) are positive in all contexts. The polarity of the other adbverbs depends on the context. To understand the function of these context dependant adverbs, the post position of the term is examined in this research. It was observed that the polarity of the post position of the adverb "අනිවායර්යෙන්" (compulsory) is more negative than positive. However, the impact of the word "තරයේ" (thoroughly) is insignificant as it is always followed by a neutral word.

An experiment was conducted using adverbs (only) next to examine the effectiveness of Sinhala adverbs on sentiment classification without considering the context.

Table 7.5: Classification by Adverbs

Features and Weighting	Accuracy		F-Measure	
	Naïve Bayes	**SVM**	**Naïve Bayes**	**SVM**
Adverbs & binary	52.893	58.677	0.470	0.499
Adverbs & *tfidf*	61.303	61.303	0.559	0.500
Adverbs & polarity score	61.487	62.838	0.584	0.576

The results of the experiments show that the impact of adverbs on classification is significant as the accuracies improved for both Naïve Bayes and SVM (Table 7.5). The highest accuracy is achieved when using prior polarities. The results indicate that the adverbs alone are better than the adjectives alone. With this promising result, the next step is to identify the type of adjectives that contributed the most to the classification accuracy. To find out this, a classification was carried with each type of adverb independently using the prior polarity weighting method. The prior polarities method was used in these experiments because it gave the highest accuracy for sentiment analysis using all adverb types. The results are tabulated in Table 7.6.

Table 7.6: Classification using different adverb types

	Accuracy		F-Measure	
Features and Weighting	**Naïve Bayes**	**SVM**	**Naïve Bayes**	**SVM**
Adverb of Manner	62.626	62.626	0.597	0.625
Adverb of Degree	62.201	58.373	0.605	0.583
Adverb of Time	59.909	59.909	0.545	0.598

The results in Table 7.6 reveal that the impact of each of the different adverb types on Sinhala sentiment classification performance is significant o when compared with the combined adverbs (Table 7.5). Thus, it was concluded that in the case of Sinhala all adverb types are important in sentiment analysis.

Given these findings, it was decided that it would be interesting to examine the success of the classification with both adjectives and adverbs.

Table 7.7: Classification by Adjectives and Adverbs

	Accuracy		F-Measure	
Features and Weighting	**Naïve Bayes**	**SVM**	**Naïve Bayes**	**SVM**
Adjectives + Adverbs & binary	53.359	54.183	0.524	0.492
Adjectives + Adverbs & *tfidf*	53.889	56.444	0.516	0.548
Adjectives + Adverbs & polarity score	59.935	55.339	0.535	0.508

The results presented in Table 7.7 indicate that there is a significant drop in classification performance when both POS (adjectives and adverbs) are used as features when compared to using only adjectives or only adverbs. This finding contradicts the results of the research of Benamara, Cesarano, & Reforgiato (2007) who carried out an evaluation of the POS for English sentiment classification. For English, it was reported that, when used with some scoring axioms, adjectives and adverbs combined to give better classification results than using adjectives alone. One of the reasons why combined POSs are not giving the better results for Sinhala opinions is possibly the inappropriate feature weighting method in the study. In the above analysis for Sinhala opinions, the weighting was based purely on the occurrences of the adjectives and adverbs using statistical measures. The dominance of these features can be measured extensively using advanced contextual as well as scoring techniques and it is likely that a contextual approach would give better results.

7.3 Role of Negation

Having gained some knowledge of how the main parts of speech, namely adjectives and adverbs, influence sentiment classification in Sinhala it was decided to next investigate the impact of negations in classification. Negation affects polarity determination extensively. Therefore, it is important to consider the impact of these lexical constructions in sentiment classification. The investigation of negation was carried out on three aspects namely word detection, the level of representation, and scope of negation.

Negation word identification is a quite complex task because the polarity of a word is not only reversed by the syntactic negations (such as “no’, “not”, etc.), but also a negation calculation which involves knowledge of the lexical patterns of prefixes, suffixes, and contextual valence shifters. Contextual valence shifters include intensifiers and diminishes. These shifters can increase or decrease the intensity of a word and hence determine the overall polarity of a sentence. In this study, possible syntactic negations in the Sinhala language are first identified and listed. With the help of expert Sinhala linguists, and based on the author’s understanding of the language, widely used negation words were identified. In addition to the syntactic (functional) negators (explained in the chapter 4 section 4.5.2), a spoken word “එපා” (do not) was found to be frequently used in the sample opinion data. Syntactically, this “එපා” (do not) negator is used after a verb in a sentence. The sentence below is a negative opinion and the negator “එපා” (do not) changes the neutral verb “දෙන්න” (give) to negative.

“අපේ රටේ කිසිම සාකච්ඡාවකට අවසර දෙන්න එපා” (Do not allow for any discussion in our country)

All effective negators in the Sinhala language, from a linguistic point of view, are words that twist the polarity of an opinion. Effective negators can be classified into two groups. In this thesis, the negators have a direct impact on a word are referred to as **based negators** and those that impact on the phrase of a sentence are called **contextual negators**.

The following sentence illustrates the contextual negator in an opinion.

> එයා එන්නේ නැත්නම් වැඩ කරගන්න පහසුයි
> (It is easy to work in his absence)

The opinion contains two phrases, both combined, and the first one is the explanation of negativity of the second one. The negators of both groups are presented in Table 7.8.

Table 7.8: Negators in Sinhala

Base Negators	නෑ නෑ, නැහැ, නැත, නැති (all means no or not) බෑ, බෑ, බැහැ, බැරි, බැරිය (all means can't) එපා (don't)
Contextual Negators	නැත්නම් (or), නැතැයි (no), නැතිනම් (unless), නැතත් (or not), නැතොත් (unless), නැතහොත් (or), නැතුව (without)

7.3.1 Impact of the base negators by artificial feature modeling

The impact of the base negators is mainly on a single word in pre-position whereas, contextual negators effect the post-position of a phrase. For example, "හොඳ නෑ" (not good) is a negative expression where the negator "නෑ" (no) operates on pre-position word "හොඳ" (good).

Both categories were first analysed separately in order to measure their impact on classification accuracy.

One of the simplest methods of finding the impact of negators is by introducing an artificial word that represents a negation (Wiegand, Balahur, Klakow, Roth, & Montoyo, 2010). In this approach, an artificial word for example "w" is followed by a negation NOT. Consequently, an artificial word is created "w_NOT". Pang et al. (2002) considered every word in the sentence after the NOT and replaced each word with the artificial word until the next punctuation mark was reached. In this initial experiment, only the word affected by the negation word is considered. The context of the negators experimented in the scope modeling section (7.3) of the chapter. In this research, the impact of base negation words was tested using a bag-of-word representation including the artificial word as one of the feature terms.

The study of the Sinhala negation began by identifying the polarity of the words adjacent to the base negators. That is, the experiment tries to identify the scope of the base negators by examining the adjacent words. Linguistically, all negators except for "බැරි" (can't) function on the pre-position adjective, noun or verb. The negator "බැරි" (can't) effects on both pre and postpositions. The scope of the part of speech of adjacent words for these base negators has already been explained in table 4.5.2 in chapter 4. The investigation discussed in this section extends the previous experiments in order to identify the polarity of the pre-position words. By scanning through the sample of opinions used in this study it was discovered that 80% of pre-

position lexical items were positive. That is, predominantly base negators immediately affect the word preceding them and change the polarity of the affected word. However, there are some cases where negative words also follow negators. In such a situation, the polarity of the word also reverses and becomes positive. In general, a positive polarity is changed to negative if a base positive word is followed by a base negator, and vice versa. Additionally, an interesting pattern was discovered where a word; "කමක්" (no translation word in English) followed by a negator its polarity reverted to positive. This pattern is observed largely in spoken dialect. These patterns/rules are illustrated in figure 7.3.

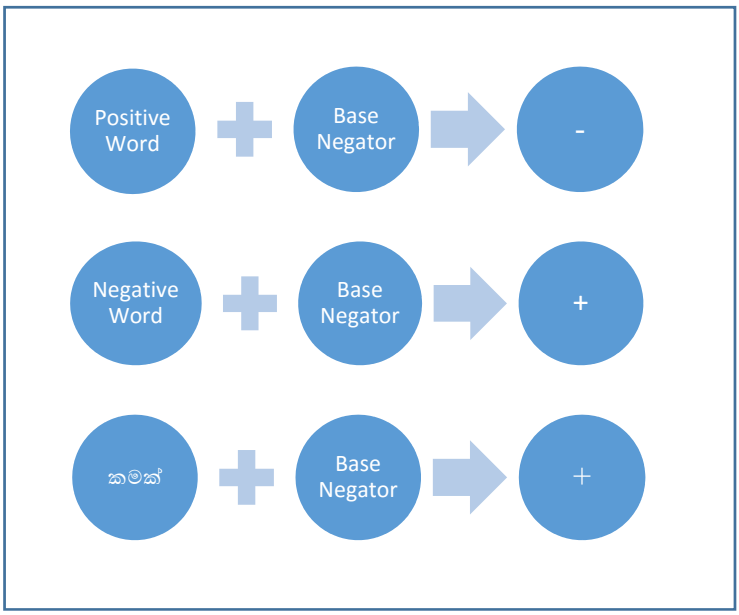


Figure 7.3: Negators in Sinhala

In this research, all the base negators in the opinions were tagged by introducing an artificial word; POS_NOT, NEG_NOT, and KMK_NOT. These tags were formed by combining the base negator with the affected word using the following proposed rules. These novel rules were tested using supervised classification methods.

Rule 1: If the pre-position of the base negator is positive then both the affected word and the base negator are replaced by POS_NOT

{eg: හොඳ (good) නෑ (no) was tagged as POS_NOT}

Rule 2: If the pre-position of the base negator is negative then both affected word and the base negator are replaced by NEG_NOT

{eg: වරදක් (mistake) නෑ (no) was tagged as NEG_NOT}

Rule 3: If the pre-position of the base negator is "කමක්" then both affected word and the base negator are replaced by: KMK_NOT

{eg: කමක් නෑ (no) was tagged as KMK_NOT}

Rule 4: If the pre-position of the base negator is a noun or verb (neutral) then both affected word and the base negator are replaced by PURE_NOT

{eg: ගෙදර (home) නෑ (no) was tagged as PURE_NOT}

The impact of the base negators was investigated using three approaches. In the first approach, the effectiveness of the items was measured without incorporating any other linguistic features. The purpose of the experiment was to understand the impact of the negators on classification in isolation. The approach used was the simplest as it considers the bag-of-word method where the features consist of the keywords in the sample including the base negators. Again the Naïve Bayes and SVM classification algorithms with tfidf feature weightings were used. The experimental results are given in Table 7.9.

Table 7.9: Impact of base negators

Test Case	Naïve Bayes	SVM
(a) All keywords including base negators	57.143	60.114
(b) Feature Selection (correlation)	58.913	62.326
(c) Keywords+ POS_NOT + NEG_NOT	56.827	61.568
(d) Keywords+ NEG_NOT + KMK_NOT	57.081	57.585

The initial experiment was conducted with 464 keywords and the base negators tagged in the test data as explained in figure 7.3. The keyword list was selected based on a relative frequency using a minimum threshold that was also used to select the base negators. The inclusion of base negators was found to effective as it shows average performance.

In the next step, features were selected using the feature selection algorithm; correlation based algorithm as it was found to be the best feature selection for unigram features (section 6.4.3, chapter 6).

