SENTIMENT ANALYSIS OF SINHALA TWEETS

Warna Ieshaka Karunarathne

(189328H)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2020

SENTIMENT ANALYSIS OF SINHALA TWEETS

Warna Ieshaka Karunarathne

(189328H)

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: …… Date: …….

Name: W.I.Karunarathne

The supervisor/s should certify the thesis/dissertation with the following declaration.

The above candidate has carried out research for the Masters Dissertation under my supervision

Signature of the supervisor: ……. Date: ……..

Name: Dr. Uthyasanker Thayasivam

ACKNOWLEDGEMENT

I would like to convey my genuine appreciation to my supervisor Dr. Uthyasanker Thayasivam for his determined efforts, continuous supervision and assistance. I have been extremely fortunate to have him as my supervisor, whose guidance and enthusiasm helped me to improve. His motivation, persuasion and patience helped me to overcome many crisis situations and complete this research successfully.

My heartfelt appreciation is rendered to all my friends for their continuous assistance and encouragement given to me during this challenging and hectic endeavour.

Most significantly, this wouldn't have been possible without the love and support of my parents. They have been a continuous source of love, concern, support and strength all these years. I am grateful to my parents for believing in me, cheering for me and being patient and supportive during this critical period in my academic life, which motivated me to successfully complete the research.

ABSTRACT

Sentiment analysis has become a popular topic since the last decade. The increase in the use of internet has led to the increase of user-generated content. This has played an important role in making sentiment analysis more popular among researchers. The user-generated content can provide some valuable insight about the public opinion to the government and various industries.

This research has mainly focused on sentiment analysis of Sinhala language. Sinhala is the most spoken language in Sri Lanka. With the increased use of the internet and social media, there is a considerable amount of information communicated via Sinhala. This has presented a good opportunity to mine the information presented in Sinhala language. Performing Sinhala language sentiment analysis has some difficulties, as Sinhala is morphologically rich and is a language of free order compared to English. Lack of Sinhala language resources has brought challenges from gathering and generating data sets to stemming / lemmatizing algorithms. This research has tried to address the above challenges by developing a Sinhala dataset suitable for sentiment analysis and by developing a stemming algorithm for Sinhala. The dataset is developed by collecting Tweets from Twitter and it has been manually annotated.

In addition to the resource creation, sentiment analysis of Sinhala language is also performed using word embedding as features. Several sentiment analysis experiments are performed by using several machine learning techniques. The accuracy as well as precision and recall are used to identify the best performing model. The problems faced when conducting sentiment analysis for Sinhala language are discussed in the research. The research has discussed the difference between the user-generated content in English and Sinhala.

Table of Content

Table of tables

Table of figures

Chapter 1

Introduction

- Sentiment Analysis
- User Generated Content
- Web Content in Sinhala Language
- Sentiment Analysis for Sinhala Language
- General Approaches to Sentiment Analysis
- Main Challenges in Sentiment Analysis
- Motivation
- Applications
- Objectives of the Research
- Contribution of the Research

The development of World Wide Web has resulted in a rapid increase in content generated by users. This generated content is a vital resource needed to perform sentiment analysis. This information is generated from blogs, online newspapers, reviews and social media networks such as Facebook and Twitter. With content created by users growing over the last decade, sentiment analysis has become a common subject among the researchers.

The field of study which investigates the views, feelings, attitudes of people towards various things such as goods, services, organizations, individuals is known as sentiment analysis or opinion mining.[1] Sentiment analysis can be classified into three parts as follows,

- Document level sentiment analysis – It is used when classifying the sentiment of the whole document.
- Sentence level sentiment analysis- It is used when there are many sentences containing several opinions.
- Aspect based sentiment analysis- It is used when opinions about multiple features are expressed.

In 2009, Agarwal, Mckeown and Biadsy reported that the task of analysing sentiments has shifted from analysing document level to analysing sentence level. [2]

1.1 User-Generated Content

Development of the internet has led to a swift growth in the content created by the users. Users are provided with various platforms to interact through the internet. [3] Some of these platforms are mentioned below,

- Blogs – A blog is a website continually updated and owned by a single person or small group of people. Blog is used to express their views about some relevant topics in the society.
- Online newspapers – All most all the newspapers have a website and all the news in the paper are also published online.

- Reviews – Product reviews are given by customers more often these days.
- Social media – Facebook, Twitter and G+ are the most popular social media networks. These social media networks are used by millions of people to express their views.

Through these platforms, a lot of user-generated content are created. Sentences in User-generated content can be divided into two parts as, [3]

- Subjective sentences- An opinion is carried out through the sentence
  Example: - It was a good match.
- Objective sentences – A fact is carried out through the sentence instead of an opinion.
  Example: - The match was between Sri Lanka and Zimbabwe.

Subjective user-generated content can be classified further into three sections based on the sentiments expressed. They are, [3]

- Positive - Example: - Good performance by the team.
- Negative - Example: - You behaved very badly.
- Neutral – Example: - I usually get sleepy at night.

1.2 Web Content in Sinhala Language

Sinhala is the most spoken language in Sri Lanka. With the introduction of Unicode (UTF-8) standards for Sinhala language, the number of web pages in Sinhala language has gone up. Internet has reached out to more people within the country, therefore the number of users and the contributors also have increased rapidly. Some examples for Sinhala websites are, http://sinhala.adaderana.lk/, http://www.gossiplankanews.com/ , http://topsinhalablog.com/ etc. The use of Sinhala in social media sites such as Facebook and Twitter has also increased in recent years. Most people have used social media sites to express the opinions and therefore social media has become a valuable resource for researchers to analyse the feelings and the opinions of the general public towards a desired topic.

1.3 Sentiment Analysis for Sinhala Language

As discussed above, the number of users and the web content for Sinhala language has increased during the past few years. While sentiment analysis has become a very prevalent subject of research in the past decade, the study of sentiment for Sinhala language is still in the infant stage. This is due to the fact that Sinhala is a low resource language and therefore the most essential resources needed to perform Sinhala sentiment analysis such as an annotated Sinhala datasets and Sinhala sentiment lexicons are unavailable. [4] Various forms of content created by users could be used for an analysis of sentiments.

Sentiment analysis for Sinhala language has to face many challenges as Sinhala is a less resource language. The accessibility of tools, annotated corpus and other resources are limited or are in the development phase for a low resourced language. This research has focused on developing resource for Sinhala language while performing sentiment analysis.

1.4 General Approaches to Sentiment Analysis

Although sentiment analysis has become a very common subject of study, the majority of the work is done for English only. [4] Very little work has been done for less resource languages such as Sinhala. Some of the common methods used for an analysis of sentiments are described below. [3]

- Using subjective lexicon- List of words of a given language where for each word, a sentiment score is given specifying the positive or negative sense of the word, is called a subjective lexicon. In this approach, each word in a text is assigned the respective sentiment score given in the subjective lexicon and finally summed up to get the total sentiment score. The text is categorised as positive if the total is positive and if the total is negative the text is categorised as negative.

- Using N-Gram modelling- In this methodology, an N-Gram (Uni-Gram, Bi-Gram, Tri-Gram, etc.) model is fitted using training data and classification is performed on the test data using the model formed.
- Using machine learning techniques- Supervised learning techniques like Naïve Bayes, Support Vector Machine (SVM) and maximum entropy are also used for evaluating sentiments.

Most researchers have used N-Gram modelling and machine learning algorithms to accomplish sentiment analysis. The first Sinhala sentiment analysis study, done by Medagoda, Whalley & Shanmuganathan in 2013,[4] has used the machine learning methods to perform sentiment analysis. They noted that the task of mining opinion and classifying sentiments is complex, and that more research is required to develop more efficient algorithms that can be applied to various languages.

1.5 Main Challenges in Sentiment Analysis

Challenges faced when performing sentiment analysis is discussed in this section.

1.5.1 Challenges in Sentiment Analysis overall

Challenges faced by researchers while performing sentiment analysis for any language are listed below. [3]

- Noise (slangs, abbreviations) – Data from the internet have got noise. Abbreviations and slang words are used by most people. For example, “gud nyt”. This has made the analysis more complex.
- Contextual information – same word can have different understanding depending on the context it is been used in. For example

  “මේක දිග පාරක්”

  “මේකට දිගු ආයු කාලයක් තියනවා”

  In both the sentences, the meaning of the words “දිග” and “දිගු” is the same, but the first sentence is a negative comment and the second sentence is a positive

comment, therefore it has become important to identify the context of the sentence.

