Sentiment Analysis of Tweets to predict Sri Lankan Election Results using Supervised Learning Techniques

A Dissertation Submitted for the Degree of Master of Computer Science

W.M.H.D Gunasiri

University of Colombo School of Computing

2021

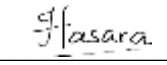
DECLARATION

I hereby declare that the thesis is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: W.M.H.D Gunasiri

Registration Number: 2017/MCS/033

Index Number: 17440331



Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. /Ms. W.M.H.D Gunasiri under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. M. G. N. A. S. Fernando

29/11/2021

Signature of the Supervisor & Date

I would like to dedicate this thesis to my husband….

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. M. G. N. A. S. Fernando for helping me throughout the research work. He guided me in the right direction by providing his valuable assistance.

I am grateful to all the researchers who made available their research outcomes which I have used as a source of my research.

My special thank goes to my parents and parents — in — laws for their patience and support during this research period.

Finally, I would like to express my heartful gratitude to very special person, my amazing husband for his continues support, patience and understanding during this period. I really appreciate the sacrifice he has made during this study.

ABSTRACT

Due to the current pandemic situation in the world, people are spending their leisure time mostly on social media platforms. They more tend to express their selves openly when they are behind the keyboard. This user behavior has created a huge advantage for researchers and analyzers to analyze people's opinions, behaviors and predict certain outcomes.

This research study is used to get the best out of aforesaid user behavior and conduct the prediction-based analysis using the Twitter — social media platform. When we consider the election prediction using sentiment analysis, there were many researches done based on the languages English, Chinese, Arabic, Hindi etc. But this is a novel application area for Sinhala language(de Silva, 2020). Even though there are several studies available for Sinhala language, they cannot be directly used for Election prediction since sentiment analysis is highly application dependent. Applications which develop for one domain cannot be used for another domain in sentiment analysis. And another issue is, same methodologies and technologies which are used for other languages, cannot be directly used for Sinhala language due to language differences. So, the focus here is to create a domain specific research for the area of Election prediction and to introduce new resources to the text analysis community which will be helpful for their further studies.

In this research, prediction — based system was developed using Sinhala tweets. Automatic labelling was used to predict the election results for each candidate. These predicted results were compared with the actual presidential election results in Sri Lanka – 2019. Suitable model was developed by applying text preprocessing and feature extraction techniques. Supervised learning classifiers were trained against the developed model to find the best classifiers for predictive sentiment analysis in Sinhala language.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 - INTRODUCTION

In this pandemic situation, millions of people got stuck at home due to multiple lockdowns in all over the world, yet they have to continue their day — to — day lifestyles by adapting to the current situation. Most of the tasks has to be completed online and their leisure time has also increased due to the time savings specially from transport and etc. As a matter of fact, people lean towards to online platforms and they are utilizing their leisure time mostly on social media platforms like Facebook, Twitter, Instagram, etc.

The number of mobile connections has been increased by 612,000 in Sri Lanka between 2020 and 2021 period (Kemp, 2020). This is mostly due to the social media usage since people are more attractive towards the social media platforms. As per the statistics in Sri Lanka, (Kemp, 2020) the number of social media users have been increased by 1.5 million between 2020 and 2021, and in a percentage wise this is a 23% increase. This implies how the social media usage increases with the current lockdown in this country. People more tend to express their selves openly in social media through several type of platforms. For an example one will expose their day — to — day activities through “vlogs”, some will post their likes/dislike on certain things, some will express their political opinions on Facebook, Twitter, or similar platforms without even not considering the privacy limits, as they all are heroes “Behind the keyboard”.

There are several privacy threats involved with the rapid increase of social media usage, but this user behavior has also created a huge advantage for researchers, analyzers and other interested parties, to study and analyze about people’s opinions, behaviors and predict certain outcomes. Sentiment analysis or opinion mining becomes a key area when it comes to analyzing the user thoughts and ideas in social media. The main purpose of this research is to get the best out of aforesaid user behavior and conduct the prediction — based analysis using social media platforms.

Sentiment analysis is useful in multiple aspects such as track feedbacks, provide personalized services, brand monitoring, and predict behaviors. Among these advantages, predict user behavior is the key advantage of Sentiment analysis. Users’ behaviors can be predicted in various areas such as in tourism, stock market, election etc. Out of above areas, Election prediction will be considered for this study.

1.1 Problem

When it comes to internet users, their internet usage can be categorized into different categories such as social networking, electronic business, entertainment, telecommuting, crowdsourcing, collaborative publishing etc.

Among these multiple categories, social networking plays a huge role which allows users to be socialize and to interact with others. As of today, social networking has covered 3.96 billion users worldwide ("Digital 2020: Sri Lanka — DataReportal – Global Digital Insights," n.d.). Nowadays, people are addicted to social network platforms such as Facebook, twitter, blogs, wikis, etc. due to many attractive features in these environments. Most of the people in this platform are bound by these attractive features, so that they are not certain about the content that they are sharing and there is no gap between their public and private lives. They are sharing their experiences, opinions, complaints, achievements, knowledge, suggestions, and many by using this platform. Due to this behavior of the users, they have created a large data pool of their behaviors unintentionally.