Interestingly the feature set (by correlation based algorithm) included all base negators. It was indicated (see Table 7.9(b)) that base negators were important in the classification as shown by the increase of accuracies for both Naïve Bayes (58.913) and SVM (62.326). Moreover, the effect of pure negators in the analysis is examined by dropping the NEG_NOT and KMK_NOT which are positive in the experiment (Table 7.9, (c)). The classification accuracies show that these negators affect the accuracies as it decreases, but the difference is not significant. The confusion matrix generated for the SVM classification showed that 87% of the positive

opinions were correctly classified. On the other hand, the Naïve Bayes algorithm having base negators in the classification vector correctly classified 94% of the negative opinions. Both SVM and Naïve Bayes classification labeled more than 92% of the negative opinions, correctly identified as negative even though base negators were not in the classification feature vector. This confirmed that the base negators are essential for classification.

In the second approach, the additional linguistic features of adjective and adverbs were introduced. Adjectives and adverbs were included in the classification vector in order to examine the interaction of negators on the basic parts of speech in sentiment classification. In this experiment, six features were included in the vector; adjectives (Adj), adverbs (Adv), and the negators POS_NOT, NEG_NOT, KMK_NOT, and PURE_NOT. The classification accuracies are given in Table 7.10.

Table 7.10: Impact of Base Negators on POS

Test Case	Naïve Bayes	SVM
Adj + Adv	58.248	59.170
Adj + Adv + POS_NOT + PURE_NOT	57.143	59.631
Adj + Adv + NEG_NOT+ KMK_NOT	58.157	58.065
Adj + Adv + POS_NOT + PURE_NOT+ NEG_NOT + KMK_NOT	57.604	58.710
Adj + POS_NOT + NEG_NOT	58.065	58.525

No improvement over keyword-based classification was observed by adding the linguistic features of adjectives, adverbs, and negators. In fact, there was a decrease in classification accuracies across the board. On the other hand, adjectives and adverbs by themselves (without base negators) performed better. In further investigation, a tree based algorithm (J48) was used setting the parameters to default sate to examine the linguistic feature dependencies in sentiment classification. Figure 7.4 provided a simple visualisation of the resultant decision tree generated by Weka using tfidf weights.

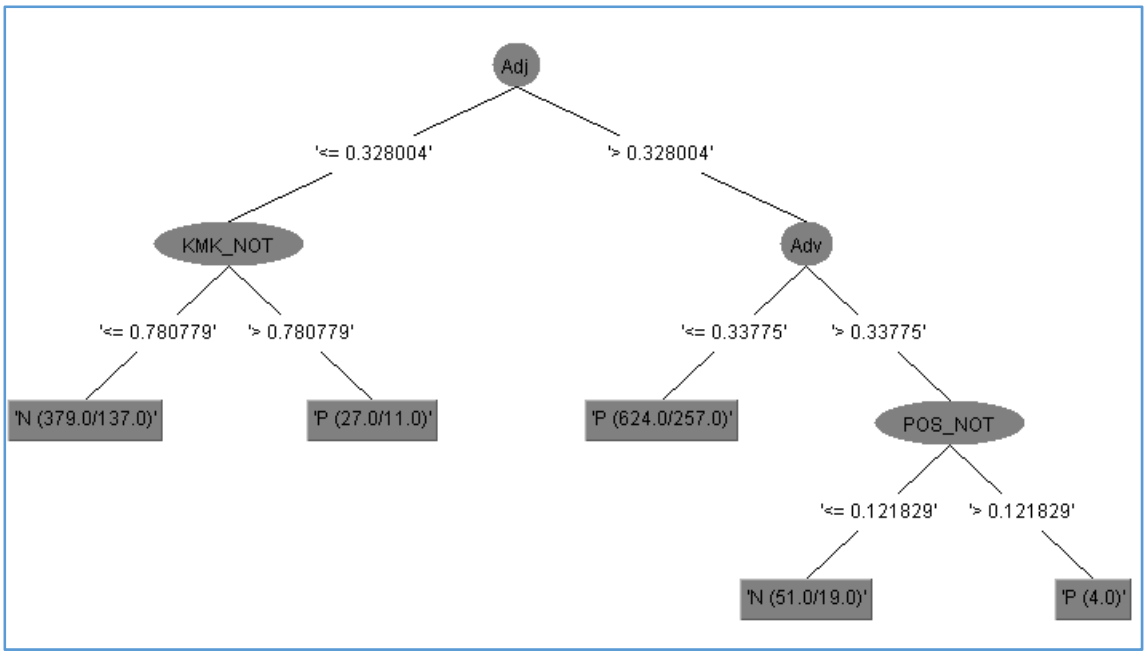


Figure 7.4: J48 Decision tree illustrating the impact of base negators on adjectives (Adj) and adverbs (Adv)

Figure 7.4 reveals that adjectives (Adj) initially determine the polarity as the feature or attribute which provides the highest normalized information gain. The next level of the decision tree again determined based on normalised information gain, use adverbs (Adv) for one branch and the negative marker KMK_NOT for the other branch of the tree. The right hand branch is further split based on the POS_NOT negator.

The branch of the tree Adj (root) → Adv (node) results in a total of 624 instances being classified as positive at the leaf of which 367 are correctly classified as positive (58.8%) and 257 were misclassified as positive. This branch of the tree correctly classifies the majority of the positive opinions in the data set (84%). This concludes that Adv(adjectives) more effective in classifying positive opinions. The Adj→KMK_NOT branch classifies the majority of the negative opinions in the data (45%). The figures indicate that KMK_NOT is influencing the negative opinion classification even though it is a positive tag.

The third approach in this series of experiments incorporates the polarity of adjectives and adverbs in the classification. Adjectives and the adverbs were annotated with their polarity as either positive (1) or negative (-1) using the positive, negative list generated in chapter 5, section 5.6. The base negators were also tagged as either positive (1) or negative (-1) as denoted in figure 7.3. That is, KMK_NOT and NEG_NOT are tagged as positive (1), and PURE_NOT and POS_NOT are tagged as negative (-1). The classification accuracy achieved in these experiments are presented in Table 7.11.

Table 7.11: Impact of base negators with polarity of adjectives and adverbs

Test Case	Naïve Bayes	SVM
Adjectives + Adverbs	60.615	60.718
Adjectives + Adverbs + All base negators	62.359	63.897
Adjectives + Adverbs + POS_NOT + PURE_NOT	62.359	64.205
Adjectives + Adverbs + NEG_NOT+ KMK_NOT	60.410	61.539

The results show that the impact of the polarity of the linguistic features on classification accuracy is higher than other statistical based measures such as base negators. Adjectives and adverbs with base negators give better accuracies than adjectives and adverbs alone. The experiment using only negative bearing features, POS_NOT and PURE_NOT, with adjectives and adverbs achieved the highest accuracy. The accuracy decreased when the classification vector included the positive bearing negators; NEG_NOT and KMK_NOT. This revel that adding these base negators not improve the classification. On the other hand, POS_NOT and PURE_NOT highly impact for the accuracies.

With this improvement, all the sample opinions are scanned through to examine the context of the base negators and their distribution (positive/negative). Table 7.12 gives the distribution of base negators based on the polarity of opinions and the negators.

Table 7.12: Distribution of base negators

Opinions (%)	POS_NOT	NEG_NOT	KMK_NOT	PURE_NOT
# Positive Opinions	83.48	0.87	14.78	0.87
# Negative Opinions	85.04	2.99	8.12	3.85

Based on the distribution it can be concluded that the base negator POS_NOT is equally present in both positive and negative opinions (Table 7.12). The high presence of POS_NOT in the positive comments reveals that there could be another possible linguistic feature that determines the polarity of these comments. The negator KMK_NOT, which shifts the polarity from negative to positive appears in a considerably higher proportion of positive comments than negative comments. Also, the context of the dominant base negator, POS_NOT on the context where it appears was examined. The context of adjectives and adverbs was taken into consideration in terms of the pre-and post-position of the negator. Table 7.13 presents the context of the adjectives.

Table 7.13: Context of POS_NOT on adjectives

Opinions (%)	Post-position Adjectives	Pre-Position Adjectives
# Positive Opinions	39.19	60.81
# Negative Opinions	37.60	62.40

The table 7.13 shows that the impact of the POST_NOT on pre-position adjectives was much higher than in the post-position in both positive and negative opinions. The similarity in the frequency of pre- and post-position adjective in both positive and negative opinions mean that it is difficult to identify a rule based on the context of base negators such as POS_NOT that governs the polarity of an opinion. With the aim of mining a rule that determines the polarity of a comment, the polarity of the word that precedes the POS_NOT is investigated. It has been noted that POS_NOTs do not always precede an adjective or an adverb. The preceding lexical items can also be nouns or verbs. All adjectives (100%) which appeared before the POS_NOT negator were positive for all the positive comments which included the negator. Similarly, all the adverbs before the negator are positive for positive opinions. Furthermore, negative adjectives mostly follow a base negator. However, negative adverbs do not have a regular pattern. No significant pattern is shown in the negative comments which included base negators such as POS_NOT.

In summary, this section investigated a method for handling negation where an artificial word (feature) is included in the comments with the purpose of evaluating the impact of negators on classification. The drawback of this method is the immediate consideration of the context of the negation, neglecting the conjunction of two phrases when present in an opinion. In other words, the scope of the negation is limited in this modeling approach. The improvement in classification accuracy achieved using this method is insignificant when compared with the simple bag of words approach (chapter 6, table 6.4) where negation is not considered at all.

Despite the fact that introducing artificial words for negation is not an appropriate approach for negation modeling as seen in the above section, the experiment helped to identify the importance of negation to word polarity. The improvement shown in Table 7.11 compared to 7.9 is promising and opens a path for further investigation using other methods for including negation using polarities in the classification.

7.3.2 Negation modeling using subjective lexicon

In the experiments reported in this section, the polarity of the word is incorporated into negation modeling. The proposed approach uses the subjective lexicon constructed in the chapter 5 section 5.5. In this method, as the initial step, polarity scores are assigned for each word in the sample opinions. Then the following rules are applied to represent the impact of the negators' for each opinion.

> *Rule 5: If a word was followed by the negator then the score was multiplied by -2*
>
> *Rule 6: If a word was followed by the KMK_NOT (කමක් නෑ) then the score was multiplied by 2 otherwise the score remains unchanged*

Rule 5 adds the impact of the pure negators' effect on the words before the negators. Rule 6 considers the significance of the KMK_NOT in sentiment classification. Finally, the classification vector was defined with two features one for a total positive score and the other feature calculates the total of a negative score. Classification by SVM and Naïve Bayes gave promising results that explain the impact of the negators and KMK_NOT. The highest accuracy of 61.6% was achieved using SVM classification, and more than 77% of the positive comments were correctly classified suggesting that the negator KMK_NOT was reasonably well handled by this approach.

7.4 Scope Modelling

Exploring the impact of the different linguistic features other than the main part of speeches such as noun, verb, adjectives and adverbs is an important task in sentiment analysis. The influence of intensifiers and shifters in the context of an opinion is known as scope modeling. In addition to negation scope modeling, the scope of intensifiers and shifters were investigated and are elaborated on in the following section.

7.4.1 The impact of Contextual Intensifiers

As mentioned in chapter 4 section 4.5.2, contextual intensifiers are of two types; increasing and decreasing. The six increasing intensifiers include වඩා (more), වඩාත් (more), ගොඩක් (more), විශාල(big), ලොකු (huge), ඉතා (very), ඉතාමත් (very) බොහොම (huge). These intensifiers have an effect on the next (following word the) by increasing the polarity of the next word. The intensifiers, කුඩා (little), සුළු (small), පොඩි (small), and පොඩ්ඩක් (a little) are

diminishing intensifiers that weaken the polarity of the words. The impact of both types of intensifier on sentiment classification are investigated by applying the following rules in addition to the rules outlined in section 7.3.2.

Rule 7: If an increasing intensifier is followed by a positive word then the sentiment score is multiplied by 2

Rule 8: If an increasing intensifier is followed by a negative word then the sentiment score is multiplied by -2

Rule 9: If a decreasing intensifier is followed by a positive word then the sentiment score is multiplied by -2

Rule 10: If a decreasing intensifier is followed by a negative word then sentiment score is multiplied by 2

Otherwise, the score remains unchanged.