- Sarcasm detection – Sarcasm detection can be a hard task for even a human being, making a computer understand sarcasm has become even more difficult.
- Lack of resources – When dealing with non-English languages, lack of adequate resources, tools and annotated corpora are major disadvantages.

1.5.2 Challenges in Sentiment Analysis for Sinhala

Apart from the above mentioned challenges, sentiment analysis of Sinhala language has faced additional challenges,

- Unavailability of a Gold standard dataset
- Unavailability of Lexicons
- Unavailability of resources such as POS tagger etc.
- Most Sinhala comments are written using the English letters

1.5.3 Challenges in collecting Sinhala Tweets

There are few challenges which are specific when collecting Sinhala tweets from twitter.

- Presence of emojis, mentions and URLs
- Presence of English words in the tweet

1.6 Motivation

With the advent use of the internet throughout the country, feelings expressed in Sinhala over the internet have increased. Commenting in Sinhala has been enabled in most of the social media networks and e-commerce websites. This has created a lot of content created by users which can be analysed and used to understand the state of the mind of the general public towards something.

For example, by analysing the user-generated content, a government can understand the state of mind of the public towards the government.

Limited amount of work is conducted in the field of Sinhala language sentiment analysis, therefore Sinhala user-generated content is not appropriately analysed. Thus this research would contribute towards the development of sentiment analysis of Sinhala language.

1.7 Applications

Below are several places where sentiment analysis can be used.

- Review systems - Restaurant and hotel review systems can be created by analysing the user comments about the particular restaurant or the hotel.
- Product analysis – Companies can use sentiment analysis to analyse user comments about their products to ensure better sales by correcting the mistakes pointed out by the public.
- Analysing open ended questions in a survey questionnaire – Most of the time these open ended questions are not analysed. With sentiment analysis, the answers to these questions can be at least classified in to positive, negative or neutral states.

1.8 Objectives of the Research

It is a tough task to conduct Sinhala language sentiment analysis with the scarce resources. Creating suitable resources to implement sentiment analysis for Sinhala language and performing sentiment analysis on the Sinhala dataset can be considered as the main focus of this research.

The primary objective is achieved using the following sub objectives,

- Creating a Sinhala dataset suitable to train and validate algorithms for Sinhala sentiment analysis.
- Performing sentiment analysis using machine learning techniques.

1.9 Contribution of the Research

This research is mainly focused on Sinhala language. As described above, Sinhala has lacked adequate resources to perform sentiment analysis. As a part of the research, an

annotated Sinhala dataset is created by using the sentences collected over the internet and a Sinhala word embedding is created.

Chapter 2

Literature Review

- History and Growth of Sentiment Analysis
- General Methods Used in Sentiment Analysis
- Sentiment Analysis for Other Languages
- Sinhala Language Sentiment Analysis
- Summary

Sentiment analysis has become a trend in the research community from the last decade. Study of sentiments for English has been carried out more but the amount of research performed for non-English languages is very small [5]. As Agarwal, Biadsy & Mckeown in 2009 [2] have suggested, sentiment analysis has progressed from document level to sentence level analysis. The work done in the past within the field of sentiment analysis is discussed in this chapter.

2.1 History and Growth of Sentiment Analysis

User produced content has become an important information source to explore the sentiments of individuals about different products and services. With the development of technology, internet has become more accessible to people. This has resulted in generation vast amount of user generated content. Therefore it has become more important to extract sentiments from these user generated content. [3] Sentiment analysis has become a trend in the research community from the last decade. Sentiment analysis for English has been carried out more but the amount of research done for non-English languages is very small.[4]

Sentiment analysis has been investigated mainly at 3 levels. [1]

- Document level – It can be determined whether the entire document communicates a positive or a negative view.
  E.g.: Product review system where the system decides whether the review states a favourable or unfavourable general opinion on the product.
- Sentence level – This level goes to the phrase and decides whether each phrase expresses a positive, negative or neutral view.
- Entity and Aspect level – Neither the document nor the sentence level sentiment analysis determine what precisely people loved or did not love. Aspect level sentiment analysis does more granular level examination.
  E.g.: "The phone's look and feels is good, but the battery life is low"

The above sentence has two contrasting opinions about two feature of the same product. Identifying these opinions is known as Aspect level sentiment analysis.

Opinions also have two categories.

Regular opinions – only states a sentiment on a certain product or aspect.

E.g.: “Coke tastes good”

Comparative opinions – This will compare multiple products or aspects.

E.g.: “Fanta tastes better than lemonade”

As suggested by Agarwal, Biadsy & Mckeown in 2009, sentiment analysis has moved from document level to sentence level analysis.[2]

2.2 General Methods Used in Sentiment Analysis

As mentioned in the Introduction section, Some of the common methods used for the analysis of sentiments are described below [3]

Using subjective lexicon

Using N-Gram modelling

Using machine learning techniques

2.2.1 Sentiment analysis using subjective lexicon

The most important markers of emotions are nostalgic terms which are widely used to convey positive or negative emotions. A collection of such words and expressions is called a lexicon of sentiment.[1] The words in these lexicons are marked with their prior polarity. This prior polarity may be different from the contextual polarity of the expression or word. [6]

In 2005, Wilson, Wiebe and Hoffmann [6] have proposed a new experiment on automatic contextual and prior polarity distinction. Starting with a comprehensive set of features marked with a prior polarity, the contextual polarity of the sentences containing the occurrences of those features in the corpus was identified. They also implemented new manual annotations of contextual polarity and a constructive analysis of the inter-annotator agreements. They have also applied contextual polarity assessments to prevailing annotations in the Multi-perspective Question Answering (MPQA) Opinion Corpus to build a corpus. They have manually annotated 15,991 subjective expressions

of which 28% had no subjective expressions, 25% had only one expression, and 47% had two or more expressions. They have used 2,808 subjective expressions as the training set, used for exploratory data analysis and feature engineering and the second set of 13,183 subjective expressions in 10-fold cross-validation experiments. In this paper they have used a lexicon of over 8,000 clues with subjectivity and expanded it further by using a dictionary and a thesaurus. To identify the usefulness of the prior polarity alone in contextual polarity classification, they have built a model using just the prior polarities and evaluated the performance was evaluated using the development set. They have managed to achieve 48% accuracy for this simple classifier. As the next experiment they have used a two-step process that employs BoosTexter AdaBoost machine learning algorithm and a wide range of features. Each expression having a clue is classified as polar or neutral in the first step. The contextual polarity classification in to positive, negative or neutral is done for the phrases marked as polar in the second step. This has allowed the system to automatically define the contextual polarity of a large subset of sentimental expressions. This method has achieved an accuracy of 75.9%. As future work, they have mentioned that trying to identify expression boundaries might improve performance.

The conventional approach for predicting subjective sentence contextual polarity is to use a lexicon of terms with prior polarity, to first set prior polarity on focused expressions and then use the semantic and syntactic information in and around the sentence to generate the final result.. [2] They had used a multi-perspective question answering opinion corpus for their study. They had used DAL (Dictionary of Affect in Language) to derive the lexical scores. DAL technique has allowed them to dynamically score the majority of words in their input without having to manually tag them. To identify the effect of context, they have supplemented lexical scoring with n-gram analysis. Each word was given 3 kinds of scores, namely evaluation, activeness and imaginary and the scores have a range from 1(low) – 3(high). Evaluation is the measure of polarity, activeness is the measure of activation and imaginary is the extent of the simplicity with which a word forms a mental depiction. Further they had suggested that,

since it includes different scores for various types of word inflection; morphological parsing and the likelihood of resulting errors are avoided. They have used Weka to implement logistic regression classifier to perform the classification. In a 3-way classification of positive, negative, and neutral phrases, they had managed to achieve high accuracy. They had some limitations in their study, the system requires accurate expression boundaries and the impact of the connectives such as “but”, “although” had not been accounted for.

Agarwal, Xie, Vovsha, Rambow & Passonneau in 2011 [7] had used a Part of Speech (POS) specific prior polarity fetures to create a lexicon. They selected the number of features relying on the word's prior polarity. They had used Dictionary of Affect in Language (DAL) and then used Word Net to expand it to acquire the prior polarities. They had normalized the evaluation score of DAL which had about 8000 English words. Then, words with a polarity below 0.5 are then regarded as negative and words with a polarity above 0.8 are regarded as positive and the others as neutral. If the word was not specifically included with the dictionary, then all synonyms have been extracted from Word Net, and checked every synonym in DAL. If the word was there in DAL, they had given the same evaluation value as its synonym and if none of the synonyms were there in DAL, then the word was not given any prior polarity. They had found a prior polarity of approximately 81% of words directly for the given data, and found a polarity of another 7.8% of words using the Word Net. They have thus succeeded in discovering a prior polarity of around 88.9% of English words.