This data pool provides information to travelers/hotel owners (hotels/restaurants reviews/comments/ratings), consumers/product owners/service providers (product reviews/comments/ratings), researchers, politicians, companies, and different kind of online users for different purposes ("Everything There Is to Know about Sentiment Analysis," n.d.). But the usage of this information is difficult due to heavy online data load with lots of unnecessary information. These data need to be analyzed to cater only the relevant information. There are many research areas which builds up for the above requirements and one such main area is Sentiment analysis. Even though sentiment analysis has become a trending area, most of the researches are focused only on the linguistic rich languages such as English language. Hence this creates a variety of sentiment analysis tools for English language. But there are no sufficient tools for Sentiment analysis which could be used for Sinhala language. There is a necessity to identify more tools which are more suitable for Sinhala language for the area of Sentiment analysis.

Nowadays, due to this pandemic situation, people are spending more and more hours in internet ("Global Digital Overview — DataReportal – Global Digital Insights," n.d.). Their internet usage has increased due to the multiple lockdowns in the world and work from home situations. During the election period also, people are using social media platform frequently and updating their opinions. They will express their opinions truthfully when their identity is anonymous.

Most of them tend to create multiple accounts and share their thoughts through cyberspace. Analyzing this information will be useful to politicians and to the society for further actions. So, there is a necessity to predict the user behavior, based on the analyzed information. During this study, the user's opinions will be analyzed based on different criteria and the output will be provided as the analysis for the presidential Election results in Sri Lanka, which is the novel research application area in Sinhala language (de Silva, 2020).

1.2 Motivation

As a country, which uses the Sinhala language as a mother tongue, it will be very important to identify characteristics and features of Sinhala context since the amount of sentiment analysis tools specific for Sinhala language is lesser compared to English language. Also, it is difficult to construct sentiment analysis tools for Sinhala Language from the existing analysis tools as they are more oriented towards the English language and it will create language specific issues. One of the purposes of this research is to focus on finding the sentiment analysis tools which are more suitable in Natural language processing for Sinhala language.

Sentiments are domain specific and application dependent. Sentiments in reviews systems (පිරිසිදු, කාරුනික, ස්ථානය, දිවා ආහාරය) are different from sentiments in product analysis (කල් පවතින, බැටරිය, තත්වය, මිල)). Due to the aforesaid behavior, the application, which is developed for one domain, will not be suitable for another. This will create domain specific sentiment research areas. Therefore, another focus of this research is to conduct a domain specific research for the area of prediction and trend analysis to predict the election results in Sri Lanka.

1.3 Problem domain

This research is based on different areas in computer science such as Machine learning and Natural language processing. During this study, above main areas are further divided into sub areas such as Supervised learning and Sentiment analysis. Below section will cover these few topics to get an idea about the domain of this study.

1.3.1 Machine Learning (ML)

Machine learning is a data analysis method, and it is a subset of Artificial Intelligence (AI). In Machine learning, machines learn from data, identify hidden patterns, and make predictions (Brownlee, 2016). Today, Machine learning plays a huge role in areas such as search engines, product personalized recommendations, Social media services, customer supports etc. Some of

the example services are YouTube, Google, Facebook, and Twitter which use machine learning to predict the user behavior. During a machine learning process, there are basically four steps, create a dataset, preprocessing/clean data, train, and test model. These basic steps will be covered during this research.

Machine learning methods can be divided into three categories such as Supervised learning, Unsupervised learning, and semi — supervised learning (Brownlee, 2016).

Supervised learning

Supervised learning uses labelled dataset to train the model. During this method, set of input and output data are defined and the predictions are made based on the labelled dataset and the learnings during the training process. Supervised learning can be divided in to two categories, Classification and Regression (IBM Cloud Education, 2020). Classification is used to predict the discrete(category) values and Regression is used to predict the continues(numerical) values. There are different types of classification algorithms such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, Decision tree etc. Simple linear regression, Multiple linear regression, Decision tree regression, Random forest regression are examples to the types of regression algorithms (Wilson, 2019). There are multiple usages in Supervised learning such as fraud detection, trend analysis, automation etc. Supervised learning algorithms will be used during this study as a sentiment analysis technique.

1.3.2 Natural language processing (NLP)

There are around 6500 human languages in the world. Natural language processing is the ability of computers to read, understand, analyze, interpret these human languages. Natural language processing supports the interaction with the computers and human languages (Sharma, 2020). As a subset of Artificial Intelligence (AI), computers are programmed to analyze and derive meanings from large number of human languages. There are two main techniques in Natural language processing, Syntactic analysis and Semantic analysis (Garbade, 2018). During the syntactic analysis, grammar of the sentence is considered and do the analysis. During the semantic analysis, the meaning of the sentence is considered for the analysis. In today, there are many applications which uses the natural language processing. Speech recognition, Personal assistant applications, language translations, chatbots, search engines are some of the few examples which uses the natural language processing.