The classification results obtained indicate that the impact of the intensifiers is significant in sentiment classification in Sinhala. The classification accuracies are increased by 4% (Table 7.14) when compared to the performances achieved using the negation rules explained in section 7.3.2.

Table 7.14: Impact of Intensifiers

Algorithm	Accuracy
Naïve Bayes	65.758
SVM	65.529

According to the table7.14, approximately equal performances by Naïve Bayes and SVM when handling the intensifier in Sinhala. More than 66% of negative opinions were correctly identified by the proposed rules. Approximately similar accuracies were observed (63%) for positive comments.

In exploring the sample data, it was noted that shifter features also occur in comments that combine two phrases. In the next section, the impact of the contextual shifters on sentiment classification is examined.

7.4.2 Impact of Contextual Shifters

The contextual shifters considered in this study are negation and flow shifters. Seven negation shifters and 15 flow shifters have been identified in the Sinhala language. A detailed explanation of both shifters is given in chapter 4 section 4.5.2. During this research, the influence of negation shifters in sentiment analysis was investigated. It was found that the

context of the negation shifters is a critical component of polarity change in a sentence (section 4.5.2). The following table gives the context of each negation shifters on pre and post position of a phrase.

Table 7.15: Context of negation shifters

Negation Shifter	Effective position
නැත්නම් (or not)	pre
නැතිනම් (unless)	post
නැතත් (whether)	pre
නැතුව (without)	pre
නැතැයි (if not)	pre
නැතොත් (unless)	pre
නැතහොත් (or)	pre

Defining a rule for polarity determination for the opinions which contain the negation shifter නැත්නම් (or not) was comparatively more complex than for other linguistic features. This shifter is used to justify the main argument of the opinion. As an example, in the following sentence the, shifter නැත්නම් (or not) is used to explain the reason for the preceding phrase (it is easy to work).

> එයා එන්නේ **නැත්නම්** වැඩ කරගන්න පහසුයි
> (It is easy to work if he did not come)

In this case, the sentiment of the phrase (or words) before the shifter is negligible. However, in the following sentence, the impact of the shifter is important to the overall polarity of the comments.

> රට දියුණු වෙන්න තිබුනා යුද්දයක් ඇති කලේ **නැත්නම්**
> (The country would have developed unless a war is created)

In this sentence, the first phrase "රට දියුණු වෙන්න තිබුනා" (The country would have developed) is a positive phrase however,, the overall comment is a negative. This example illustrated the importance of identifying the correct phrase that is impacted by the shifter and that this identification is a highly complex problem. In the first example, it is clear that the phrase that affected the shifter is before the negation shifter. However, in the second example the impact on the second phrase which is difficult to separate from the first unless a punctuation mark mentioned is added in between the phrases. The phrases can be identified using sentence parsing or chunking methods based on the dependency relation of morphemes (Jang & Shin, 2011). As Sinhala is a considered as a less resourced language in the context of language

processing these methods are yet to be developed. Due to these complexities the investigation of these negation shifters is left for future research in sentiment classification in Sinhala.

7.4.3 Impact of Flow Shifters

The research has identified 15 flow shifters in Sinhala which will intensify or weaken the polarity of a comment at the phrase level. As mentioned in chapter 4 section 4.5.2, only seven shifters are frequent in the sample considered in the study. The impact of the flow shifters is more on the phrases than the single word as negation shifters. Since the influence of the flow shifters is on the phrases, the classification was also performed using compositional semantics. Compositional semantics explores the structural inferences of a sentence by considering the phrases (Liang & Potts, 2015). The impact of each flow shifter on the pre- and post-phrase of the comments identified linguistically are tabulated in the table 7.16

Table 7.16: Context of flow shifters

Flow Shifter	Effective position
එහෙත්(but)	pre
ඒත්(but)	pre
හැබැයි(but)	post
නම්(if)	pre
නිසා(because)	pre
බැවින්(because)	pre
අනුව(according to the)	pre

The first four flow shifters change the polarity of the phrase which occurred before the shifter. On the other hand, rest of the features explain the author's empathetic point on the comment giving a reason in the phrase before the shifter. The impact of these flow shifters in sentiment classification is linguistically not significant, and therefore the investigation is carried out only using the first four linguistic features. The rules for understanding the shifters are compiled by investigating the occurrence of each in the comments. The first three shifters tended to weaken the polarity of the first phrase giving more weight to the second phrase. In the following example opinion, the negativity of the first phrase (F_Phrase) is diminished by the shifter while at the same time adding positive sense to the second phrase (S_Phrase).

{කොමිස් ගැහිල්ල කවුරු ආවත් කරනවා} (F_Phrase)
එහෙත්
(S_Phrase) {ඉන්දියාවට වඩා චීනය අපට ජාත්‍යන්තරව ගොඩාක් උදව කරනවා}
(Everybody takes a commission, ***but*** *China helps us more than India)*

However, in a second example given below, the first phrase is positive and the effective content of the comments in the second phrase is negative.

{සොභාවික ආරක්ෂාව ගැන කියාලා තිබුනා}
එහෙත්
{එය ක්‍රියාත්මක කිරීමෙන් නිලදාරින්ට කොමිස් ලැබෙන්නේ නැහැ.}
(had told the natural protection ***but*** *not getting the commission for the officials by implementing it)*

A similar pattern was observed for the shifters ඒත්(but), ඒත්(but), හැබැයි(but) and නම්(if).

Having gained knowledge of how the shifters act on comments, the following rules were introduced to handle these shifters. These rules are implemented on the phrases basis total sentiment rather than on their lexical semantics. The rules that describe the compositional semantics are as follows;

tot_{F_Phrase} = total sentiment score of the phrase before the shifter

tot_{S_Phrase} = total sentiment score of the phrase after the shifter

Rule 7: If $tot_{F\_Phrase} < 0$ *and* $tot_{S\_Phrase} > 0$ *then*

tot_Positive Score = $1.5 * tot_{S\_Phrase}$

tot_Negative Score = tot_{F_Phrase}

Rule 8: If $tot_{F\_Phrase} > 0$ *then* $tot_{S\_Phrase} < 0$ then

tot_Positive Score = tot_{S_Phrase}

tot_Negative Score = $1.5 * tot_{F\_Phrase}$

Else: tot_Positive Score = Score calculated by Rules 1 to 6

tot_Negative Score = Score calculated by Rules 1 to 6

It has been noted that only 8% of the opinions in the sample contained shifters. However, by applying the above compositional semantic rules, it was observed that 68% of the opinions that contained එහෙත් (but), ඒත් (but), හැබැයි (but) or නම් (if) were correctly classified by the SVM classifier (Table 7.17).

Table 7.17: Impact of flow shifters

Algorithm	Accuracy
Naïve Bayes	67.974
SVM	68.278

Thus, it can be concluded that the impact of flow shifters on classification is positive and the proposed Sinhala flow shifter compositional semantic rules gave promising results. It is also important to note that the accuracy of classification using the compositional semantics rules may have been affected by the poor representation of these shifters in the overall sample.

7.5 Morphological Approach

One of the limitations of the dictionary-based approach for sentiment classification is insufficient coverage of sentiment words. That is the emotional terms stated in the comments are not available in a sentiment lexicon with polarity scores. In morphologically rich languages, this limitation is important as inflection and derivational terms are not included in the subjective lexicon. As an example; the polarity word හොඳ (good) can have many inflectional and derivational forms such as හොඳට (adverb), හොඳක් (adverb) හොඳින් (adverb), හොඳම (adjective). The base form හොඳ (good) is an adjective, and it is significantly changed their grammatical category in the other forms coined by derivation or inflection. However, the polarity of all forms remain unchanged in the dictionary based approach. As an example the polarity of හොඳට (adverb), හොඳක් (adverb) හොඳින් (adverb), හොඳම (adjective) and හොඳ (good) are equal. However, in reality the strength of each form differs. In order to represent the different morphological forms, the ideas of inflection and derivation are important.

7.5.1 Inflection

Word formation by adding affixes to the root form of a word without changing the meaning of the base form is defined as inflection. Such word formation contributes to syntactic condition information such as case number and gender. For example, හොඳම (excellent) is an inflectional form of හොඳ (good) formed by adding the morpheme “ම”.

7.5.2 Derivation

Derivation refers to creating a new word by adding affixes to the base form. In this type of formation, the meaning and the grammatical category of most of the root forms are changed. The affixes can be added to the beginning or the end of the base form. The adverb හොඳක්

(good to be) is a derivational form of the base adjective word හොඳ (good) where the derivational morpheme ක් is added to the end. The adjective නොහොඳ (not good) is also derived from the same base form හොඳ (good) but the meaning as well as the polarity of the base form is changed. However, in this case the grammatical category of the new word does not change.

7.5.3 Generating morphological dictionary

It is understood that the lexicon considered so far in the study may not contain the full list of all the derivatives of (morphological changed) emotional words with their polarity scores. As explained above (7.5.1 and 7.5.2), it cannot be expected that all forms of inflectional as well as derivational forms are included in the list. The reason is that the dictionaries used to develop the subjective lexicon may not contain all the morphological forms for each entry. The method of generating such forms is complex, and the use natural language techniques are required to coined the words. Many researchers have used the theory of finite automata to generate all inflectional forms for a given lemma (Mlaenovic, Mitrovic, Jrstev, & Vitas, 2009). In this study, this problem is overcome by using with a previously generated sentiment lexicon and using a distinct word list extracted from a text corpus representing 10 million words (Language Technology Research Laboratory, 2011). The distinct word list comprises more than 400k distinct entries which were extracted from a 10-million-word corpus. Thus, it is reasonable to assume that the distinct word list generated contains most of the inflectional forms.

A new sentiment lexicon was constructed using the following novel algorithm including the base forms and all morphologically derived forms with polarity scores. The sentiment lexicon of lemmas (base) was compiled using a dictionary of base forms of comprised of four main lexical categories; nouns, verbs, adjectives, and adverbs. This lexicon contained all the lemmas with associated sentiment scores which are the same as detailed in the sentiment lexicon constructed in chapter 5. Once the lexicon was established, then all possible inflectional forms for all of the base forms are extracted from the distinct word list. The same sentiment score of the base form is assigned to the corresponding inflectional derivatives. This novel sentiment lexicon expanding process is depicted in the following algorithm.

Algorithm: Expansion of a sentiment lexicon for inflectional forms

```
Input:  DistinctWordList, SentimentLexicon
Output: SentimentLexiconInflectional

For each Lemma in the SentimentLexicon
        Inflections = findInflectionsInDistinctWordList(Lemma)
For each Inflection in Inflections
        AddToSentimentLexicon(lemma, infection, sentimentScore)
return SentimentLexiconInflectional
```

The algorithm returned an expanded list of 45,225 unique sentiments with their polarity scores which are approximately a 12 times increase of the original sentiment lexicon. However, there were some lexicon items that occurred with typographical errors in the distinct list. For example, the emotional word "ඉතාඑ" (good) inflected to 77 words where some were generated due to typographical mistakes. Such instances were removed from the list. It was also noted that some colloquial and spoken words were now included in the newly generated lexicon. This novel corpus based approach is an efficient technique as it can capture the majority of inflected words that are difficult to generate using linguistic rules.

Sentiment classification was carried out using Naïve Bayes and SVM to investigate the impact of including the morphological impact of sentiment terms.

Table 7.18: Classification accuracy using inflection and the expanded lexicon

Algorithm	Accuracy
Naïve Bayes	74.1429
SVM	72.857

The improvement observed using the expanded sentiment lexicon by adding inflected terms to the list is significant (Table 7.18). These results prove the importance of considering morphological features in Sinhala to achieve more accurate sentiment classification results.