2.2.2 Sentiment analysis using n-gram modelling

Pak and Paroubek [8] received a corpus of 300,000 tweets from Twitter in 2010, divided equally across three classes; positive, negative and neutral. They have used n-gram as a binary feature, and trialled with uni, bi, and tri-grams. With these n-grams as features they tested the models SVM (Support Vector Machine), Naïve Bayes and CRF (Conditional Random Field), and the Naïve Bayes model yielded the best results. They have trained two Bayes classifiers, one uses part-of-speech (POS) information and the

other uses existence of n-grams. N-gram based classifier had used the availability of an n-gram as binary feature in the post. The classification algorithm focused on POS information has estimated the likelihood of existence of POS-tags within various collections of texts and has used it to compute the subsequent probability. Although POS depends on the n-grams, the hypothesis of conditional independence of n-grams and POS data has been made for the ease of calculation. In order to improve the model's accuracy, they have proposed elimination of the common n-grams that don't infer any strong sentiment or infer sentence objectivity. Two approaches were introduced by them to differentiate common n-grams. The first approach focuses on measuring the entropy of the probability of an n-gram occurring in various datasets (different sentiments). The high entropy rate has suggested that an almost uniform distribution of the existence of an n-gram in distinct sentiment datasets. Such an n-gram therefore doesn't contribute much to the classification, whereas a low entropy value indicates that an n-gram has occurred more often in certain datasets than in others, and may therefore illustrate a sentiment. By setting a threshold value, they have governed the accuracy by extracting n-grams above threshold with entropy. They have also introduced a word "salience" for the second approach which is determined for each n-gram. The calculation given accepts a value from 0 to 1. The low value designates a lower n-gram salience, and such n-grams should be distinguished. They have regulated the system performance by fine-tuning the threshold value as with the entropy. They have obtained their best outcomes by using bi-grams, since bigrams provide a better balance among coverage and the ability to recognise the trends of sentiment speech. Finally they have recommended that, even though this research is done for English language this method can be extended to other languages as well.

In 2002 Pang, Lee & Vaithyanathan[9] used a syntactic strategy to classification of sentiments using N-Grams. They have used Part of Speech (POS) data together with the conventional n-gram method as a feature for implementing machine learning to assess the polarity. Different variants of n-gram method like unigrams existence, unigrams with frequency, unigrams plus bigrams, bigrams, unigrams plus POS, adjectives, most

frequent unigrams and unigrams plus positions are used in their study. They eventually came to the conclusion that the frequency of matching n-gram could be a variable that could minimize accuracy. They had got a maximum accuracy of 82.9% for unigrams presence approach on SVM among all the other experiments they had performed.

2.2.3 Sentiment analysis using machine learning techniques

Agarwal, Xie, Vovsha, Rambow, & Passonneau in 2011 [7] have presented POS-specific prior polarity features and have used an unigram technique, a feature centered technique and a tree- kernel centered technique to perform sentiment analysis on a data set collected from Twitter. They have obtained from a commercial source, 11,875 manually annotated tweets. They have developed a tree representation of tweets in one compact representation to combine several categories of features. They have used Partial Tree (PT) to measure the similitude between two trees. They had suggested that this method outperform both feature based method and the uni-gram model by a substantial margin.

Because of insufficient contextual knowledge, the study of sentiments in small texts like small single sentences and tweets is said to be difficult. Effectively solving this problem requires strategies that combine the small text information with previous knowledge and use more features than just bag of words.[10] In 2014, Santos and Gatti[10] introduced a novel deep convolutional neural network that uses knowledge from sentence to character level to interpret short texts' sentiments and is named Character to Sentence Convolutionary Neural Network (CharSCNN). The suggested architecture has two convolutional layers that are used to retrieve appropriate features from phrases and words of any size, and can effortlessly take advantage of the wealth of word embedding generated by unsupervised pre-training. They have also performed studies that demonstrate CharSCNN's efficacy in evaluating the sentiments of texts in the domains of movie reviews and tweets. The proposed program calculates a value for each sentiment category according to the given sentence. The arrangement of words in the sentence is taken as the input to generate the score to the sentence. The score is passed

through a series of layers, where character level features are extracted up to the sentence level. The key innovation in the proposed design of the network is the addition of two convolutional layers, allowing it to handle sentences and words of any scale. The network is trained by minimizing a negative likelihood over the training set. For the movie review dataset, they have managed to achieve state of the art performance of accuracy 85.7% while an accuracy of 86.4% for the Tweeter dataset.

In 2011, Glorot, Bordes and Bengio have suggested a deep learning method to the issue of sentiment classifier domain adaptation.[11] The Stacked De-noising Auto-encoder is the basic framework for their models. Auto-encoders are typically trained to minimize error loss. They have had access to unlabelled data for the tests from numerous domains, and labels from one domain only. They tackled the issue of domain adaptation with a two-step technique for sentiment classifiers. First, an extraction of higher-level features is learnt in an unsupervised way from the text reviews of all readily available domains using the above-mentioned auto-encoder with a rectified code layer. In the following phase, a linear classifier is trained on the source domain's transformed labelled data. They have used a linear SVM with squared hinge loss. This classifier is ultimately evaluated on the target domains. Their studies have shown that the existing state-of-the-art classifiers are defeated by the linear classifiers equipped with this highly learned feature representation of reviews.

2.3 Sentiment Analysis for Other Languages

Most researchers have used techniques used in sentiment analysis for the English language for non-English language, but with some restricted use of properties which are specific to a language such as morphological variations.[5] They have also noted that the application domains tend to be confined to a specific domain and there has been considerably less cross-domain research. In this research paper they had focused on two non-English languages namely Hindi and Russian. For Hindi language they had reviewed 2 research papers. One paper had used Support Vector Machine (SVM) as the first approach to evaluating the polarity. The second method that the reviewed paper

mentioned was to use Google translation to translate the Hindi corpus into English. Then the translated English corpus was given as the input to the classifier. The researchers of this reviewed paper have created a senti Word Net for Hindi by mapping the Synset belonging to English in senti Word Net to the corresponding to synset in Hindi as their third approach. The Hindi senti Word Net contained 16253 sysnsets which consist of adjectives, adverbs, nouns and verbs. Classification under the resource dependent method has been carried out when changing the several structural features such as modifying n-grams and with or without stemming. They had stated that the Google Machine translation based method produced poor results due to the translational errors.

A subjective lexicon has been developed by the second paper reviewed for Hindi language by Medagoda, Shanmuganathan and Whalley [5]. The lexicon has been developed by using a seed list comprising 45 adjectives and 75 adverbs. Findings of that research paper stated that the system of scoring has beaten the method of unigrams presence. The key drawback of the mentioned algorithm was Word Net's failure to perform disambiguation of the word sense. As the stem of a certain words contributed for computation of the polarity value, the results would have been influenced by morphological variation of the Hindi language.

Table 2-1 Review of two Hindi Sentiment analysis papers

Study	Application Domain	Classification Method	Features	Accuracy
I	Movie reviews: Corpus of 250 reviews	In-language using SVM	Term frequency	74.57%
			Term presence	72.57%
			TF-IDF	78.14%
		Machine translated using SVM	TF-IDF	65.96%
		Resource based using SVM	Most common sense	56.35%
			All sense	60.31%
II	Product review dataset translated from English to Hindi	Using subjective lexicon	Unigram presence	77.34%
			Simple scoring method	79.03%

For Russian language also Medagoda, Shanmuganathan and Whalley [5] had chosen 2 research papers to review. The focus of the first comparative research has been to check how lemmatization affects the accuracy of sentiment classification while the primary goal of the second study was to classify sentiments into five, three and two classes. In the first research, to improve the accuracy rate, a bagging algorithm was implemented into a Naïve Bayes classifier. The language-independent method tested in the second

research, with the Support Vector Machine (SVM) classifier, has been entirely reliant on attributes which are based on features such as n-gram, POS tags and parsing of dependence. They have also introduced a new feature in addition to n-grams which was close to n-grams called d-grams. D-grams are built from a parser tree of dependency, where syntactic relations connect the words. They have finally confirmed that the performance of sentiment classification algorithms in experiments that target a specific domain is greater than the experiments that don't target a specific domain.