Sentiment analysis

Sentiment analysis is also called as opinion mining or emotion extraction which is used to extract data/emotions/opinions from the written content. This is a text analysis and a Natural language processing (NLP) technique ("Everything There Is to Know about Sentiment Analysis," n.d.). This has become one of the fastest growing research areas today. This helps the researchers to analyze the users written contents, find opinions in their text and understand hidden information such as their likes, dislikes, opinions, and expressions thoroughly. Sentiment analysis is proven to be the best method for evaluating humans' emotions and opinions in many application areas such as business, decision making, financial analysis, predictions analysis and trend analysis (Gupta, 2018). Considerable amount of work has been conducted for this area recently and the presidential election prediction will be based on the Sentiment analysis.

1.4 Research contribution

1.4.1 Goal

The goal of this research is to predict the winner party of an election, based on the Sinhala tweets in the period of 2019 presidential election. This will analyze the user opinions on candidates and categorize them according to the users positive, negative, and neutral feedback. This model will be used to provide the visualization output for each of the candidate and predict the majority vote in election.

1.4.2 Objectives of the study

During the Election period in most countries, almost all political parties are working with different kind of Intelligent services to get a rough idea about the outcome of the election. Based on those predictions, politicians tend to alter their approaches towards the election. Most of this information is gathered by based on the people's behaviors, such as ideas they share on their office, in public places, with communities etc. But to be speaking frankly, most of the people do not like to share their true opinions with others due to the post — election situations in countries like ours. Because no one likes to be the supporter of the losing party after the election. But when the people are behind the keyboard (especially when their identity is kept anonymous), they are more likely to express their thoughts genuinely, especially for topics like election. This behavior can be used to get a more accurate prediction rather than sticking into old — fashioned intel gathering method.

As per the statistics for languages in Sri Lanka ("Sri Lanka Demographics Profile," n.d.), Sinhala language is the widely used language which is spoken by the 87% population in Sri Lanka. Therefore, election related twitter — based data in Sinhala language will be identified, analyzed, and predicted the output based on supervised learning algorithms.

Limited number of researches were conducted for Sentiment analysis of Sinhala language (Amali and Jayalal, 2020; Chathuranga et al., 2019; Demotte et al., 2020; Jayasuriya et al., 2020; R. Jenarthanan et al., 2019; Medagoda et al., 2015; R. Jenarthanan et al., 2019; Chathuranga et al., 2019) and it is revealed that this particular area related to prediction analysis in politics is not covered by the existing sentiment analysis researches in Sinhala.

The objectives of this study are listed below.

- Extract Sinhala tweets which are related to the Sri Lankan presidential election – 2019
- Analyze and use of Sentiment analysis techniques and text preprocessing techniques for Sinhala language for the purpose of sentiment identification and unnecessary data removal
- Develop automatic labelling approach to predict the presidential Election results
- Develop a suitable model and train it to find the best classifiers for predictive sentiment analysis in Sinhala language
- Evaluate text analysis results with the use of different classifiers to determine the accuracy of the proposed model
- Visualization of presidential election result prediction with use of different perspectives and conduct a comparison study with the actual presidential election results in Sri Lanka

1.5 Scope

The main extent of this venture is to analyze and predict the human's behavior for the Presidential election using the Machine learning algorithms such as Naïve Bayes, Support vector machine, decision trees etc. These results will be evaluated based on the past presidential election results and by comparing the different sentiment analysis techniques in Sinhala language. This research will be conducted based on the assumption that the Sinhala comments in twitter is written only by Sri Lankans.

There are different steps in sentiment analysis phase such as data collection, preprocessing, sentiment words detection, sentiment classification and output presentation ("Everything There

Is to Know about Sentiment Analysis," n.d.). This research will cover all the mentioned phases and present the output to the user by polarity level (positive, negative, and neutral) and by a graphical representation. Feature extraction phase can be divided further into two categories such as language dependent (POS tagging, Negation) and language independent (Bag of words model, n — gram, term frequency) features. This research will be focused mostly on language independent features due to the limitation of the availability of Sinhala language related tools.

Dataset will be obtained from twitter, based on election — based opinions during the year — 2019. It is supposed to obtain the sufficient dataset for training and testing in order to facilitate better classification results. This research will be conducted based on the assumption that; this model is valid only for pure Sinhala texts. Then the Sinhala tweets are extracted accordingly. During the data text analysis, it was identified that the data set consists of comments as well as news. This news related sentences were also included in the dataset due to the time limitation for text mining. If this research study focuses on text mining, objective of this research will be narrowed down to text mining area. So, currently the main focus was to use the manual method to remove the news from the comment section. When the dataset was investigated manually, it was also identified that some sentences have indirect meanings as per the below example sentences.