7.5.4 An examination of misclassification cases

Even though an improvement is shown by taking into account morphology in classification, it is essential to analyze the misclassification instances in order to further improve accuracy.

Initially, the coverage of the lexicon constructed using the classification vector where the features that were not available were investigated. It was observed that the sentiment words were not included in 11% of the comments. Further investigations revealed that the complex

morphologically inflected words were available in some of these comments. For example, the negative emotional word "නොලැබුනොත්" (if not received) was found in the comments but, was not found in the sentiment lexicon. However, the base form of the word "ලැබුනා" (received) and its base negation "නොලැබුනා" (not received) are available in the list generated using all base words and the novel algorithm developed in this research. The rule(s) to generate such complex words is difficult to implement and therefore, not identified by the proposed algorithm. In addition, the absence of any sentiment word(s) in a comment was observed to be high in the sample. It was also observed that there was a large amount of spoken emotional words in the comments. The word "නාවට" (not coming) is a colloquial spoken form of the expression "not coming". Another coverage effect noticed in the sample was that the comments were too complex and the polarity was found to be dependent on the context. For these comments a human can easily code them as either positive or negative. However, training a model to classify these comments is a rather difficult task. For example, consider the following opinion:

"දේවින්දට ගිනි තියන්න වෙන්නේ, වාහන ගෙන්වන්නේ ඉල්ලුම හා සැපයුම අනුවයි"
(vehicles are imported based on supply and demand, Devidnda has to fire (the vehicle))

This negative comment is complex to classify by just considering emotional terms. The word "ගිනි" (fire) can be considered to be neutral in general but it can be negative in a different context.

Misclassified comments on available words in the sentence was investigated. Two important observations were made through this investigation. The first one is that the net sentiment score is zero in some comments where the number of positive and negative words were same. On the other hand, in some cases, the total positive score was equal to the total negative score even though the number of positive and negative words were different. Secondly, there were some complex sentences in the sample, where more positive (or negative) sentiment words appeared than the negatives (or positive) ones even though the comments were positive (or negative). In the following sentence, its feature vector represents two positives and one negative emotional word.

"අනෙක් අයටත් පාඩමක් ඉගෙන ගන්න හොඳ දඩුවමක් දෙන්න ඕනෑ" (Should give a strong punishment so it will be a lesson for others)

However, in this example, the word “හොඳ” (good) functioned as an intensifier to increase the polarity of the negative word “දඩුවමක්” (punishment). Generally, the word “හොඳ” (good) acts as an adjective. In this sentence, identifying the positive word “හොඳ” (good) as an intensifier is complex and such contexts typically only happen in the spoken form of sentences not the written form.

7.6 Chapter Summary

The chapter investigated the impact of Sinhala linguistic features in sentiment classification. The findings of the chapter contribute to sentiment classification for any morphologically rich language by considering and applying linguistic features to sentiment classification. To the researcher’s best of knowledge this study is the first case study that extensively investigates of linguistic features for a morphologically rich language.

Initial experiments revealed that among the relevant POS units’ the adverb was the highest influencing linguistic item for Sinhala sentiment analysis. Descriptive adjectives were frequent in Sinhala opinions. However, they made no significant contribution to the classification. Additionally, there was no significant advantage in using both categories together even though other languages have been shown to benefit from combining them (Hu & Hatzivassiloglou, 2003).

The impact of negators was also investigated by introducing artificial features and new rules for the classification model, and it was found that the negators were essential for the analysis. Furthermore, it was also noted that no significant improvement to the classification model was shown when combining base negators with parts of speech (adjectives and adverbs). However, linguistic features with numeric polarity assignment for each feature including base negators gave promising results. Base negators that immediately flips the negative polarity of the adjacent word to a positive polarity were identified as a dominating feature for positive comments.

The analysis of the feature was also carried out using the polarity score extracted from the subjective lexicon. By incorporating proposed hand-written rules, it was observed that contextual intensifiers play a significant role in polarity determination in Sinhala opinion classification. The impact of flow shifters examined by incorporating compositional semantics rules showed that the features highly influenced the classification. The influence of morphologically inflected lexical items on sentiment analysis for Sinhala was investigated by expanding the subjective lexicon, which was built in the study. Even though lexicon based

techniques were used to include the morphologically inflected sentiment words in the classification, there were complex derivatives that cannot be identified by the proposed model.

Chapter 8: Discussion, Recommendation, and Conclusion

8.1 Introduction

This research set out to build a framework for sentiment analysis for morphologically rich languages. The investigation was carried out in the Sinhala language, one of the morphologically rich languages in South Asia. To the best of the researcher's knowledge, this is the first attempt at building lexical resources for sentiment analysis and applying sentiment classification methods for Sinhala. The findings of study contribute to the field of sentiment analysis for other morphologically rich languages and natural language processing tasks. The results achieved demonstrated the feasibility of sentiment analysis in morphologically rich languages.

The research mainly focuses on key two research themes;

1. Automatically generating lexical resources using already existing dictionaries for morphologically rich languages.

2. Adaptation of Bayesian algorithms for sentiment classification.

The experiments were carried out using the opinions written for news articles, which were extracted from the online newspaper as a case study. The impact of linguistic features on sentiment classification was extensively investigated.

This chapter summarizes the research achievements, draws some conclusions, discusses the limitations of the research and explores possible avenues for future work.

8.2 Thesis contributions

8.2.1 Research Theme 1

Within the main objective of building a framework for sentiment analysis for morphologically rich languages, the first research question was formulated to focus on building a sentiment lexicon for morphologically rich languages. The research question addressed the issues of applying contemporary lexicon building techniques and methods of evaluating them.

(a) How can effective and efficient lexical resources be automatically generated?

In this direction, the thesis proposed three novel methods for sentiments lexicon generation for morphologically rich languages. The methods presented in the thesis exploited appropriate language resources that are freely available for English. However, electronic dictionaries are available for Sinhala and most of other languages. Because Sinhala and many non-English languages do not have sufficient electronic resources, it was necessary as part of this research to develop techniques for automatically and efficiently generating language resources suitable for data mining and sentiment analysis.

The first method dubbed the cross language approach, combined an existing sentiment lexicon for English and a bilingual dictionary for Sinhala to construct the sentiment lexicon for Sinhala. The outcome was a list of adjectives and adverbs with their polarity scores extracted from the English sentiment lexicon. The generated lexicon is the first ever sentiment lexicon for the Sinhala language. Moreover, the generated lexicon can be used as the baseline repository for any future sentiment classification study using dictionary-based methods. Additionally, the classification accuracies achieved can be considered as the baseline sentiment classification accuracy for Sinhala. To the best of the researcher's knowledge, this is the first attempt at combining foreign language resources to construct a sentiment lexicon for a morphologically rich language.

In the second method, the thesis generated a sentiment lexicon with positive, negative words. The objective of the experiment was to build a lexicon considering the linguistic features of the languages. Unlike the first approach, the method proposed is independent of the external sources such as lexicon for another language. The study identified 17 affixes in Sinhala that can use to develop the positive/negative list. The identified affixes naturally reverse the polarity or the sentiment of a word when combining with another possible word. The complexity of the affixes that generate polarity bearing words was analysed extensively and exploited in the evaluation process.

(b) Can contemporary lexicon building techniques be adapted to morphologically rich languages?

The third method of building resources, the thesis tested the suitability of contemporary graph based methods for building the sentiment lexicon. By explicitly capturing relationships between entries of the dictionary the coverage of the generated lexicon by graph based approach proved to be much higher than with the lexicons built with the two previous

approaches. The proposed approach is the first ever use of a graph based sentiment lexicon construction for a morphologically rich language.

(c) How can newly generated lexical resources in sentiment classification be evaluated?

The success of the lexicons was mainly evaluated by classifying written opinions on online news articles. The promising results (60%) by this first ever attempt demonstrated the feasibility of generating sentiment lexicons by cross-linguistic approaches.

The appropriateness of the composed positive, negative words lexicon was tested using expert's (linguist) knowledge that available in literature such as books and supervised classification methods as well. It was revealed that the proposed method is a feasible solution for generating such a list for any morphological rich language.

By evaluating graph based lexicon using supervised classification approach and linguistic expert's knowledge, observed that negative words correctly generated by the graph assigning correct polarity scores.

8.2.2 Research Theme 2

The second research question was set with the objective of evaluating the Bayesian sentiment classification for morphologically rich languages. The impact of language features especially; the morphological variations were investigated to answer this question.

(a) How can Bayesian sentiment classification algorithms for morphologically rich languages be evaluated?

The significant accomplishment of the research is investigating the Bayesian sentiment classification for Sinhala. The feasibility of applying classical text mining techniques for a morphologically rich language, Sinhala tested successfully in this research achieving significant results (83%) by applying two classification algorithms such as Naïve Bayes and SVM. These algorithms investigated on the standard sentiment classification approaches such as document level and domain specific analysis. The experiment was initiated by investigating statistical approaches to feature selection that relied on word frequency. Then effects of using n-gram features were investigated next. The tfidf weighting scheme for unigram features with SVM achieved better performance than with Naïve Bayes. However, a binary weighting scheme for both unigram and bigram for Naïve Bayes would give higher classification accuracies than SVM. It was found that correlation based feature selection methods are more

applicable for Sinhala sentiment classification. Domain specific classification outperformed the domain independent accuracies as in the case of other languages such as English. In next step, the thesis investigated the linguistic features in classification.

(b) How can word level morphological features be applied to Bayesian sentiment classification in the context of morphologically rich languages?

One of the main contributions of this thesis is identifying the language specific features for Sinhala that have the potential to improve sentiment classification. To researchers interested in investigating linguistic features this work represents the first attempt, in any language morphologically rich or not, at using such linguistic feature such as rules generated for discourse analysis using pre and post positions for shifter features. Of the linguistic features examined, adverbs were shown to make a vital contribution to classifying Sinhala opinions into negative or positive classes. As with other languages, the impact of negative words was discovered to be highly influential in classification. The research introduced new terminology for handling negators such as base negators which are purely negative. Negators were tagged with new symbols; POS_NOT, NEG_NOT, KMK_NOT which indicate the context of the negator's usage. A special negation word combination; කමක් නෑ (Its OK, dubbed KMK_NOT) was identified in the analysis and it was noted that this combination does not exist in many other languages including English.

This research paid special attention to semantic discourse analysis in sentiment classification for Sinhala. In this analysis, the semantic discourse of contextual intensifiers and flow shifters were investigated extensively using manually written rules. It was found that, in addition to the morphological features, contextual intensifiers and flow shifters contributed highly to the determination of polarity of Sinhala comments. Further experiments considered morphologically inflected/derived word that was added to the Sinhala lexicon. It was discovered that the inclusion of morphologically inflected words in the sentiment lexicon not only improved classification accuracies but also increased the coverage of the lexicon constructed.

8.3 Limitations of the Research

In this section, the discussion of the limitations of this research is categorized into themes based primarily on the research questions. These themes include the limitations or constraints of:

- methods used to construct the lexicons,

- features used in Bayesian classification,
- adapting linguistic information for sentiment classification.

The initial lexicon constructed using a cross-linguistic approach may not be the optimal lexical sentiment resource for Sinhala. Primarily, the resource compiled by the proposed method is based on several assumptions. These assumptions include equality of the sense and equality of POS of a word in both languages. It was also assumed that the sense of the bridging word is same for both languages. It is clear that this assumption is not valid in some cases. As an example, the word "acceptable" was mapped to the Sinhala word "ප්‍රිය" (nice) with positive sentiment score of 0.625. However, the sense of the word "ප්‍රිය" (nice) is not exactly same as "acceptable" depending on the context. Additionally, the POS for the languages was assumed as equal and this is a possible limitation of the study as there is a chance of differing lexical categories in two languages. The sentiment word "අවසානය" (cessation) is noun in Sinhala but the translation or the mapping to the English adjective "close". Although the POS of the two words are different, they have close meaning and it can be generated an adjective by dropping the morpheme "ය" of the inflected form "අවසානය" (cessation). The sentiment scores borrowed from English for use in Sinhala is one of the drawback of the proposed method as the score was calculated solely using the linguistic rules for English.