Table 2-2 Review of two Russian Sentiment analysis papers

Study	Application Domain	Classification Method	Features	Accuracy
I	Bank customer reviews	Naïve Bayes	Bagging multinomial model	87.69%
		SVM	Length > 2	88.21%
II	Product reviews	SVM	N-gram + TF-IDF	90.4%
			D-gram + TF-IDF	91.3%

2.4 Sinhala Language Sentiment Analysis

This research has focused on the sentiment analysis of Sinhala language. As indicated in 2016 by Medagoda, Shanmuganathan and Whalley, [4] the attention obtained by low-resource languages such as Sinhala is significantly lower as the advancement of any sentiment analysis algorithm rest on on the accessibility of resources such as special lexicons and Word Net like tools.

As far as the Sentiment analysis of Sinhala language is concerned, the research done by Medagoda, Shanmuganathan & Whalley in 2015 was the first one. They had stated that eventhough sentiment analysis has become a prevalent research topic, the attention received by low resourced languages such as Sinhala is signifiacantly less. They had introduced a less complex yet successful technique in this research which could be used to classify comments of languages with less amount of resources. With the help of an sentiment lexicon of English (Senti Word Net 3.0), they have built a Sinahala sentiment lexicon. They had used a Sinhala dictionary which was online to translate the English lexicon to Sinhala. The dataset to evaluate the Sinhala sentiment lexicon was created by using 2083 mannualy annotated news article opinions. They had used Naïve Bayes algorithm, Support Vector Machine (SVM) and J48 decision tree algorithm as the classification algorithms. They had suggested that Naïve Bayes algorithm, which is based on Bayesian theorem, was more suited when the inputs are dimensionally high. Further they had suggested that the SVM algorithm was the best binary classification method and it uses an iterative training algorithm to construct an optimal hyper plane. The J48 decision tree algorithm, as suggested by them, is a univariate decision tree that makes use of knowledge gain to construct trees. They had managed to get an accuracy of 56%-60% for binary classification for the three algorithms, which is a significantly promising result for sentiment analysis of Sinhala.

In 2018, Liyanage and Ranatunga have performed sentiment analysis for Sinhala language in the News domain. [12] They have selected www.lankadeepa.com as their main data source and have written a web crawler to retrieve the data. The collected dataset was annotated by few annotators. They have used the directly available features to simplify the feature selection method. For initial experiments they have used bag of word model. Then tf-idf, 2-gram word vectors and finally word embedding techniques with different aggregation methods are also used. In their experiments they have used Support Vector Machine (SVM), Naïve Bayes and other deep learning techniques. With these experiments they have managed to prove the effectiveness of the word embedding features and performance of SVM and RNN in general text classification tasks. All of

their experiments have passed the baseline accuracy while Support Vector Machine (SVM) and Recurrent Neural Networks (RNN) have performed best.

2.5 Summary

Table 2-3 Summary of Sentiment Analysis papers

Study	Application Domain	Classification Method	Features	Accuracy
[6]	MPQA opinion corpus	Simple classifier that assumes that the contextual polarity of a clue is same as the prior polarity	Prior polarity	48%
		BoosTexter AdaBoost algorithm	Word tokens, word POS, word context, prior polarity	75.9%
[2]	MPQA opinion corpus	Logistic Regression	Chance baseline	33.33%
			N-gram baseline	59.05%
			DAL scores only	59.66%
			DAS + POS	60.55%

			DAL + Chunks	64.72%
			DAL + N-gram	67.51%
			All features included	70.76%
[9]	Movie reviews	Naïve Bayes	Unigram frequency	78.7%
			Unigram presence	81.0%
			Unigrams + bigrams	80.6%
			bigrams	77.3%
			Unigrams + POS	81.5%
			Adjectives	77.0%
			Top 2633 unigrams	80.3%
			Unigrams + position	81.0%
		Maximum Entropy	Unigram frequency	N/A
			Unigram presence	80.4%
			Unigrams + bigrams	80.8%
			bigrams	77.4%
			Unigrams + POS	80.4%
			Adjectives	77.7%

			Top 2633 unigrams	81.0%
			Unigrams + position	80.1%
		SVM	Unigram frequency	72.8%
			Unigram presence	82.9%
			Unigrams + bigrams	82.7%
			bigrams	77.1%
			Unigrams + POS	81.9%
			Adjectives	75.1%
			Top 2633 unigrams	81.4%
			Unigrams + position	81.6%
[7]	Twitter data	SVM	Unigram	71.35%
			Senti-feature	71.27%
			Tree Kernel	73.93%
			Unigram + Senti-feature	75.39%
			Kernel + senti-feature	74.61%
[10]	Stanford Sentiment Tree Bank	CNN	Word Embedding	85.7%

	Stanford Twitter Sentiment	CNN	Word Embedding	81.9%
[5] Hindi 1st paper	Movie reviews	In-language using SVM	Term frequency	74.57%
			Term presence	72.57%
			TF-IDF	78.14%
		Machine translated using SVM	TF-IDF	65.96%
		Resource based using SVM	Most common sense	56.35%
			All sense	60.31%
[5] Hindi 2nd paper	Product review dataset	Using subjective lexicon	Unigram presence	77.34%
			Simple scoring method	79.03%
[5] Russian 1st paper	Bank customer reviews	Naïve Bayes	Bagging multinomial model	87.69%
		SVM	Length > 2	88.21%
[5]Russian 2nd paper	Product reviews	SVM	N-gram + TF-IDF	90.4%

			D-gram + TF-IDF	91.3%
[4]	Comments from Lankadeepa	SVM	Sinhala Lexicon	43%
		Naïve Bayes	Sinhala Lexicon	44%
		J48	Sinhala Lexicon	44%
[12]	Comments from Lankadeepa	Logistic Regression	Word presence	84.22%
			2-gram word presence	71.60%
			TF-IDF	85.02%
			W2V word presence	84.23%
			W2V TF-IDF	83.58%
		SVM	Word presence	82.63%
			2-gram word presence	67.81%
			TF-IDF	85.32%
			W2V word presence	84.43%
			W2V TF-IDF	84.08%
		Random Forest	Word presence	82.18%
			2-gram word presence	64.37%
			TF-IDF	80.88%

			W2V word presence	83.73%
			W2V TF-IDF	83.53%
		Naïve Bayes	Word presence	76.89%
			2-gram word presence	73.35%
			TF-IDF	75.79%
			W2V word presence	77.69%
			W2V TF-IDF	77.29%
		Decision Trees	Word presence	75.24%
			2-gram word presence	63.97%
			TF-IDF	75.19%
			W2V word presence	76.54%
			W2V TF-IDF	75.79%
		Hybrid CNN + SVM	W2V skip gram	83.13%
		RNN LSTM	W2V skip gram	86.45

Chapter 3

Methodology

- Sinhala Dataset
- Sinhala Word Embedding
- Stemming
- Sentiment Analysis

The Sinhala dataset and Sinhala word embedding are the main resources needed to perform the sentiment analysis in this research. The methods and theories that are used to generate these resources are discussed in this chapter.

3.1 Sinhala Dataset

In this research, a Sinhala dataset suitable for Sinhala sentiment analysis is created by collecting Sinhala comments from Twitter. The dataset has contained Sinhala tweets across all domains such as politics, sports and reviews. All the sentences are manually annotated as -1, 0 and +1. -1 for negatively sensed sentences, +1 for positively sensed sentences and 0 for neutral sentences. The annotation process is done by 3 separate annotators and the score with the majority vote is considered as the score of the sentence.

The unwanted information in tweets, such as emojis, mentions and urls are removed by using a pre-processing library available in python. The tweets that contain English words are removed when doing the annotation.

Sentiment analysis is performed on the generated dataset, by using Sinhala word embedding generated through this research.

3.2 Sinhala Lexicon

Two Sinhala lexicons are also created as a part of the research; one lexicon for only positive Sinhala words and one lexicon for only negative Sinhala. They are created by translating 2 very popular English lexicons [13]. The "Ingiya" dictionary developed by the University of Colombo Computing School's (LRL) Language Resource Lab is used for translation purposes [14]. The English words in the respective English lexicons are used as the search key to construct the sentiment lexicons. There are several Sinhala synonyms for some English words, therefore all possible Sinhala synonyms for a given English word are inserted into the Sinhala lexicon. A sentiment score of + 1 is given to all words in the positive Sinhala lexicon, and all words in the negative Sinhala lexicon are given a sentiment score of -1. The three assumptions that Medagoda,

Shanmuganathan, & Whalley [15] had considered in developing their lexicon are also considered when developing these two Sinhala lexicons as well. The assumptions are,

- For the 2 languages, the sense of the word is the same.
- Sentiment score for the Sinhala word is same as the score for the English word.
- Part of Speech (POS) is the same for both the languages.