රටට වසන්තය උදාකරන්නැයි ජනතාව දෙවියන් යදින්නේ නැත. ජනතාව ආයාචනා කරන්නේ ගෝඨාභය රාජපක්ෂ ට රට බාර ගන්නා ලෙසයි.

හැමදාම මැව් පරදිනවා නම් වෙන කෙනකුට බැට් කරන්න අවස්ථාව දිය යුතු බවත් සජිත්, නවීන් වැනි පිරිසකට අවස්ථාව දිය යුතුව ඇති බවත් ඕනෑම භාණ්ඩයක් හෝ පුද්ගලයකු කල් ඉකුත් වන බවද දේශපාලන පක්ෂ නායකයන්ටද එය වලංගු බවද මාධ්‍ය හමුවකදී සඳහන් කළේය.

These types of statements can only be identified in the manual investigation methods; hence it will not be possible to use automatic labeling for these type of sentences as it will not output the hidden meaning. In this study, it was assumed that the extracted data set consists only direct sentences, not indirect sentences.

Sentiment analysis techniques can be categorized in to two main techniques; Lexicon based and Machine learning as per the below figure 1. Machine learning algorithms are further divided into two categories as Supervised and Unsupervised learning. As per the reference ("Supervised vs Unsupervised Learning: Key Differences," n.d.), supervised learning algorithms are highly accurate, while unsupervised learning algorithms are less accurate. According to Ingedata (Ingedata, n.d.), "Classification, categorization, problem solving supervised algorithms are still kings of their realms". As mentioned in the previous sections, Sentiment analysis is domain

specific as well as application dependent. For example, application develop for marketing will not be suitable to use as an application for election prediction. Machine learning algorithms which will be used for sentiment analysis should provide the domain specific approach rather than providing generalized model. Hence, the focus will be to use supervised learning algorithms with the labeled training data. Since this is the domain specific research in Sinhala language, it would be convenient to use the labeled data for text classification.

The focus of this research will be based on supervised learning approaches with the below highlighted classifiers.

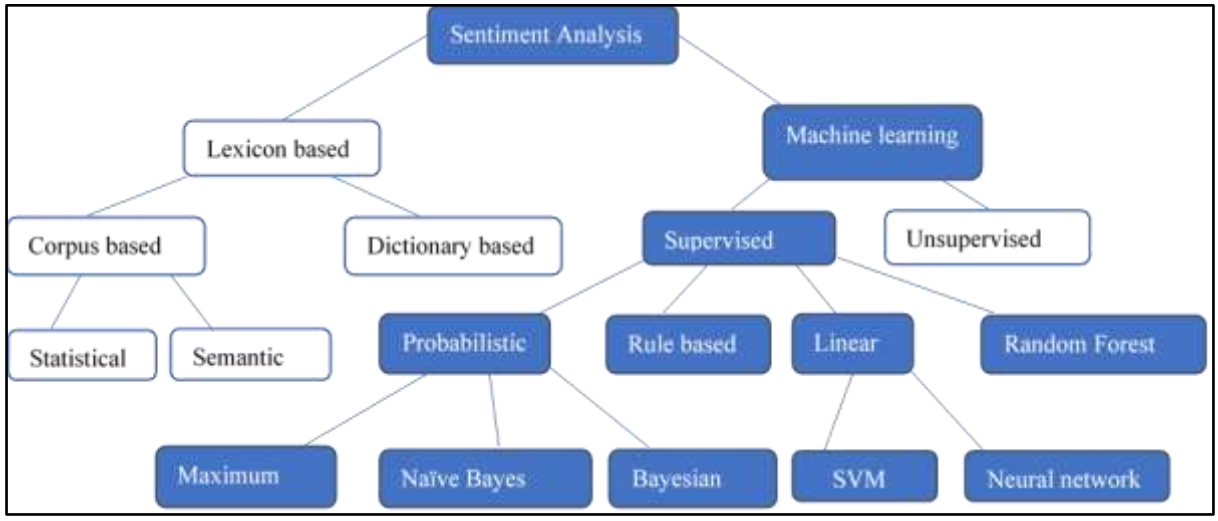


Figure 1. Available Sentiment analysis techniques ("Figure 2 Sentiment classification techniques.," n.d.)

1.6 Structure of the thesis

Chapter 2 of this thesis demonstrates existing researches related to the sentiment analysis and machine learning algorithms. Chapter 3 describes the methodology which used to predict the election result behavior by using visual techniques. Chapter 4 is about the evaluation process. Future work and implementations related to this thesis will be described in Chapter 5.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

As mentioned in Chapter 1, this research is based on different areas in Computer science such as sentiment analysis, supervised learning etc. Users will be able to get thorough knowledge about Sentiment Analysis and identify its areas / gaps by reviewing the literature review section. And it would be beneficial for the user to know the background of the study when it comes to the Chapter 3, methodology section.

Nowadays, Sentiment analysis has become the most trending area in Natural language processing. Many researches have been conducted to analyze the sentiment contents and most of them are based on English language. There are some limited researches carried out for Sinhala language as well. In this section, Sinhala and other language related researches are explained briefly.