The second method proposed for sentiment lexicon construction was based purely on the affixes (morphemes) of the Sinhala language. The insufficient coverage of affixes can be a limitation for generating positive/negative words. Even though the affixes were listed with the consultation of an expert and experienced linguist, the number of affixes can change as always languages are evolving. Hence listing the complete list of affixes is a limitation of the proposed method, as the list of affixes will eventually become out of date. Unexpectedly, it was found that adding affixes did not always generate a negative word. There are cases, as explained in Chapter 5; where adding affixes will generate positive words. These situations need to be investigated by manual inspection and it is challenging to develop rules. Another limitation of this approach is that emotional terms (positive or negative) that do not follow the general rule of generating a negative sentiment by adding affixes may not be automatically included in the lexicon and need to be added manually by an expert. For example, the word "නරක, nɑːrʌkʌ" (bad) is a pure negative word that cannot be generated by affixes and was not included in the lexicon and therefore had to be added manually by the researcher in consultation with an expert Sinhala linguist.

The use of the single dictionary to compile the graph based lexicon can also be seen as a limitation of the graph based lexicon approach. Hammer et al. (2014) shown that a good coverage lexicon can be developed using different dictionaries and label propagation. In addition, the proposed method used the only relationship between synonyms to construct the graph. Other relationships such as antonyms, hyperonymy and meronyms were not accounted for the lexical generation is a limitation of the proposed method. Furthermore, listing the words using the shortest path algorithm is not the optimal solution for retrieving all of the possible sentiment words for inclusion in the lexicon. On the other hand, consideration of all possible paths would increase the search space drastically, thus increasing the computational complexity of lexicon construction. Moreover, the inability to consider the context of the terms in this approach results in adding irrelevant words to the path by the proposed method.

In contemporary text classification techniques, features are only the highest frequent words for the classification vector and they include mostly the emotional terms of expressions. This inclusion is a limitation in classifying neutral opinions that contain non-emotional terms. Also of note is that the spelling variation of terms might have affected the classification accuracy. In Sinhala, the spelling variation is significant for standard and casual writing such as online opinions. As an example, the sentiment term හොඳ (good) can be written in two forms in Sinhala. Even though හොදැ (good) is not accepted in the standard writing in Sinhala, ordinary writers are not concerned with the correct spelling, especially on social media. The frequency of these terms are less and may not be included in the highest frequency list which comprises the feature set for classification. On the other hand, if they are included, the high dimensionality of the feature vector that would increase the complexity of the classification. Additionally, these spelling differences affect to the semantic inferences when integrating the discourse to the sentiment classification. As an example, the word "ගණ" refers to "group" with neutral sentiment and "ගන" gives the meaning of "hard" assigning positive or negative sentiment in different context.

Primary in this research the Bayesian based classification used the bag-of-words model to set the feature vector for the classification. Even though the research handles the discourse, information for this model at the semantic level the bag-of-words model fails to include the discourse relations by connectives and conditions in phrase level. The phrase connectives such as "කෙසේ නමුත්" (however) requires advanced syntactic based discourse analysis.

As mentioned in chapter 7, identification of the part of speech was carried out using a part of speech tagger. This study incorporated a corpus-based approach to tag the part of speech. The tagging may not be the complete answer for detecting the lexical items such as adjectives and adverbs. Some lexical categories were not considered in this study due to the complexity of the linguistic functions involved. For example, the impact of verbs and nouns was not considered however, these lexical categories may have a significant impact on classification accuracies.

8.4 Future Works

This research presents a foundation and seminal study into sentiment analysis using linguistic features to classify sentiments expressed in a morpheme rich language, and as such, there are many potential avenues for future research the three main avenues are discussed here.

8.4.1 Expanding the sentiment lexicon

It is possible that the sentiment lexicon constructed using the cross-linguistic approach can be improved in terms of its coverage and quality by considering the gloss of both of the languages in cases where the context of two words match. The gloss provides the relevant meaning of the word that both languages agree on in different discourses. For examples, the discourse of the term “නියත” (definite) in “නියත ජයග්රහණයක්” (definite win) and “නියත දිශාවක්’ (definite direction) are different. The first case tends to be positive, the second phrase is actually neutral, and the gloss explains the degree to which phrase one or two are positive. By assigning different scores for the word “නියත” in these situations considering the gloss, the complexity of the discourse can be solved successfully at lexicon level.

In addition to the synonym relationship used in the graph based sentiment construction, this research led to the suggestion of the use of other relations that are available in WordNet such as antonyms, hyperonymy, and meronyms. Thus, a translation model could be utilized to translate a word from a morphologically rich language to English. This translation should then help to identify a relevant synset (Miller, 1995) word. Using the extracted synset and the WordNet for English the lexical relations for the target language may be established.

The thesis also proposed refining the general purpose sentiment lexicon constructed in the study by corpus based approach. The opinion corpus used in this study could be enriched by adding opinions from other domains such as movie, book and product reviews collected from social media. The proposed Bayesian classification algorithms could then classify these newly added opinions. Then the keywords included in the positive/negative comments initially treated

as positive/negative words, and the relevant sentiment score could be calculated using word co-occurrence statistics such as pointwise mutual information (Chaudhari, Damani, & Laxman, 2011). This pointwise mutual information measures the association between two terms. The association between new term from the corpus and the known words form available lexicon can be calculated using pointwise mutual information. Then the sentiment score of the highest associated word will be the sentiment score of the new term.

8.4.2 Further enhancement of Bayesian classification using linguistic structures

The proposed Bayesian classification methodologies can be tested for other application areas, other than online opinions of new paper articles, such as opinions from product reviews, film reviews, etc. The evaluation of the effectiveness of the constructed lexicons in these application domains is one possible avenue of future work.

The discourse and the context knowledge of the reviews greatly contributed the polarity determination. Incorporating context information as features in Bayesian classification methods is worth exploring. The polarity of a word is usually most often bound to the context of location (including geo-location), time and author information (Vosoughi, Zhou, & Roy, 2015). The phrase "විශාල කාලයක්" (huge time) express the context of time in a negative sentiment. While, "විශාල ප්‍රදේශයක්" (huge region) refers to the context of location and it may be negative or positive based on the other phrases in the sentence. Bayesian based classification methods performance should be improved by using a bag-of-word model that includes contextual information related to space and time.

The discourse analysis of opinions could be further investigated at a syntactic level for morphologically rich languages. Discourse or dependency parsers are used to identify the discourse relations within an opinion (Mukherjee & Bhattacharyya, 2012). However, these parses show good performances in structured texts than the unstructured texts so applying parses for opinions is a challenging task especially for morphologically rich languages. Alternatively, a discourse annotated corpus can be used to identify the discourse at the sentence level. Such a corpus can be used to, "model the discourse structures as predicate-argument identifications where predicates are discourse connectives" (Mukherjee & Bhattacharyya, 2012, p. 1849). These corpora are tree based structures for which Penn Discourse Treebank (PDTB) and Rhetorical Structure Theory (RST) Discourse Treebank are available resources that may be adopted for use with morphologically rich languages. For example, it is expected that coherence relation such as contrast relation expressed by Sinhala "අනෙක් අතට" or "එසේ

වුවත්” (by contrast) can be identified by these tree structures. This idea for furture work is support by Prasad et al. (2008) who suggested that corpora for morphologically rich languages could be developed by adopting lexically grounded approach using PDTB or RST.

8.4.3 Classifying opinions into different levels

The sentiment level of an opinion can be scaled for a given range by considering the scale of emotional words. Investigating the impact of the scale of emotional words in sentiment analysis is proposing as future work. Some words are inflected with morphemes and generate word intensifying the valence of the word. The word “හොඳ” (good) inflected to “හොඳම” (very good) by intensifying the sentiment of the noun affected by the emotional word “හොඳ” (good). The opinions include these forms can be scaled in a range based on the emotional word intensifying.

This research succeeded in classifying Sinhala opinions to a high level of accuracy, but it is clear that future works could lead to further improvements. The finding of the linguistic features to enhance sentiment classification of morphologically rich languages has proven to promising and it is hoped that this research will spur further research particularly in the key areas identified above.

References:

Agarwal, B., Mittal, N., Bansal, P., & Garg, s. (2015). Sentiment Analysis Using Common-Sense and Context Information. *Computational Intelligence and Neuroscience*.

Al-Rfou, R., Perozzi, B., & Skiena, R. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning CoNLL'2013* (pp. 183–192). Sofia, Bulgaria: Association for Computational Linguistics.

Arora, S., Mayfield, E., Ros, C. P., & Nyberg, E. (2010). Sentiment classification using automatically extracted subgraph features. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 131-139). Stroudsburg, PA, USA: Association for Computational Linguistics.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary & Linguistics Computing*, 1-16.

Badaro, G., Baly, R., & Hajj, H. (2014). A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining. P*roceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)* (pp. 165–173). Doha, Qatar: Association for Computational Linguistics.

Bakliwal, A., Arora, P. & Vrma, V. (2012). Hindi Subjective Lexicon: A lexical Resource for Hindi Polarity Classification. *The eighth international conference on Language Resources and Evaluation (LREC).* Hyderabad.

Balamurali, A. R., Joshi, A., & Bhattacharyya, P. (2011). Harnessing WordNet senses for supervised sentiment classification. *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural* (pp. 1081-1091). Stroudsburg, PA, USA: Association for Computational Linguistics.

Benamara, F., Cesarano, C., & Reforgiato, D. (2007). Sentiment Analysis: Adjectives and Adverbs are better that Adjectives Alone. *International Conference on Weblogs and Social Media (ICWSM).* Boulder, Colorado.

Benerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *IJCAI*, (pp. 805–810).

Bespalov, D., Bai, B., & Qi, Y. (2011). Sentiment Classification Based on Supervised Latent n-gram Analysis. *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 375-382). ACM.

Bharti, K. K., & Singh, P. (2014). A Three Stage unsupervised dimension reduction method for text clustering. *Journal of Computer Science*, 156-169.

Bhowmick, P., Mitra, P., & Basu, A. (2008). An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text. Proceedings of the workshop on Human Judgements in Computational Linguistics (pp. 58 - 65). Manchester: Coling.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Chandrasekaran, R. M., & Vinodhini, G. (2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 282-292.

Chaudhari, D. L., Damani, O. P., & Laxman, S. (2011). Lexical Co-occurrence, Statistical Significance, and Word Association. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (pp. 1058–1068). Edinburgh, Scotland, UK.

Chen, Y., & Skina, S. (2014). Building Sentiment Lexicons for All Major Languages. *52nd Annual Meeting of the Association for Computational Linguistics* (pp. 383-389). Baltimore, Maryland: Association for Computational Linguistics.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). Using verbs and adjectives to automatically classify blog sentiment. *AAAI Spring Symposium.*

Chetviorkin, I., & Loukachevitch, N. (2012). Extraction of Russian Sentiment Lexicon-Domain. *Proceedings of COLING 2012*, (pp. 593–610). Mumbai.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 22-29.

Covell, M., & Baluja, S. (2013). Efficient and Accurate Label Propagation on Dynamic Graphs and Label Sets. *International Journal on Advances in Networks and Services*, 246-259.

D'Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications*, 26-33.

Das, S. R., & Chen, M. Y. (2001). Yahoo! For Amazon: Sentiment Parsing from Small Talk on the Web. *EFA 2001 Barcelona Meetings.* Barcelona.