Since Sinhala is morphologically rich, some Sinhala words have got several word forms. It is a very difficult task to have all these word forms in a single lexicon. So the better approach is to convert the different word forms of a given word into a single form.

3.3 Stemming

Since Sinhala is morphologically rich, some Sinhala words have got several word forms. When we create a Word2Vec model, it will create one vector for each of these word forms. So the better approach is to convert the different word forms of a given word into a single form. In this research, a stemming algorithm is developed to accomplish the above mentioned task. Stemming is the act of trimming off the ends of words and also involves dropping of derivational affixes.

3.3.1 Stemming Method

First, all the distinct Sinhala words in the dataset will be arranged in the alphabetical order. Stem of the very first word of the unique word list is chosen as the word itself. Then the next word is checked to see if that word starts with the stem of the previous word. If it has started with the stem of the previous word, then the remaining part of the word is checked in the suffix list [16]. If that word part is in the suffix list, then that word is also given the same stem as the previous word. If the word part is not in the suffix list or the new word does not start with the stem of the previous word, its stem is considered to be as the word itself. Then the words in the dataset are replaced by the stemmed word of the respective word. Flow chart of the stemming algorithm used is shown in the figure 3.1.

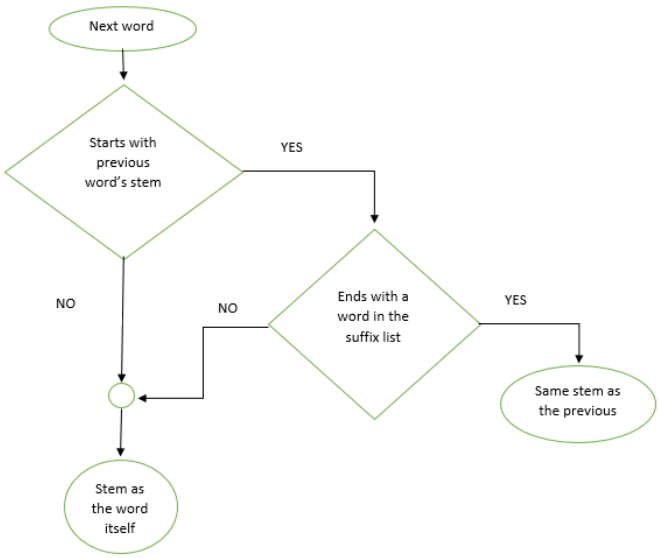

Figure 3.1: Flow chart of the stemming algorithm

3.4 Sinhala Word Embedding

As discussed in the Chapter 2, word embedding has become the latest approach for representing text in Natural Language Processing. Embedding algorithms such as Word2Vec has achieved state of the art performance. In the Word2Vec method, a high dimensional one-hot style representation of words is mapped to a lower dimensional vector while maintaining the word context. The effect of presence of slang words and abbreviations is minimized when using word embedding compared with the lexicon based method. There are several python libraries available to create a Word2Vec model such as "Gensim" and "Tensorflow". Gensim library is tried in this research.

3.5 Sentiment Analysis

Maximum entropy, SVM and Naïve Bayes (NB) are the supervised learning methods widely used in the study of sentiments. Therefore in this research, supervised learning algorithms such as NB, SVM, lightGBM, adaboost and XGBoost are considered while taking word embedding as the feature. Accuracy as well as Area under the curve (AUC) and weighted F1 score are being used to compare the trained models.

Chapter 4

Implementation

- Sinhala Dataset
- Sinhala Sentiment Lexicons
- Stemming Algorithm
- Sinhala Word Embedding
- Sinhala Sentiment Analysis
- Problems faced during implementation

The way that the research is implemented, and the problems faced during implementation are discussed in this chapter.

4.1 Sinhala Dataset

Sinhala dataset suitable to perform sentiment analysis is created as one of the main aims in this research. The dataset is created by collecting Sinhala tweets from Twitter. A total of 10,000 Sinhala tweets are collected across all domains. Then the dataset is manually annotated. A positively sensed sentence is given ‘+1’, a neutral sentence is given ‘0’ and a negatively sensed sentence is given a ‘-1’ value as shown in the figure 4.1. The annotation is done by 3 annotators.

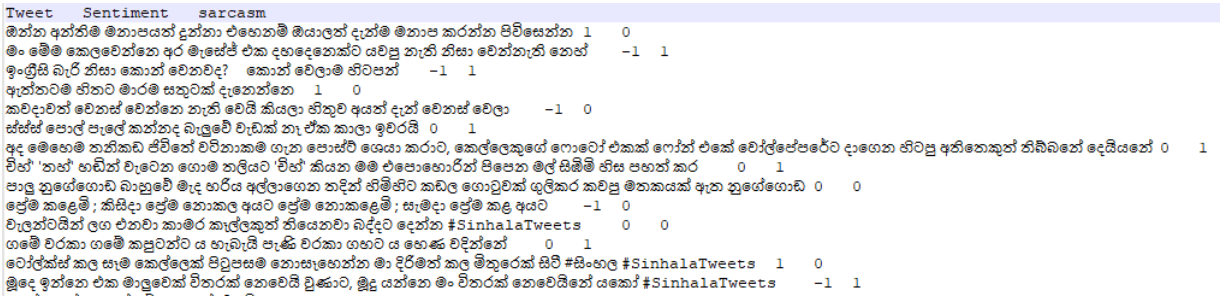

Figure 4.1 Annotated Sinhala Tweets

4.2 Sinhala Sentiment Lexicons

Two Sinhala sentiment lexicons (positive and negative) are created by translating popular English lexicons [13] using the "Ingiya" dictionary developed by the University of Colombo Computing School's (LRL) Language Resource Lab [14]. A part of “Ingiya” dictionary is shown in the figure 4.2.

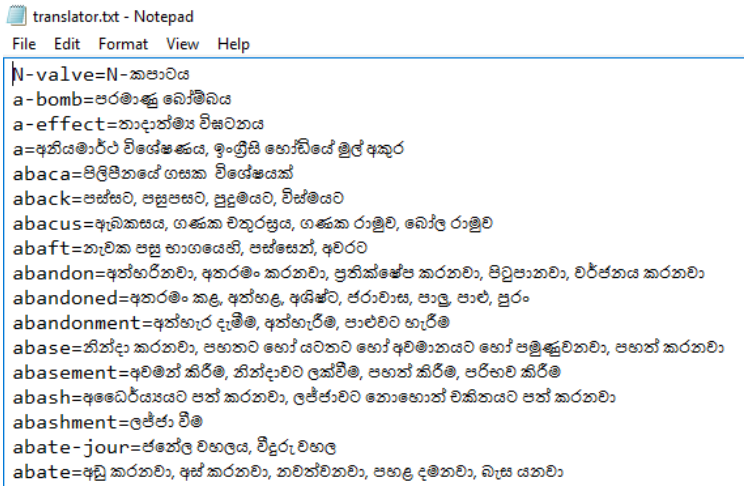

Figure 4.2 A part of "Ingiya" Dictionary

All words in the positive Sinhala lexicon are given a sentiment score of '+1' and a sentiment score of '-1' is given to all words in the negative Sinhala lexicon. The Sinhala lexicons created are shown in the figure 4.3.

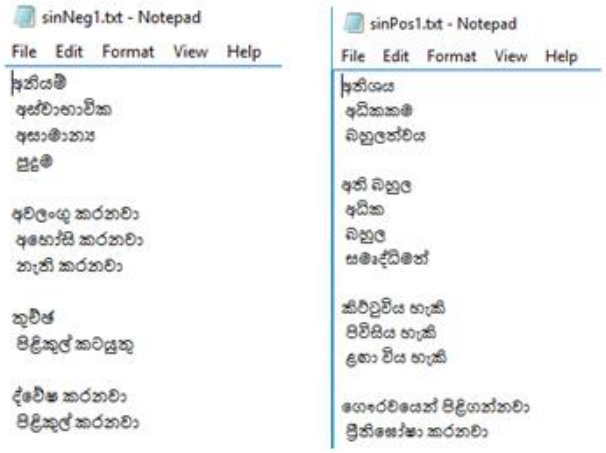

Figure 4.3 Sinhala lexicons with word combinations

4.3 Stemming Algorithm

Since Sinhala is a morphologically rich language, it is better to perform stemming or lemmatization before creating the word embedding. The method mentioned in section

3.3.1 of the Methodology is used in this research to do the stemming. The results from the stemming algorithm are shown in the figure 4.4.