2.1.1 Literature related to sentiment analysis and sentiment analysis techniques in Sinhala language

Medagoda N., Shanmuganathan S., Whalley J. have proposed an algorithm (Rajenthiran Jenarthanan et al., 2019) for constructing Sentiment Lexicon for Sinhala language. This was claimed as the first attempt for generating Sinhala sentiment lexicon even though there are different language — based sentiment lexicons. English sentiment lexicon (SentiWordNet 3.0) was used as a baseline for this study. Sinhala word dictionary and English word dictionary were mapped with the words by using translations and assigned sentiment value for a Sinhala word and its synonyms based on an English word sentiment value. Experiment was carried out for the created Sinhala lexicon using 2,083 articles collected from online newspapers. These articles were categorized into three categories such as positive, negative, and neutral.

In the first experiment, Naïve Bayes, J48(Decision Trees) and SVM (Support Vector Machine) classifiers performance was measured for all three sentiment levels: positive, negative and neutral. This results the highest accuracy of 48% which is less than the benchmark values for English and other Asian languages.

In the second experiment, binary classification was conducted with only two sentiment levels: positive and negative. This resulted in 16% accuracy improvement for all classifiers. Among these three classifiers, Naïve Bayes provides the highest accuracy with up to 60%. This research was done based on multiple assumptions such as:

- Sinhala and English word senses are same
- Part of Speech (POS) tagging and sentiment score for both languages are same.

The above assumptions are arguable because of the linguistic differences between multiple languages. The results of this algorithm were not publicly available. In this research, the publicly available sentiment lexicon – "helasentilex" (Karunanayake, n.d.) will be used.

Jayasuriya P., Ekanayake S., Munasinghe R., Kumarasinghe B., Weerasinghe I., Thelijjagoda S. (Jayasuriya et al., 2020) have performed Sentiment classification for Social media – YouTube contents in Sinhala Language. The focus of this research was to classify the domain based(sports) social media content in to positive and negative polarities by using machine learning algorithms, lexicon based and hybrid approaches.

YouTube video comments related to specific domain(sports) were considered during this methodology. 2210 comments were collected from YouTube and grouped them in to positive and negative categories. Data preprocessing was carried out by removing unnecessary characters and stop words. Stop word list was taken from the customized stop word list which has already defined for Sports domain. These stop words and characters were identified and removed them from the comment list. Sentiment analysis was conducted by using three classifier methods: Machine learning based classifiers, Lexicon based classifiers and Ensemble classifiers.

In machine learning based classification, Naïve Bayes, Logistic Regression and Support vector machine classifiers were used with the unigram, bigram, and trigram feature extraction methods. Each word in comments were scored based on the polarity value and the sentiment value of the comment was calculated based on the total sentiment score in lexicon — based classification. During the sentiment analysis using Ensemble classifier, Machine learning based and lexicon — based classifiers were combined using a majority voting option. Accuracy and F1 — Score were calculated and used as the evaluation metrics. When comparing machine learning and lexicon — based approaches individually, machine learning approach provides more accurate results. But Hybrid approach is more accurate when compared with the three approaches for social media sentiment classification.

Amali H.M.A.I and Jayalal S. have proposed a method (Amali and Jayalal, 2020) to classify cyberbullying comments in Sinhala language for Social media contents. Data was gathered using tweepy (a python library) which is used to access the Twitter REST API. Since the Standard Twitter API supports data up to 7 days only, the python program was developed to

gather data for more than 7 days. Yourswear.com has suggested an offensive word list for Sinhala language. This list was used to extract data from twitter comments and will be saved as a CSV file. Total 652 records were extracted. 5 rules were used to identify the cyberbullying words.

- Percentage of the offensive words (If the percentage is higher than the 10%, that comment is considered as a bullying comment)
- Types of combination of the pronoun and the offensive words

By using above rules, twitter comments were labelled into four categories as very cyberbullying, cyberbullying, non — cyberbullying and very non — cyberbullying. Manual labelling process was used to label these content using crowdsourcing. And the next step was to conduct the data preprocessing. Removal of unnecessary characters, outliers and stop words, was performed at this stage. Feature extraction was carried out with the help of already defined rules. As the final step, classification was done with the 70% and 30% data percentage ratios for training and testing. Three classifiers (SVM, KNN, Naïve Bayes) were used to train and test the model and observed that the SVM with RBF kernel provides the highest F1 — score with the 91% value.

Liyanage I.U (Iu, 2018) has performed a sentiment analysis of Sinhala news comments. Data was collected from the comment section in online newspapers. Logistic regression, Decision tree, Naïve Bayes, SVM and Random Forest classifiers were used for the experiment. Out of these five classifiers, Logistic regression provides better results with the highest accuracy. Another performance measurement was done by removing punctuation marks. This increases the classifiers performance. Feature extraction methods such as n — grams, TF — IDF were used and tested with these five classifiers. TF — IDF method outperformed other extraction methods. Another testing phase was carried out to test the effectiveness of word embedding features such as Word2Vec and Bag of word model. Word2Vec outperformed other word embedding features.