Dave, k., Lawrence, S., & Pennock, D. K. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *In Proceedings of 12th International Conference on World Wide Web,* (pp. 519-528). Hungary.

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentiment Analysis. *Proceedings of the 2008 IEEE 24th international conference on data engineering workshop*, (pp. 507-512). Cancun.

Deng, Z. H., Luo, K. H., & Yu, L. H. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert System with Applications*, 3506-3513.

Dictionaries Translation and Language Resources. (2000). Retrieved July 12, 2014, from Online bilingual and multilingual dictionaries: http://www.lexicool.com/#dictionary-search

Dileep, C. (2010). *Sinhala.* John Benjamins Publishing.

Dissanayake, J. B. (2014). *Pada Nirmanaya (පද නිර්මාණය).* Colombo: Sumitha Prakashakyo.

Duwairi, R., & Orfali, M. E. (2013). A Study of the Effects of Preprocessing Strategies on Sentiment Analysis for Arabic text. *Journal of Information Science*, 1-14.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *In Proceedings of the 5th Conference on Language Resources and Evaluation*, (pp. 417 - 422). Genoa - Italy.

Esuli, A; Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Managemen* (pp. 617- 624). ACM.

Ethnologue. (2016). *Statistical Summaries: Summary by world area*. Retrieved 04 05, 2014, from Ethnologue: Languages of the World: http://www.ethnologue.com/statistics

Fernandez, J., Guti´errez, Y., G´omez, J. M., & Martinez-Barco, P. (2014). GPLSI: Supervised Sentiment Analysis in Twitter using Skipgrams. *Proceedings of the 8th International Workshop on Semantic Evaluation*, (pp. 294–299). Dublin, Ireland.

Fisher T. (2009). ROI in social media: A look at the arguments. *Journal of Database Marketing & Customer Strategy Management*, 189-195.

Franky, B. O., & Veselovská, k. (2015). Resources for Indonesian Sentiment Analysis. *The Prague Bulletin of Mathematical Linguistics*, 21- 41.

Freeworldmap.net. (n.d.). Retrieved from Where is Sri Lanka Located on the World Map?: http://www.freeworldmaps.net/asia/srilanka/location.html

Fu, G., & Wang, X. (2010). Chinese Sentence–Level Sentiment Classification Based on Fuzzy Sets. *The 23rd International Conference on Computational Linguistics (COLING 2010)*, (pp. 312-319). Beijing.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. *COLING '04 Proceedings of the 20th International conference on Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Gavilanes, M., Lopez, T. L., Martınez, J. J., Montenegro, E. C., & F.J.G, C. (2014). GTI: An Unsupervised Approach for Sentiment Analysis in Twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, (pp. 533–538). Denver, Colorado.

Ghwanmeh, S. H. (1998). Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language. *International Journal of Information Technology*, 168-172.

Hall, A. M. (1999). *Correlation-based Feature Selection for Machine Learning.* Retrieved from Dept of Computer Science, University of Waikato.: http://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Hall, M. A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 10-18.

Hammer, H., Bai, A., Yazidi, A. & Engelstad, P. (2014). Building sentiment Lexicons applying graph theory on information from three Norwegian thesauruses. *Norsk Informatikkonferanse (NIK)*.

Han, E., Karypis, G., & Kumar, V. (2001). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. *Advances in Knowledge Discovery and Data Mining*, 53-65.

Han, J., Kamber, M., & Pei, J. (2012). *Dat Mining Concepts and Techniques.* MA: Morgan Kaufmann.

Hatzivassiloglou, V., & McKeown, K. P. (1997). Predicting the Semantic Orientation of Adjectives. *EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174 -181). ACL.

Hidayatullah, A. F. (2015). The Influence of Stemming on Indonesian Tweet Sentiment Analysis. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015).* Palembang, Indonesia.

Hilpert, M. (2006). Auxiliaries in spoken Sinhala. *Functions of Language*, 229–253.

Hu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *EMNLP '03 Proceedings of the 2003 conference on Empirical methods in natural language processing*, (pp. 129-136).

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168 - 177). ACM.

Huang, S., Niu, Z., & C, S. (2014). Automatic Construction of Domain-specific sentiment lexicon based on constrained label propagation. *Knowledge0Based systems*, 191-200.

Jain, D., & Cardona, G. (2014). *The Indo-Aryan Languages.* New York: Routledge.

James, W. G., & Lust, B. (1998). *Studies in South Asian Linguistics: Sinhala and Other South Asian Languages.* New York: Oxford University Press.

Jang, H., & Shin, H. (2011). Effective Use of Linguistic Features for Sentiment Analysis of Korean. *24th Pacific Asia Conference on Language, Information and Computation*, (pp. 173-182).

Jia, L., Yu, C., & Meng, W. (2009). The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1827-1830). ACM.

Jindal, N., & Liu, B. (2006). Identifying comparative sentences in text documents. *29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 244 - 251). ACM.

Joshi, A., Balamurali, A. R., & Bhattacharyya, P. (2010). A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. *International Conference on Natural language Processing (ICON).* Kaha.

Jovanoski, D., & Pachovski, V. (2015). Sentiment Analysis in Twitter for Macedonian. *Proceedings of Recent Advances in Natural Language Processing*, (pp. 249–257). Hissar, Bulgaria.

Kaji, N., & Kitsuregawa, M. (2007). Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational* (pp. 1075–1083). Prague: Association for Computational Linguis.

Kampas, J., Marx, M., Mokken, R. J., & Rijke, M. D. (2004). Using WordNet to measure semantic orientations of adjectives. *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, (pp. 1115–1118).

Kariyawasam, D. S. (2013). *Sinhala Basha Sampradaya (සිංහල භාෂා සම්ප්‍රදාය).* Maharagama, Sri Lanka: Media House.

Kasthuriarachchy, B. H. (2012). A Review of Domain Adaption for Opinion detection and Sentiment Classification. *The International Conference on Advances in ICT for Emerging Regions*, (pp. 209 - 213). Colombo.

Kennedy, A., & Inkpen, D. (2005). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence, 22*, 110-125.

Khawaldeh, F. T. (2015). A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic. *World of Computer Science and Information Technology Journal*, 124-128.

Khoo, H. C., & Chan, S. (2003). Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction. *14th ASIS SIG/CR Classification Research Workshop*, (pp. 42-48). California.

Kim, S. M., & Hovy, E. (2004). Determining the Sentiment of Opinions. *COLING '04 Proceedings of the 20th International conference on Computational Linguistics.* Geneva: Association for Computational Linguistics.

Kim, S., & Hovy, E. (2006). Automatic identification of pro and con reasons in online reviews. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 483-490). Sydney: Association for Computational Linguistics.

Kobayasi, N., Inui, K., & Matsumoto, Y. (2007). Opinion Mining from web documents: Extraction and structurization. *Transaction of the Japanese Society for Artificial Intelligence*, (pp. 227-238).

Kudo, T. & Matsumoto, Y. (2004). A Boosting Algorithm for Classification of Semi-Structured Text. *Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing.*

Kwok, J. (1998). Automated Text Categorization Using Support Vector Machine. *International Conference on Neural Information Processing (ICONIP)*, (pp. 347-351).

Language Technology Research Laboratory. (2011). Retrieved 08 30, 2015, from Language Technology Research Laboratory: http://ucsc.cmb.ac.lk/ltrl/

Li, S., Lee, S., Chen, Y., Huang, C., & Zhou, G. (2010). Sentiment Classification and Polarity Shifting. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, (pp. 635–643). Beijing.

Liang, P., & Potts, C. (2015, January). Bringing Machine Learning and Compositional Semantics Together. *Annual Review of Linguistics 1(1)*, pp. 355-376.

Liao, Y., & Vemuri, V. R. (2002). Using Text Categorization Techniques for Intrusion Detection. *Proceedings of the 11th USENIX Security Symposium*, (pp. 51-59). Berkeley, CA.

Liddy E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science.* New York: Marcel Decker, Inc.

Liddy, E. (2001). *Natural Language Processing.* NY.

Liu B. (2012). *Sentiment Analysis and Opinion Mining.* Chicago.

Liu, B. (2010). Sentiment Analysis and subjectivity. In *Appear in Hand Book of Natural Language Processing.*

Liu, B., Hu, M., & Chen, J. (2005). Opinion observer analyzing and comparing opinions on the web. *In Special interest tracks and posters of the 14th* (pp. 342-351). Chiba, Japan: ACM.

Liu, F. W. (2010). Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, (pp. 309–312). Los Angeles, California.

Makki, R., Brooks, S., & Milios, E. E. (2014). Context-Specific Sentiment Lexicon Expansion via Minimal User Interaction. *Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP)*, (pp. 178-186).

Manning, C., & Schütze, H. (1999). Text Categorization. In *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Martineau, T., & Finin, J. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. *Proceedings*, (pp. 258-261).

Maynard D. & Funk A. (2011). Automatic detection of political opinions in Tweets. In F. D. Castro F., *The Semantic Web: ESWC 2011 Workshops* (pp. 88 - 89). Springer.

Maynard D., Bontcheva. K., & Rout D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of @NLP can u tag #usergeneratedcontent? Workshop at LREC.* Istanbul, Turkey.

McCarthy, J. (1959). Programs with Common Sense. *Proceedings of the Symposium of the National Physics Laboratory* (pp. 77- 84). London, U.K: Her Majesty's Stationery Office. Reprinted in McC90.

McCarthy, J. (1989). Artificial Intelligence, Logic and Formalizing Common Sense. In R. Thomason, *Philosophical Logic and Artificial Intelligence.* Klüver Academic.

Medagoda, N., & Weerasinghe, R. (2011). An application of Document clustering for Categorizing open-ended Survey Responses., (pp. 102-110). Colombo.

Medagoda, N., Shanmuganathan, S., & Whalley, J. (2013). A Comparative Analysis of Opinion Mining and Sentiment Classification in non-English Languages. *14th International Conference on Advances in ICT for Emerging Regions December.* Colombo.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A Survey. *Ain Shams Engineering Journal*, 1093–1113.

Miller, A. G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 39-41.

Mittal, M., Agarwal, B., Chouhan, G., Bania, N., & Pareek, P. (2013). Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. *International Joint Conference on Natural Language Processing*, (pp. 45-50). Nagoya, Japan.

Mittal, N., Agarwall, B., Chouhan, G., Bania, N., & Pareek, P. (2013). Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. *International Joint Conference on Natural Language Processing*, (pp. 45–50). Nagoya, Japan.

Mlaenovic, M., Mitrovic, J., Jrstev, C., & Vitas, D. (2009). An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds. *In Proceedings of Lexis and Grammar Conference*, (pp. 204-212). Bergen.

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating High-Coverage Semantic Orientation Lexicons From Overtly MarkedWords and a Thesaurus. *Conference on Empirical Methods in Natural Language Processing.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Mukherjee, S., & Bhattacharyya, P. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. *Proceedings of COLING 2012, 24th International Conference on Computational Linguistics:Technical Papers*, (pp. 1847-1864). Mumbai, India.

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *In Proceedings of Conference on Empirical Methods in Natural Language Processing* (pp. 412--418). Barcelona, Spain: Association for Computational Linguistics.

Murphy, K. P. (2006). *Naive Bayes classifiers.* University of British Columbia.

Na, J., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. *Proceedings of the Eighth International ISKO Conference*, (pp. 49-54). Wurzburg, Germany.

Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. *n Intelligent Data Engineering and Automated Learning–IDEAL*, 194-201.

Ng, V., Dasgupta, S. & Arifin, S. M. N. (2006). Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 611–618). Sydney: Association for Computational Linguistics.

Ng, V., Dasgupta, S., & Arifin, S. M. (2006). Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 611–618). Sydney: Association for Computational Linguistics.