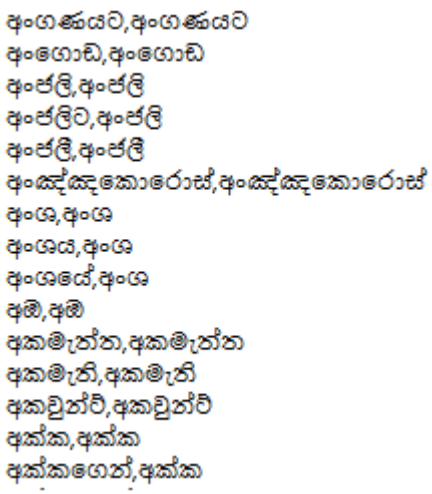

Figure 4.4 Sinhala word and its stem

The accuracy of this stemming algorithm is tested manually by using the 100 most frequent words in the dataset. The figure 4.5 has shown the 100 most frequent words in the dataset and 87% of them had the correct stem.

```
print(count.most_common(100))
```

```
[('නෑ', 1006), ('නම්', 538), ('මේ', 400), ('එක', 391), ('කියලා', 344), ('මට', 287), ('දැන්', 276), ('ඒ', 266), ('මම',
247), ('ගැන', 246), ('එකක්', 242), ('බන්', 227), ('වගේ', 225), ('මන්', 204), ('තමා', 203), ('කරන්න', 200), ('අද', 19
8), ('ඔව්', 194), ('වෙලා', 186), ('අපි', 180), ('හරි', 180), ('ඒක', 179), ('ඇති', 179), ('නේ', 173), ('නැති', 166), ('නි
සා', 166), ('එපා', 159), ('ඉන්න', 158), ('යන්න', 147), ('ඔබ', 146), ('වෙයි', 144), ('එහෙම', 144), ('ගන්න', 143), ('
වෙන්න', 143), ('එකේ', 134), ('බෑ', 134), ('වෙන', 131), ('තියෙනවා', 129), ('මං', 127), ('මගේ', 126), ('එක්ක', 119),
('උඹ', 113), ('වඩා', 112), ('ඇයි', 112), ('කරන', 108), ('උන්', 107), ('තාම', 106), ('ඔය', 105), ('කියන්නේ', 104), ('
ඕන', 104), ('වැඩ', 103), ('මේක', 102), ('ඉතින්', 101), ('ඕනේ', 101), ('ඔයා', 100), ('ද', 98), ('දන්නේ', 98), ('අය',
97), ('ආදරේ', 96), ('වල්', 96), ('නැද්ද', 96), ('බලන්න', 95), ('ගිය', 91), ('අනේ', 89), ('කතා', 89), ('වෙනවා', 88),
('අපිට', 88), ('එන්න', 88), ('නවත්', 87), ('වලට', 86), ('නෙමෙයි', 85), ('ජනපති', 84), ('ඉන්නේ', 83), ('තියෙන', 82),
('හා', 82), ('තියෙන්නේ', 82), ('ගිහින්', 81), ('ඒවා', 80), ('කරයි', 80), ('අපේ', 79), ('ඉන්නවා', 79), ('කරලා', 78), ('
විතරයි', 76), ('කියයි', 75), ('වැඩ්', 75), ('දෙන්න', 74), ('යනවා', 74), ('ගෙදර', 73), ('වෙන්නේ', 72), ('කියන්න', 71),
('හුළුවන්', 71), ('කළ', 70), ('යන', 70), ('මහ', 70), ('හෙට', 69), ('නැතුව', 67), ('ආදරය', 66), ('ඔබට', 66), ('මහින්
ද', 66), ('එකට', 65)]
```

Figure 4.5 100 frequent words in the dataset

Even though the algorithm has a good accuracy for the above 100 words, it has error that are shown in the figure 4.6.

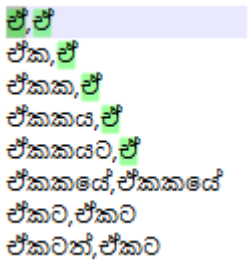

Figure 4.6 Errors in the stemming output

4.4 Sinhala Word Embedding

10,000 annotated Sinhala tweets, two Sinhala lexicons and the UCSC Sinhala News Corpus[17] are selected as the data and the Gensim library in python is used to create the Sinhala word embedding. Stemming has not been performed on the dataset as the output has some errors. Using the above mentioned sources, Sinhala word embedding of dimensionality 200 is created. This Sinhala word embedding is used to perform Sinhala sentiment analysis in this research.

4.5 Sinhala Sentiment Analysis

Sinhala Sentiment analysis is performed in two ways in this research. They are two-way sentiment (Experiment 1) analysis and three-way sentiment analysis (Experiment 2). Two-way sentiment analysis is conducted by only considering the positive and negative tweets and three-way sentiment analysis is conducted by considering the neutral tweets also.

Sentence vectors are created using the output of the word embedding. For this mean of all the word vectors in the sentence is used.

4.5.1 Two-way Sinhala Sentiment Analysis (Experiment 1)

The dataset for this experiment is created by dropping the neutral tweets from the tweets data and combining the remaining with the two Sinhala lexicons. Then the dataset has contained 4591 negative tweets, 2944 positive tweets, 11054 negative words and 5101 positive words. From this complete dataset 2000 tweets are randomly chosen as the test set leaving 21690 sentences as the training set. Then Naïve Bayes, SVM (linear), SVM

(rbf), lightGBM, adaboost and XGBoost algorithms are trained on the mentioned dataset. For this experiment AUC value is chosen as the model evaluator since the dataset is imbalanced and AUC curves both the precision and recall metrics.

4.5.2 Three-way Sinhala Sentiment Analysis (Experiment 2)

The dataset for this experiment has contained all 3 states of annotated tweets and the data from the two Sinhala lexicons. Then the dataset has contained 4591 negative tweets, 2944 positive tweets, 2296 neutral tweets, 11054 negative words and 5101 positive words. From this complete dataset 3000 tweets are randomly chosen as the set leaving 22986 sentences as the training set. Then lightGBM, adaboost and XGBoost algorithms are trained on the mentioned dataset. Weighted F score is used as the evaluating metric in this experiment as the dataset is imbalanced and F score covers both the recall and precision.

4.6 Problems Faced in Implementation

In this section, the problems faced when creating the Sinhala dataset and the three Sinhala sentiment lexicons are discussed.

4.6.1 Problems Faced When Collecting Sinhala Dataset

Even though the use of Sinhala to express views of the public via internet has increased, there are some problems in collecting these data to perform sentiment analysis. They can be expressed as follows,

- Most people have used English letters to express their views in Sinhala rather than using Sinhala letters. For example, people have used “Eka Lassanai” instead of ”ඒක ලස්සනයි“as it is easy to type in English. This has made it difficult to collect a good amount of data in a limited time.
- Some Sinhala sentences have contained misspelled Sinhala words.
- Most of the Tweets have contained emojis and urls.
- Some Sinhala tweets have contained English words in the middle of the sentence.

4.6.2 Problems Faced When Creating Sinhala Sentiment Lexicons

The problems faced when creating the Sinhala lexicon are mainly due to the translation of English to Sinhala. The main reason for this is that, a single English word can be translated into several Sinhala words as shown by the figure 4.7. This has led to the following problem.

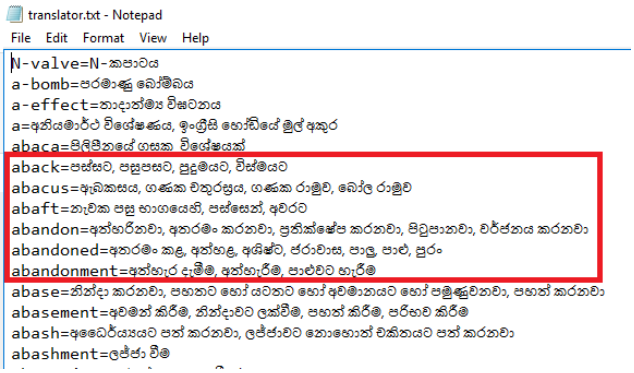

Figure 4.7 Part of "Ingiya" translator

- Some Sinhala words are contained in both the positive and negative Sinhala lexicons. For example, the word “අධික” is in both the lexicons. This is because a single Sinhala word can have several English words. An English word that could be translated to “අධික” has been in the original positive English Sentiment lexicon and another English word that could be translated to “අධික” has been in the original negative English sentiment lexicon.