Jenarthanan R, Senarath Y, Thayasivam U have proposed (R. Jenarthanan et al., 2019) an annotated corpus for Tamil and Sinhala sentiment analysis which is abbreviated as ACTSEA. This research was based on the twitter data. They have classified the emotions into 6 categories such as anger, fear, joy, sadness, surprise and disgust. Keywords of these categories were identified by using the help of linguistic professionals for two languages: Sinhala and Tamil. Year — 2018 tweets were extracted month wise, with the total of 200 tweets for Sinhala and 300 tweets for Tamil.

As the first step, preprocessing was carried out to remove unnecessary characters such as URL, spaces, hash tags etc. And also, non — Sinhala and non — Tamil sentences were removed from the dataset to eliminate the unusable data. These collected data were annotated using the Tamil and Sinhala annotators. They have evaluated the created corpus using the categories objective tweet, correctly classified, misclassified and not classified. If the data does not have any sentiment value, it was categorized as "objective tweets". If the tweets were categorized into correct category, they were treated as "correctly classified" tweets. If the tweets were not categorized into correct category, they were treated as the "misclassified" tweets. If the annotators cannot judge the tweets category, it was fallen into "not applicable" class. Reliability of the annotation process was measured by using the Cohens Kappa value. For each category Cohens Kappa value was measured and it was observed that most of the values were between 0.6 — 0.8 range which implies the reliability of the proposed corpus.

2.1.2 Literature related to the election prediction

Moh T.S., Sharma P. (Sharma and Moh, 2016) have attempted a sentiment analysis for Indian Election using Hindi twitter. Relevant Hindi language tweets were collected using twitter achiever. Tweet's data was searched with the keyword '#politicalpartyname'. Election related data was extracted from twitter. Preprocessing was conducted by removing website URLs, hash tags as the first step. Data classification was done based on the polarity levels positive, negative, and neutral. Three types of classifiers were used in this study. They have used both supervised and unsupervised learning approaches. SVM, Naïve Bayes and Dictionary tree algorithms were used for analysis and dataset was classified as positive, negative and neutral. For Dictionary based approach, 23,998 data were used. Both SVM and Naïve bayes approaches, 42,345 data were extracted, and 36,465 data were remaining after preprocessing. 5 — fold cross validation was done after the manual labelling. Data were categorized into training and testing levels and repeated the process for 5 times and average accuracy was measured. It was found that the accuracy of SVM classifier is higher with the 78.4% value than other classifiers and the election prediction was made based on the SVM classifier results.

Kuman P., Gupta Y. (Gupta and Kumar, 2019) have proposed a real time sentiment analysis approach for Punjab election — 2017 results. Twitter based data (1573) was collected using Twitter API and extracted to CSV file in real time. Punjab election was held on 11th March 2017 and the data was collected before the election which is between 13th Jan 2017 – 06th Feb 2017. Different keywords: Party names and their leaders (AAP, Congress, SAD — BJP), were used to collect the data related to the election. Labelled data was needed to train the model

during this study. Data could be manually labelled or the labelled data from different sources could be used instead. For this research study, labelled data were collected using online GitHub directory. Data was categorized into two columns: text and sentiment level with the values Positive, Negative and Neutral. As the second step, collected data was preprocessed by using below methods.

- Remove unnecessary characters (punctuation marks, Hash tags (#), URLs)
- Remove stop words
- Case lowering

As the next step, feature selection was carried out with below methods

- Unigrams
- Unigrams and bigrams
- Lemmatization
- POS tagging
- Punctuations were taken as separate unigrams
- Information gain

Model was trained by dividing the data set in to training and testing with the ratio of 70% and 30% respectively. Five machine learning classifiers (Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Maximum Entropy and Support Vector Classifier) and Three deep learning models (3 — layer Perceptron, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN)) were used in this process.

When comparing the accuracy of machine learning classifiers, Bernoulli Naïve Bayes produced the best accuracy and SVM performed as the worst accuracy classifier. F1 — score, precision and recall are measured, and Bernoulli Naïve Bayes classifier has the best performance among these metrics. And the performance was measured for deep learning models with the multiple iterations (20, 25, 30). 3 — layer perceptron was more accurate among the other deep learning models. Twitter based data was observed and the dashboard was presented as an output for the user with the real time updates by using the developed models. These results were finally compared with the actual Punjab election results and find out the system provided the results which are same as for the actual results.

Begum S.H., Nausheen F. (Nausheen and Begum, 2018) have used lexicon — based sentiment analyzer to predict US presidential – 2016 election results. Tweets related to Trump, Hilary and Bernie were collected using Twython – python library. Python code was used to collect tweets

and they were classified into three sentiment values: positive, negative and neutral. Sentiment value for each comment was observed and the sentence level sentiment was calculated by finding the total sentiment value. Data preprocessing was conducted by removing unnecessary information such as hashtag, URL, emoticons, whitespace and newline characters. Average polarity and subjectivity were measured as the performance metrics. Each of the candidate's evaluation was presented using the graphical representation, word cloud and table. The outcome was Hilary had more positive comments than the other candidates. Average subjectivity of Bernie was better than Hilary and Trump.