Niepert, C., Stuckenschmidt, H., & Strube, M. (2011). Fined-Grained Sentiment Analysis with Structural Features. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, (pp. 336-344). Chiang Mai, Thailand.

O'Connor B., B. R. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Fourth International AAAI Conference on Weblogs and Social Media*, (pp. 122 - 129). Washinton DC.

Ohana, B., & Brendan, B. (2009). Sentiment classification of reviews using SentiwordNet. *9th. IT & T Conference.* Dublin.

P´erez-Rosas, V., B. C., & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. *Eighth International Conference on Language Resources and Evaluation*, (pp. 3077-3081). Istanbul, Turkey.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh conference International Language Resources and Evaluation (LREC'10)*, (pp. 1320 -1326).

Pak, P., & Paroubek, P. (2011). Language Independent Approach to Sentiment Analysis. *LIMSI Participation ROMIP.* France.

Paltoglou, G., & Thelwall, M. (2010). A study of Information Retrieval weighting schemes for sentiment analysis. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1386–1395). Uppsala, Sweden: Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP*, (pp. 79 - 86).

Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. I*n Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271).* Association for Computational Linguistics.

Pennebaker, J. W., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and Psychometric Properties of LIWC2015.* Austin: TX: University of Texas at Austin, DOI:10.15781/T29G6Z.

Peters, C., & Picchi, E. (2014, 10 12). *Across Languages, Across Cultures*. Retrieved from http://www.dlib.org/dlib/may97/peters/05peters.html

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 1112-1130.

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, 1-10.

Politics and History of the Indian Subcontinent. (2014, August 18). Retrieved from South Asia Blog: https://southasiablog.wordpress.com/2014/08/18/south-asias-lebanon-the-demographics-of-sri-lanka/

Porter, M. (1997). An algorithm for suffix stripping. *Readings in information retrieval*, 313-316.

Potts, C. (2011). *Sentiment Symposium Tutorial: Lexicons*. Retrieved from Sentiment Symposium Tutorial: http://sentiment.christopherpotts.net/lexicons.html

Prasad, R., Husain, S., Sharma, M., & Joshi, A. (2008). Towards an Annotated Corpus of Discourse Relations in Hindi . *The 6th Workshop on Asian Language Resources*, (pp. 73 - 80). Hyderabad, India.

Priyadarsini, R. P., Valarmathi, V. L., & Sivakumari, S. (2011). Gain ratio based feature selection method for privacy preservation. *ICTACT Journal on Soft computing*, 201-2005.

Qu, L., Ifrim, G., & Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 913-921). Association for Computational Linguistics.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, 209–228.

Raaijmakers, S., & Kraaij, W. (2008). A Shallow Approach to Subjectivity Classification. *In Proceedings of ICWSM-2008*, (pp. 216–217).

Rao, D., & Ravichandran, D. (2009). Semi-Supervised Polarity Lexicon Induction. *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 675-682). Association for Computational Linguistics.

Rushdi, S. M., Martín, V. M., Reña, L. A., & Perea, O. J. (2011). Opinion Corpus for Arabic. Journal of the American Society for Information Science and Technology.

Sashi C. M. (2012). Customer engagement,buyer-seller relationships, and social media. *Management Decision*, 253 - 272.

Schouten, K., & Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge & Data Engineering*, 813-830.

Schulz, J. M., Hacker, C. W., & Mandl, T. (2010). Multilingual Corpus Development for Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 3409-3412). European Language Resources Association.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 1-47.

Severyn, A., & Moschitti, A. (2015). On the Automatic Learning of Sentiment Lexicons. *The 2015 Annual Conference of the North American Chapter of the ACL* (pp. 1397–1402). Denver, Colorado: Association for Computational Linguistics.

Sharma, R., Nigam, S., & Jain, R. (2014). Opinion Mining In Hindi Language: A Survey. *International Journal in Foundations of Computer Science & Technology (IJFCST)*, 41- 47.

Spence R. and Smith M. L. (2010). ICT, Development, and Poverty Reduction: Five Emerging Stories. *Information Technologies and International Development, 6*, 11-17.

Stoyanov, V., & Cardie, C. (2008). Annotating Topics of Opinions. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, (pp. 28-30). Morocco.

Sun, A., & Lim, E. (2001). Hierarchical Text Classification and Evaluation. *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference.*, (pp. 521 - 528). San Jose, CA.

Taboda, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics, 37*(June 2011), 267-307.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications*, 2622–2629.

Thelwall, M., Buckley, K., Paltoglou, G., & Cai, D. (2010). Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 2544-2558.

Tong, R. M. (2001). An Operational System for Detecting and Tracking Opinions in On-line Discussions. *Proceedings of the Workshop on Operational Text Classification (OTC).*

Tsarfaty, T., Seddah, D., Goldberg, Y., Kubler, S., Candito, M., Foster, J., Tounsi, L. (2010). Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages* (pp. 1-12). Los Angeles, California: Association for Computational Linguistics.

Tsoumakas, G., & Katakis, I. (2008). Multi-Label Classification; An Overview. In J. Wang, *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 64-74). New York.

Turner V., G. j. (2014, Apri). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Framingham, MA, USA.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (pp. 417-424). Philadelphia.

Umera-Okeke, N. (2007). Exploring Affixation in English. *African Research Review, 1*(3), 9-35.

Vanzo, A., Croce, D., & Basili, R. (2014). A context-based model for Sentiment Analysis in Twitter. *The 25th International Conference on Computational Linguistics (COLING 2014)*, (pp. 2345-2354). Dublin, Ireland.

Vapnik, V. (1998). *Statistical Learning Theory.* New York: John Wiley & Sons.

Velikovich, L., Goldensohn, S. B., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. *The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 777–785). Los Angeles, California, Association for Computational Linguistics.

Vikram, S. (2013). Morphology: Indian Languages and European Languages. *International Journal of Scientific and Research Publications*.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 282 - 291.

Vosoughi, S., Zhou, H., & Roy, D. (2015). Enhanced Twitter Sentiment Classification Using Contextual Information. *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*, (pp. 16-24). Lisbon, Portugal.

Wang, X., Zhao, Y., & Fu, G. (2010). A Morpheme -based Method to Chinese Sentence -Level Sentiment Classification. *International Journal on Asian Language Processing*, 95-105.

Welgama, V., Herath, D. L., L. C., Udalamatta, N., Weerasinghe, R., & T, J. (2011). Towards a Sinhala Wordnet. *Conference on Human Language Technology for Development, Alexandria*, (pp. 39 - 43). Egypt.

Wicaksono, A. F., Vania, C., Distiawan, B. T., & Adriani, M. (2014). Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. *28th Pacific Asia Conference on Language, Information and Computation*, (pp. 185-194).

Wiebe, J. M. (1990). Identifying Subjective Characters in Narrative. *COLING '90 Proceedings of the 13th conference on Computational linguistics* (pp. 401- 406). Association for Computational Linguistics.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246 - 253). ACL.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. *Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 486–497). Mexico City, Mexico.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities.*

Wiegand, M., & Klakow, D. (2009). The Role of Knowledge-based Features in Polarity Classification at Sentence Level. *In Proceedings of the 22nd International Florida Artificial Intelligence*, (pp. 296-301).

Wiegand, M., Balahur, A., Klakow, D., Roth, B., & Montoyo, A. (2010). A Survey on the Role of Negation in Sentiment Analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, (pp. 60 - 68). Uppsala.

Wiktionary. (n.d.). Retrieved from https://www.wiktionary.org/

Williams, G. K., & Anand, S. S. (2009). Predicting the Polarity Strength of Adjectives Using WordNet. *Third International AAAI Conference on Weblogs and Social Media*, (pp. 346-349).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and*

Empirical Methods in Natural Language Processing (pp. 347-354). Vancouver: Association for Computational Linguistics.

Wordnets in the World. (2014, 06 07). Retrieved from The Global WordNet Association: http://globalwordnet.org/wordnets-in-the-world/

WordNets in the World. (2014, 10 01). Retrieved from The Global WordNet Association: http://globalwordnet.org/wordnets-in-the-world/

Wu, Y., & Oard, W. (2009). Beyond topicality: finding opinionated Chinese document. *Proceedings of the 2009 Annual Meeting of the Association of the American Society for Information Science & Technology (ASIS&T)*, (pp. 6-11). Vancouver, British.

Xia, R., Zond, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 1138-1152.

Xu, G., Meng, X., & Wang, H. (2010). Build Chinese Emotion Lexicons Using A Graph-based Algorithm and Multiple Resources. *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1209–1217). Beijing: ACL.

Yang, M., Zhu, D., Mustafa, R., & Chow, K. P. (2014). Learning Domain-specific Sentiment Lexicon with Supervised Sentiment-aware LDA. *ECAI 2014 (21st European Conference on Artificial Intelligence).* Prague, Czech republic.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. *International Conference on Machine Learning*, (pp. 412 - 420).

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. *Proceedings of the Third IEEE Internati*, (pp. 427 - 434).

Yussupova, N., & Bogdanova, D. (2012). Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach. *The Second International Conference on Advances in iInformation Mining and Management*, (pp. 8-14). Venice, Italy.

Zanasi, A. (2007). *Text Mining and its Applications to Intelligence, CRM and Knowledge Management.* Southampton, UK: WIT Press.

Appendix A: Web Scrapper for extracting opinions

File Edit Format Run Options Windows Help

```
import requests
import bs4
import codecs

search_URL='http://www.lankadeepa.lk/index.php/top_story/490583' # url
oFile = codecs.open('D:\\NishanthaAUCK\\Thesis_Experiments\\outN.csv','w','utf-8') #output file
oFile.write("Date\tcomment\n")

response = requests.get(search_URL)
soup = bs4.BeautifulSoup(response.text, "html.parser")
def get_comment():
    return [span.get_text().encode('utf-8') for span in soup.find_all('span')]
def get_date():
    return [(p.get_text()) for p in soup.select('p')]

# Writing to the csv file
for j in range(len(get_date())-1):
    oFile.write("%s\t%s\n"%(get_date()[j]),(get_comment()[j]))
oFile.close()
```

Appendix B: Sample Opinions

Topic	***Opinion***
හරිත මන්ත්‌රීනියට ලංකාවේ හොඳක් නොපෙනේ	*සුමන්තිරන්ලා වගේ කොටි හම පොරවගත්ත බූරුවෝ ඉන්නකන් මේවා නවත්තන්න බෑ, ඉස්සෙලම රට ඇතුලේ ඉන්න කාලකන්නි හදල ඉන්න ඕන ඔක්කොටම කලින් එතකොට පිට රටවල එවුන් හැදෙයි.*
කාඩිෆ් පිටියට පැන්න කොටියා ළඟදී ලංකාවට	*ලාරාගේ අදහස හොඳටම ඇති. ශ්‍රී ලංකාවට පය තියන්නවත් නොදිය යුතුයි.*
බීබී ඔතන්නා මරා සල්ලි ගනී	*මෙහෙමත් ජරා මිනිස්සු*
හොරිවිල වෙද ගෙදරින් ගමගෙදරක්	*මේවා ඇත්තටම ගොඩක් හොඳ වැඩ*
සමුළු උපහාරට දැවැන්ත සිංහයෙක් රාජගිරියේ	*පසිදු, ඔයා හරියටම හරි තව සමහරු පාන් ගෙඩිය රුපියලකින් වැඩි කරහම කෑ ගහනවා ඒත් කඩේට ගිහිල්ල වැඩිපුර රුපියලක් දීල ෂොපින් බෑග් එකක් අරගෙන තවත් රුපියලක් වැඩිපුර දීල ඒ පාන් ගෙඩිය පෙති කපා ගෙනත් යනවා !!*
රියදුරු පරීක්ෂණේ ඉහළින්ම සමත් වේ	*හොඳ වෙලාවට කාරෙකේ ඇතුලේ පොන්සෙකා මහත්තයා හිටියෙ නැත්තෙ බැරිවෙලාවත්, ඇතුලේ හිටියනම්, එයා කියයි මෙක මහින්ද රාජපක්ශගේ කුමන්ත්‌රනයක් කියල ජීනීවා වලටත් කම්ප්ලේන් කරයි*
රියදුරු පරීක්ෂණේ ඉහළින්ම සමත් වේ	*බයවෙන්න එපා, දෙවැනි බලවේගය වෙන්න වැඩිය අමරුවෙන එකක් නැහැ*
සපත්තු දෙක කිලෝ අටසීයයි	*මෙයා කරන්නේ බොරු වෙන්න ඕන නැත්නම් හොස්පිටල් තමය ඉන්න වෙන්නේ කකුල් කඩාගෙන*
පොදු රද මඩුලු මහජන සමුළුව ගාල්ලේ	*මගේ රට ගැන ආඩම්බරයි ජාතියේ පිනට පහළ වූ ජනපතිට තෙරුවන් සරණයි*
පොලිසිය මාධ්‍යවේදියාට ගහලා	*බලහත්කාරකම් කර පහර දීම බොහොම වැරදියි දෙකෙන් එකක් කර විභාගය පාස් වුණහම ඊළඟ එක කළා නම් ප්‍රශ්නයක් නැහැ*