Chapter 5

Results

- Two-way sentiment analysis (Experiment 1)
- Three-way sentiment analysis (Experiment 2)

In this chapter, the results obtained through performing the 2 sentiment analysis experiments on the Sinhala tweet dataset are discussed. The validation process is done by comparing the sentiment value of the manually annotated test dataset with the sentiment score of the resulting dataset created by the experiments.

5.1 Two-way Sentiment Analysis (Experiment 1)

In this experiment, as explained in the chapter four section 4.5.1, Sinhala positive and Sinhala negative tweets are used to perform the sentiment analysis and six different algorithms are chosen as the classification algorithms. The results of this experiment are shown in the following sections.

5.1.1 Two-way Sentiment Analysis using Naïve Bayes

Classification report :

	precision	recall	f1-score	support
-1	0.60	0.82	0.70	1193
1	0.43	0.20	0.27	807
accuracy			0.57	2000
macro avg	0.52	0.51	0.48	2000
weighted avg	0.53	0.57	0.52	2000

Accuracy Score : 0.5705
Area under curve : 0.5102809553041232

Figure 5.1 Naive Bayes Model Performance in experiment 1

Figure 5.2 AUC Curve for Naive Bayes

5.1.2 Two-way Sentiment Analysis using SVM (Linear)

```
Classification report :
              precision    recall  f1-score   support

          -1       0.64      0.95      0.76      1193
           1       0.71      0.20      0.31       807

    accuracy                           0.64      2000
   macro avg       0.67      0.57      0.54      2000
weighted avg       0.67      0.64      0.58      2000

Accuracy Score   :  0.6445
Area under curve :  0.5725099220878502
```

Figure 5.3 SVM (linear) Model Performance in experiment 1

Figure 5.4 AUC Curve for SVM-Linear

5.1.3 Two-way Sentiment Analysis using SVM (rbf)

```
Classification report :
              precision    recall  f1-score   support

          -1       0.60      0.99      0.75      1193
           1       0.44      0.01      0.02       807

    accuracy                           0.60      2000
   macro avg       0.52      0.50      0.38      2000
weighted avg       0.54      0.60      0.45      2000

Accuracy Score   :  0.5955
Area under curve :  0.500765514655399
```

Figure 5.5 SVM (rbf) Model Performance in experiment 1

Figure 5.6 AUC Curve for SVM-rbf

5.1.4 Two-way Sentiment Analysis using LightGBM

```
Classification report :
              precision    recall  f1-score   support

          -1       0.67      0.80      0.73      1193
           1       0.58      0.40      0.48       807

    accuracy                           0.64      2000
   macro avg       0.62      0.60      0.60      2000
weighted avg       0.63      0.64      0.63      2000

Accuracy Score   :  0.6425
Area under curve :  0.6039105646215895
```

Figure 5.7 LightGBM Model Performance in experiment 1

Figure 5.8 AUC Curve for LightGBM

5.1.5 Two-way Sentiment Analysis using XgBoost

```
Classification report :
              precision    recall  f1-score   support

          -1       0.66      0.78      0.71      1193
           1       0.55      0.41      0.47       807

    accuracy                           0.63      2000
   macro avg       0.61      0.59      0.59      2000
weighted avg       0.62      0.63      0.61      2000

Accuracy Score   :  0.627
Area under curve :  0.5913190430339725
```

Figure 5.9 XgBoost Model Performance in experiment 1

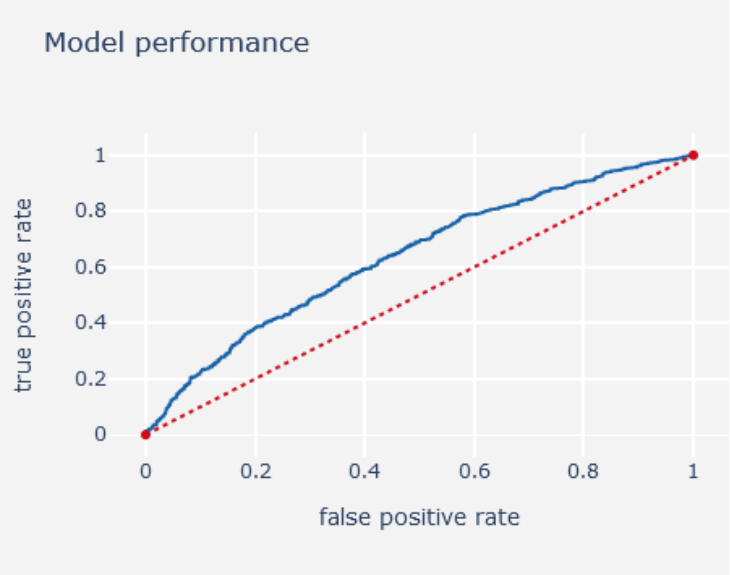

Figure 5.10 AUC Curve for XGBoost

5.1.6 Two-way Sentiment Analysis using AdaBoost

```
Classification report :
              precision    recall  f1-score   support

          -1       0.66      0.79      0.72      1193
           1       0.56      0.41      0.47       807

    accuracy                           0.63      2000
   macro avg       0.61      0.60      0.59      2000
weighted avg       0.62      0.63      0.62      2000

Accuracy Score   :  0.6325
Area under curve :  0.5959292693541737
```

Figure 5.11 Adaboost Model Performance in experiment 1

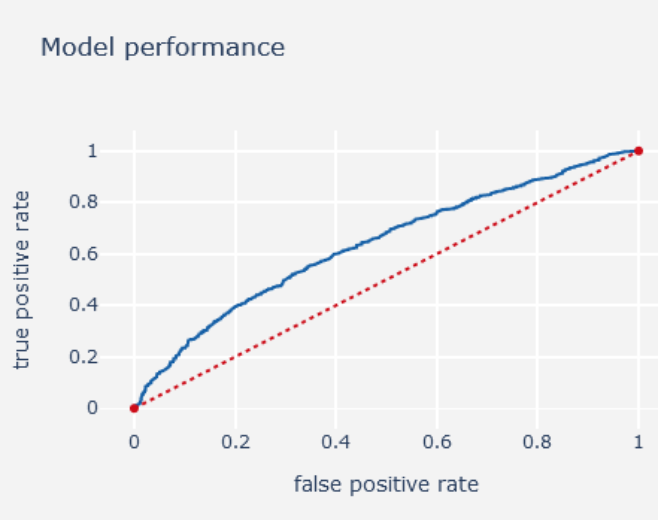

Figure 5.12 AUC Curve for AdaBoost

As shown in the figures from 5.1 to 5.12, the boosting algorithms have the highest area under the curve value. All the six algorithms have a fairly low accuracy rate but the three boosting algorithms have performed better. Among the three boosting algorithms LightGBM has slightly outperformed the other two boosting algorithms.

5.2 Three-way Sentiment Analysis (Experiment 2)

As discussed in the chapter four section 4.5.2, three-way sentiment analysis is performed by using all positive, neutral and negative Sinhala tweets and the three boosting algorithms which has shown better results in experiment 2 are used as they have shown

the ability to generalize more with the data available. The results of experiment is shown in the below sections.

5.2.1 Three-way Sentiment Analysis using LightGBM

Classification report :

	precision	recall	f1-score	support
-1	0.50	0.79	0.61	1271
0	0.54	0.13	0.21	891
1	0.40	0.37	0.39	838
accuracy			0.48	3000
macro avg	0.48	0.43	0.40	3000
weighted avg	0.48	0.48	0.43	3000

Accuracy Score : 0.47633333333333333

Figure 5.13 LightGBM Model Performance in Experiment 2

5.2.2 Three-way Sentiment Analysis using XgBoost

Classification report :

	precision	recall	f1-score	support
-1	0.49	0.77	0.60	1271
0	0.48	0.14	0.22	891
1	0.41	0.37	0.39	838
accuracy			0.47	3000
macro avg	0.46	0.43	0.40	3000
weighted avg	0.47	0.47	0.43	3000

Accuracy Score : 0.472

Figure 5.14 XgBoost Model Performance in Experiment 2

5.2.3 Three-way Sentiment Analysis using AdaBoost

```
Classification report :
              precision    recall  f1-score   support

          -1       0.48      0.72      0.57      1271
           0       0.45      0.20      0.28       891
           1       0.37      0.30      0.33       838

    accuracy                           0.45      3000
   macro avg       0.43      0.41      0.39      3000
weighted avg       0.44      0.45      0.42      3000

Accuracy Score   :  0.4483333333333333
```

Figure 5.15 AdaBoost Model Performance in Experiment 2

As shown in the figures from 5.7 to 5.9 the accuracy values are fairly low. LightGBM algorithm has the best performance out of the three boosting algorithms.