Chooralil V.S and Jose R. (Jose and Chooralil, 2016) have suggested a classifier ensemble method for sentiment analysis. Real time twitter data was used for this study to predict Delhi election results. Real time twitter data — 12000 was extracted by using Twitter Streaming API. During the data preprocessing, unnecessary data such as hash tag, URL, @ were removed. In this study, negation handling was conducted by using state variables and bootstrapping. Sentiment classification was conducted using SentiWordNet, Naïve Bayes and hidden markov model classifiers and then the ensemble approach was used for this.

As per the results, the accuracy of Ensemble approach is more significant than other individual methods. Comparison was done for the two politicians with the extracted data and results were shown in a graphical interface. And the same model was applied to conduct the comparison for newly released movies based on the twitter data.

Joseph F.J.J (John Joseph, 2019) has used Decision tree — based approach to predict the Indian General election results in 2019. Twitter based data was extracted using the Tweepy library. These tweets were extracted every day for a particular period and stored in MongoDB by using pymongo library. 5000 tweets were collected for ruling and the opposition party. As the first step, preprocessing was conducted to remove regular expressions, emoticons, stop words and punctuation marks. All the non — English words were removed. In this study, Decision tree classifier was used to predict the sentiment values. Polarity and subjectivity values were measured, and each tweet was classified as positive, negative and neutral.

Popularity = ((0 x Negative tweets) + (Neutral tweets / 2) + Positive tweets)/Total tweets

Ruling and opposition parties' popularities were calculated based on the number of seats for each party and this was tested with the actual results. And this produces results with 97% accuracy.

2.2 Research Gap

According to the literature survey, most of the existing research related to election prediction, were conducted using English, Hindi, Chinese languages but none of the researches has been carried out in Sinhala language. So, there is a necessity to conduct a prediction — based sentiment analysis for elections using Sinhala language. This is a novel research application area in Sinhala language (de Silva, 2020).

Same methodologies and technologies for other languages cannot be used for this research and need to adapt them for Sinhala language. For example, English language related stems cannot be used directly for Sinhala language. Sinhala language supports totally different kind of stems based on the language features. As De Silva(de Silva, 2020) states, Sinhala language is a resource poor language. Hence another purpose of this research is to find the sentiment analysis tools which are more suitable in Natural language processing in Sinhala language.

CHAPTER 3 - METHODOLOGY

3.1 Introduction

As described in previous chapters, the main goal of this research is to predict the presidential election results based on the twitter related data. In this chapter, the strategy used for each process will be explained thoroughly. This chapter provides information to the readers about the data collection method, sentiment analysis techniques, how to train the model etc.

Tools/software used during this research will also be discussed under this topic. Following subproblems will be further described in this chapter.

- How to conduct data collection
- How to perform automatic/manual dataset labelling
- What are the preprocessing techniques and feature extraction methods?
- How to construct a model

3.1.1 Dataset collection

During the data collection stage, presidential election (2019) related Sinhala tweets will be collected from twitter using the twitter scraper tool (ScrapeHero, 2018). The Standard twitter API provides the facility to retrieve tweets up to last 7 days only ("Overview," n.d.). Since the main objective of this research is to collect data from the previous presidential election, a 3rd party tool will be used. This tool provides an option to extract data to a csv file with a specified timeframe and keywords.

Initial plan of this research is to collect public Sinhala tweets (year — 2019) from twitter as much as possible to widen the domain space. Data Sample will be divided into training and testing dataset and they will be split in to 70% and 30% respectively. This percentage could be varied based on the user requirement. As per Brownlee (Brownlee, 2020), it is better to use the training and testing data with the ratio of 90% and 10% respectively for larger data. In this research, the expected data would be moderate since the language is Sinhala. Accordingly, the data will be split as 70% and 30%.

Data was collected from Twitter scraper tool using the keywords 'ගෝඨාභය' and 'සජිත්' for a specific date range. Standard twitter API provides data extraction up to last 7 days only. This twitter scraper tool supports the data retrieval for any given period with the specific keywords and the specific language as shown in below figure 2.

3928 and 3893 data records were extracted from the specific time period related to the keywords 'සජිත්' and 'ගෝඨාභය' as shown in below figures (figure 3 & 4).