Appendix C: Python code Calculating positive negative score for Adjective and Adverbs

File Edit Format Run Options Windows Help

```
# -*- coding: utf-8 -*-
import string, random, codecs,re, csv,cStringIO
from itertools import izip
def ajective_matrix():
    infile_SinhapaOpinionsclassified = codecs.open('D:\\Nishantha\\DCT Summer1\\Data\\dataset2full_Nostop635.txt','r','utf-16')## Opinion file
    infile_SinhalaSentimentAdjectives =  codecs.open('D:\\Nishantha\\NishanthaAUCK\\Data\\Adjectives\\AdjectivesfronSinhalaSentiwith12synsND5974.txt','r','utf-16'
    outfile=codecs.open('D:\\Nishantha\\DCT Summer1\\Data\\AdjSentscoredataset2full_Nostop635.txt.csv','w','utf-16')## Output File
    file1=list(infile_SinhapaOpinionsclassified)
    file2=list(infile_SinhalaSentimentAdjectives)
    print len(file1)
    i=0
    for lineX in file1:
        x=lineX.split('\t') # split into columns i.e a opinion
        #for x1 in x[4].split(' '):# opinion split into words, len(x[4])= length of the opinion
        Positive_adjScore=0
        Negative_adjScore=0
        for lineY in file2: # Adjectives file line
            y=lineY.split('\t') # Adjectives line split into columns
            if y[0]!="":
                pat0=' '+y[0]+' '
            if y[1]!="":
                pat1=''+y[1]+' ';
            if((re.search(pat0,x[0])) or ((re.search(pat1,x[0])is not None)):#or (re.search(pat2,x[4])is not None)or(re.search(pat3,x[4])is not None) or(re.searcl
                #or(re.search(pat7,x[4])is not None) or(re.search(pat8,x[4])is not None)or(re.search(pat9,x[4])is not None) or(re.search(pat10,x[4])is not None)):
                #print pat0,pat1
                Positive_adjScore=float((y[13]).encode('utf-8')) # y[13] Adjective positive score
                Negative_adjScore=float((y[14]).encode('utf-8')) # y[14] Adjective positive score
                outfile.write('%s\t%s\t'%(Positive_adjScore,-1*Negative_adjScore))
                #outfile.write('%s\t%s\t'%(pat0,pat1))
        outfile.write('\n')
def adverb_matrix():
    infile_SinhapaOpinionsclassified = codecs.open('D:\\Nishantha\\NishanthaAUCK\\Data\\Opinions\\opinion2085classifiedclened.txt','r','utf-16')## Opinion file
    infile_SinhalaSentimentAdverbs =  codecs.open('D:\\Nishantha\\NishanthaAUCK\\AssignscorePN\\FinalNegativeScore.txt','r','utf-16')## Positive
    outfile=codecs.open('D:\\Nishantha\\NishanthaAUCK\\AssignscorePN\\NegativeScore.txt','w','utf-16')## Output File
    file1=list(infile_SinhapaOpinionsclassified)
    file2=list(infile_SinhalaSentimentAdverbs)
    print len(file1)
    i=0
    for lineX in file1:
        x=lineX.split('\t') # split into columns i.e a opinion
        i=i+1
        Positive_advScore=0
        Negative_advScore=0
        for lineY in file2: # Positive line
            y=lineY.split('\t') # Positive line split into columns
            #print lineX
            pat0=' '+y[1]+' '; pat1=' '+y[2]+' ';pat2=' '+y[3]+' ';pat3=' '+y[4]+' ';pat4=' '+y[5]+' '
            #print pat
            if  ((re.search(pat0,lineX)is not None) or (re.search(pat1,lineX)is not None)or (re.search(pat2,lineX)is not None)or (re.search(pat3,lineX)is not None)or
            #if ((re.match(pat0,x[0])is not None) or (re.match(pat1,x[0])is not None)or (re.match(pat2,x[0])is not None)or (re.match(pat3,x[0])is not None)or (re.mat
                #x1==y[0]:#y[0]= first column i.e Adjectives

                Positive_advScore=((y[0])) # y[1] positive score
                #print y[1]
                #print lineX
                #Negative_advScore=(-1)*(float(y[6].encode('utf-8'))) # y[1] Adverb positive score
                outfile.write('%s\t'%(Positive_advScore))
                #outfile.write('%s\t%s\t%s\t%s\t%s\t'%(y[1],y[2],y[3],y[4],y[5]))
        outfile.write('\n')
    outfile.close

adverb_matrix()
```

Appendix D: Extracting all the words with prefixes or suffixes from the dictionary

File Edit Format Run Options Windows Help

```
# encoding: utf-8
# Code for develop Postive and Negative words for Sinhala
# Resources: English Sinhala online Dictionary, Sinhala Word Net
# Prefixes: "අ", "අව්", "අන","වි","නො","කු","නිර්","නිශ්","නිෂ්","සු" "දු","දුර්","දුශ්","දුෂ්","නි" \Aඅව්
# Example: අපහසු-> Negative, පහසු->Positive D:\Nishantha\NishanthaAUCK\Data
import codecs
import re

kewords= codecs.open('D:\\NishanthaAUCK\\Data\\ENSNonlinedic11sinhalawithblank0.txt','r','utf-16') #dictionary
#kewords= codecs.open('D:\\NishanthaAUCK\\Data\\From Viraj\\Sinhala Linguistic Resources\\Distinct List\\Distinct List\\DistinctListAscendingorder1.txt','r','utf-16')
#outfile=codecs.open('D:\\NishanthaAUCK\\Positive_Nagative words\\NegativeDUS_list.txt','w','utf-16') # loutput:ist flie "අ, අව්, @නො, නි, වි" \Aඅව්
Keyword_list=list(kewords)#"dictionary
Full_List= codecs.open('D:\\NishanthaAUCK\\Positive_Nagative words\\FulllistPNforgenerate.txt','r','utf-16') # for Extracting Final Positive or negative List from Dict
Full_List_list=list(Full_List) #Extracting Final Positive or negative List from Dictionary

outfile=codecs.open('D:\\NishanthaAUCK\\Positive_Nagative words\\Step3_Synsets_1_06_negative.txt','w','utf-16') # loutput:ist flie "අ, අව්, @නො, නි, වි" \Aඅව්

#Key_word= []

################ Extracting words with prefix from the dictionary: STEP 1 #############
for each in Keyword_list: # Dictionary
    word=each.split('\t') # split the dictionary line to words using tab
    if re.search(ur"^දුස්[\w.]+", word[1], re.UNICODE): # serch a word for given prefix
        outfile.write('%s\t%s\t%s\t%s\t%s\t%s\n' %(word[0],word[1],word[2],word[3],word[4],word[5])) # write the words starting with prefix with all the synonyms and E
outfile.close()

 Extracting Corssponding words eg: given word අජීරිය and get ජීරිය : STEP II ############
A_List= codecs.open('D:\\NishanthaAUCK\\Positive_Nagative words\\NegativeKU_list.txt','r','utf-16')
A_List_list=list(A_List)#list

for wordL in A_List_list: # form the list created in step 1
    rowL=wordL.split('\t') # A_list row
    #print rowL[1],rowL[1][4:]
    X=rowL[1][2:]
    #Pword=re.search(ur"^අ(.*)",rowL[1],re.UNICODE)
    #print rowL[1]
    for wordD in Keyword_list: # serching the word without prefix and all it synonyms
        rowD=wordD.split('\t')
        # print rowD[0] # for unique word list Viraj
        #if (Pword.group(1)==rowD[1]):
        if (X==rowD[0]):
            #print X
            #print rowD[0]
            #outfile.write('%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n' %(rowL[0],rowL[1],rowL[2],rowL[3],rowL[4],rowD[0],rowD[1],rowD[2]))
            outfile.write('%s\t%s\t%s\t%s\t%s\t%s\t%s\n' %(rowL[0],rowL[1],rowL[2],rowL[3],rowL[4],rowD[0])) # for unique word list viraj
outfile.close()
##########Extracting Final Positive or negative List from Dictionary
i=0
for full_word in Full_List_list: # Generated Positive or Negative List
    row_full_word=full_word.split('\t')
    i=i+1
    for dic_word in Keyword_list: # Dictionary words
        row_dic_word=dic_word.split('\t')
        #print len(row_full_word)
        if ((row_full_word[0]==row_dic_word[1]) or (row_full_word[0]==row_dic_word[2])or (row_full_word[0]==row_dic_word[3])or (row_full_word[0]==row_dic_word[4]
            #print row_full_word[1],row_dic_word[1]
            outfile.write('%s\t%s\t%s\t%s\t%s\t%s\t' %(row_full_word[0],row_dic_word[1],row_dic_word[2],row_dic_word[3],row_dic_word[4],row_dic_word[5]))
    print i
    outfile.write('\n')
outfile.close()
```

Appendix E: Tagging Adjectives and Adverbs

File Edit Format Run Options Windows Help

```
import string, random, codecs,re, csv,cStringIO
## paersing the adverb adjectives in a opinion. Read a opinion and find the adverb adjectives words (tag)
def PositiveNegative_tagging():
    infile_SinhapaOpinionsclassified = codecs.open('D:\\Nishantha\\NishanthaAUCK\\Thesis_Experiments\\opinion2083classifiedclened.txt','r','utf-16')## Opinion file
    infile_Adjective_list =  codecs.open('D:\\Nishantha\\NishanthaAUCK\\Thesis_Experiments\\Chapter6\\AdjAdv\\Adverbs-GldStd-671 - WordListtab.txt','r','utf-16')##
    outfile=codecs.open('D:\\Nishantha\\NishanthaAUCK\\Thesis_Experiments\\Chapter6\\AdjAdv\\Adv_Available.txt','w','utf-16')## Output File

    file1=list(infile_SinhapaOpinionsclassified)
    file2=list(infile_Adjective_list)
    #print len(file2)
    for lineY in file1: # Opinions
        #print (lineY.strip())
        for wordP in file2:
                #print wordP.split()[0]
                #if (wordP.split()[0] in lineY):
                if(re.search(" "+ wordP.split()[0]+" ",lineY.strip(),re.UNICODE)is not None): #and wordP!=u'\r\n'and wordP!="":
                    #print wordP.split()[0]
                    outfile.write('%s\t'%wordP.split()[0])
                #else: outfile.write('%s\t'%0)
        outfile.write('\n')
    outfile.close
PositiveNegative_tagging()
```