5.4 Summary

Table 5-1 Model performance summary table

Experiment	Algorithm	Actual Sentiment Score	Precision	Recall	F-Score	Accuracy	AUC
Experiment 1	Naïve Bayes	Negative	60%	82%	70%	57.1%	51.0%
		Positive	43%	20%	27%		
		Neutral	-	-	-		
	SVM (linear)	Negative	64%	95%	76	64.4%	57.2%
		Positive	71%	20%	31%		
		Neutral	-	-	-		
	SVM (rbf)	Negative	60%	99%	75%	59.5%	50.0%
		Positive	44%	1%	2%		
		Neutral	-	-	-		
	LightGBM	Negative	67%	80%	73%	64.2%	60.4%
		Positive	58%	40%	48%		
		Neutral	-	-	-		

	XgBoost	Negative	66%	78%	71%	62.7%	59.1%
		Positive	55%	41%	47%		
		Neutral	-	-	-		
	AdaBoost	Negative	66%	79%	72%	63.3%	59.6%
		Positive	56%	41%	47%		
		Neutral	-	-	-		
Experiment 2	LightGBM	Negative	50%	79%	61%	47.6%	-
		Positive	40%	37%	39%		
		Neutral	54%	13%	21%		
	XgBoost	Negative	49%	77%	60%	47.2%	-
		Positive	41%	37%	39%		
		Neutral	48%	14%	22%		
	AdaBoost	Negative	48%	72%	58%	44.8%	-
		Positive	37%	30%	33%		
		Neutral	45%	20%	28%		

The table 5.1 has summarized the results of the 2 experiments that have been carried out in this research. The overall accuracy of the 2 experiments is a bit low but all the algorithms have performed well for the negative Sinhala tweets.

Chapter 6

Discussion, Conclusion and Possible Future Work

- Discussion
- Conclusion
- Possible future work

Important findings obtained through the research and the conclusions that can be arrived after completing the research and the limitations and the difficulties faced while achieving the objectives are discussed in this chapter. Possible future works which can be done in Sinhala language in the domain of sentiment analysis are also discussed here.

6.1 Discussion

Sentiment analysis has been quite a popular topic in the last decade. It has helped governments, companies to understand the opinions of the general public and act according to it. With the advancement of internet throughout the country, the Sinhala user-generated content has increased rapidly increasing the need for Sinhala sentiment analysis.

One of the biggest problems faced while conducting sentiment analysis on Sinhala is the unavailability of the necessary resources. In this research a manually annotated Sinhala Tweet dataset with 10,000 sentences, is created. As part of the study, two Sinhala lexicons are also developed, one for Sinhala positive words and one for Sinhala negative words, along with a Sinhala word embedding.

Since Sinhala is a morphologically rich language, a single word has different word forms. In order to reduce theses word forms to a single form, stemming algorithm is also tested. The stemming accuracy for the most common 100 words in the dataset is 87% but most of the words in the word list are stop words in Sinhala. Stemming is not applied when the word embedding is created as stemming algorithm has needed further improvements.

Two sentiment analysis experiments are performed on this research. They are two-way sentiment analysis using six different classification algorithms and three-way sentiment analysis using three algorithms. When the six algorithms in two-way sentiment analysis are compared, the experiment with SVM (linear) algorithm has a bit higher accuracy but the performance on the minority class (positive tweets) is less compared to the experiment with the three boosting algorithms. The three boosting algorithms have had the better AUC value among the six algorithms and LightGBM algorithm has the best

AUC. The reason for this can be that Boosting algorithms can generalize more with the minority class than the SVM algorithm. Because of this reason the three-way sentiment analysis is performed using the three boosting algorithms.

When the three algorithms in the three-way sentiment analysis are compared, LightGBM has got the highest accuracy and XgBoost has also got a very similar level of performance but Adaboost has a slightly lower performance.

Overall, the two experiments have yielded somewhat low accuracy rates. The reason for this low accuracy rates is the drastic label imbalance that existed in the dataset and therefore the models has got less amount of data to train for positive and neutral classes. This can be overcome by collecting more and more data, so that there are a higher number of positive and neutral tweets to train the algorithm. The label imbalance will still be there as that is the normal behavior of the human being, but with more data for positive and neutral classes available the algorithms can perform better. Availability of words that are in both the Sinhala positive lexicon and the Sinhala negative lexicon can also have an impact on the final accuracy. This problems can be overcome by creating a Sinhala lexicon manually using the Sinhala words in a dictionary and manually annotating the words.

6.2 Conclusions

The conclusions that can be arrived after the research are stated below.

- Performing sentiment analysis for a less resource language such as Sinhala can be difficult.
- Even though the Sinhala user-generated web content has increased, it takes some time to collect a sizable dataset suitable to perform sentiment analysis.
- Some sentences can be harder to annotate even for a human being.
- Stemming algorithm tested in the research should be further improved.
- Boosting algorithms perform much better when there is a class imbalance.

6.3 Possible Future Work

The following areas can be considered for the future work carried out in this field of Sinhala sentiment analysis.

- Improving the Sinhala dataset by adding more manually annotated sentences.

- Improving the two Sinhala lexicons by correcting the problems stated in the chapter four.

- Improving the stemming algorithm

- Deep learning techniques such as LSTM can be used with more annotated data collected.

References

[1] L. Zhang and B. Liu, "Sentiment Analysis and Opinion Mining," *Encycl. Mach. Learn. Data Min.*, no. May, pp. 1–10, 2016.

[2] A. Agarwal, K. R. Mckeown, and F. Biadsy, "Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams," *Proc. 12th Conf. Eur. Chapter Assosciation Comput. Linguist. EACL 09*, no. April, pp. 24–32, 2009.

[3] P. Arora, "Sentiment Analysis For Hindi Language," no. April, pp. 1–63, 2013.

[4] N. Medagoda, S. Shanmuganathan, and J. Whalley, "Sentiment lexicon construction using SentiWordNet 3.0," *Proc. - Int. Conf. Nat. Comput.*, vol. 2016–Janua, no. May 2016, pp. 802–807, 2016.

[5] N. Medagoda, S. Shanmuganathan, and J. Whalley, "A comparative analysis of opinion mining and sentiment classification in non-english languages," *Int. Conf. Adv. ICT Emerg. Reg. ICTer 2013 - Conf. Proc.*, pp. 144–148, 2013.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings ofHuman Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005, pp. 347–354.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," *Proc. Work. Lang. Soc. Media*, no. June, pp. 30–38, 2011.

[8] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Proc. Lr.*, pp. 1320–1326, 2010.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Empir. Methods Nat. Lang. Process.*, vol. 10,

no. July, pp. 79–86, 2002.

[10] C. N. dos Santos and M. Gatti, “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts,” *Coling-2014*, pp. 69–78, 2014.

[11] X. Glorot, A. Bordes, and Y. Bengio, “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach,” *Proc. 28th Int. Conf. Mach. Learn.*, no. 1, pp. 513–520, 2011.

[12] I. U. Liyanage and S. Ranathunga, “Sentiment Analysis of news comments,” 2018.

[13] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” *KDD*, pp. 168–177, 2004.

[14] Language Technology Research Lab University of Colombo School of Computing, “Ingiya English-Sinhala dictionary database,” 2016. [Online]. Available: http://ltrl.ucsc.lk. [Accessed: 15-Jan-2020].

[15] N. Medagoda and S. Shanmuganathan, “Keywords based temporal sentiment analysis,” *2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2015*, no. May, pp. 1418–1425, 2016.

[16] Language Technology Research Lab University of Colombo School of Computing, “Suffixes-of-Sinhala,” 2016. [Online]. Available: http://ltrl.ucsc.lk. [Accessed: 04-Feb-2020].

[17] Language Technology Research Lab University of Colombo School of Computing, “UCSC Sinhala News Corpus,” *2016*. [Online]. Available: http://ltrl.ucsc.lk. [Accessed: 07-Feb-2020].