Time period –

https://twitter.com/search?l=&q=සජිත්%20since%3A**2019-01-01%20until%3A2020-01-01**&src=typd&lang=sin

Keywords –

https://twitter.com/search?l=&q=සජිත්%20since%3A2019-01-01%20until%3A2020-01-01&src=typd&lang=sin

https://twitter.com/search?l=&q=ගෝඨාභය%20since%3A2019-01-01%20until%3A2020-01-01&src=typd&lang=sin

Language –

https://twitter.com/search?l=&q=සජිත්%20since%3A2019-01-01%20until%3A2020-01-01&src=typd&**lang=sin**

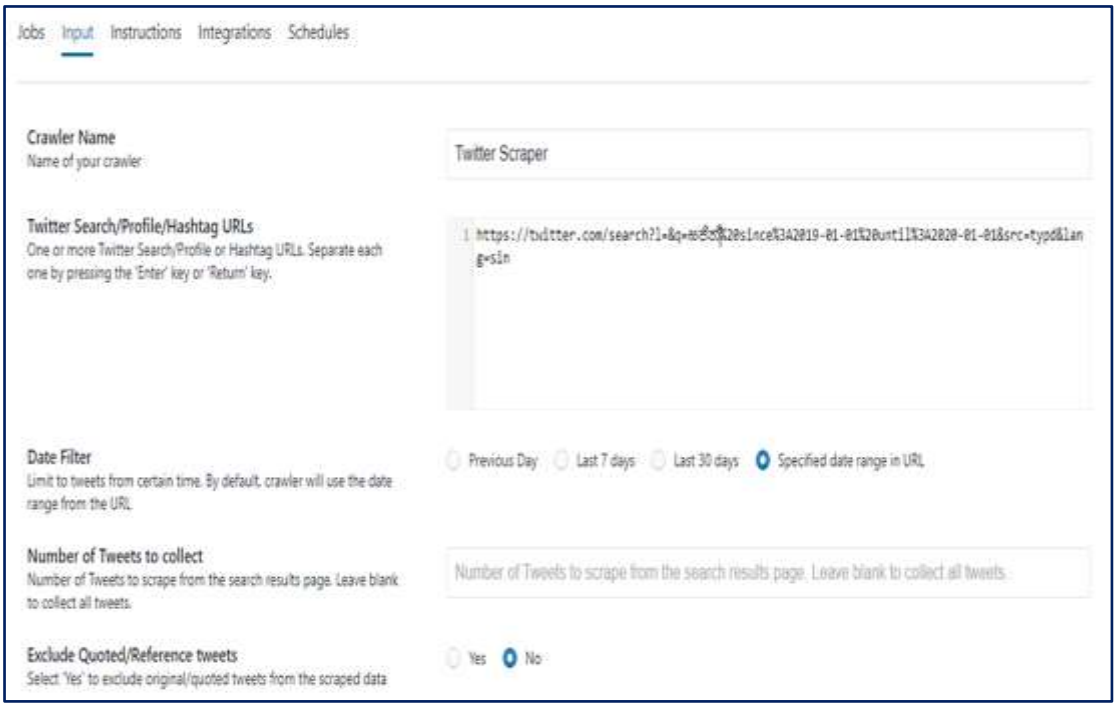


Figure 2. Input definition in Twitter Scrapper Tool

Figure 3. Scraped Data related to keyword "ගෝඨාභය"

Figure 4. Scraped Data related to keyword "සජිත්"

3.1.2 Preprocessing

It is not a good practice to use lot of junk data for direct classification hence the text preprocessing is required. Text preprocessing is considered as the most important phase in natural language processing, and it brings the text into predictable form. During this phase, unnecessary data will be removed from twitter data set ("Everything There Is to Know about Sentiment Analysis," n.d.).

In below example, extracted dataset contains lots of unnecessary data with @, numbers, non — Sinhala words, URLs etc.

e.g.

2020 අපේ එකම පැතුම ගෝඨාභය රාජපක්ෂ මහතාය..LKLK @PodujanaParty @GotabayaR @PresRajapaksa @RajapaksaNamal

සජිත් නිවටයෙක්….. එජාප නායකත්ව වෙනස පාවාදීම ගැන සජිත්ට දැඩි විරෝධය.. රනිල් විරෝධීන් සජිත් අත්හරි! අලුත් නායකයෙකු වෙනුවෙන් කදවුරක්! https://t.co/XkedBfj35G https://t.co/XgYVCsrhIy

There are various text preprocessing steps such as Data cleaning, remove stop words, stemming, lower casing etc. During this research, below highlighted preprocessing steps will be used.

- Data cleaning is the process of removing unnecessary characters such as punctuation marks, newline characters, hashtag, numbers, non — Sinhala characters and URLs. In built python functions were used to remove above unnecessary characters. Question marks can be significant during sentiment analysis. Classifier's performance was evaluated with and without the question marks in sentences. Emojis in a sentence are also added a significant importance to the sentiment of a sentence. Therefore, emojis and question marks will be handled differently, and more information related to this will be explained later.

  2020 අපේ එකම පැතුම ගෝඨාභය රාජපක්ෂ මහතාය..LKLK @PodujanaParty @GotabayaR @PresRajapaksa @RajapaksaNamal (remove .., 2020, LKLK @PodujanaParty @GotabayaR @PresRajapaksa @RajapaksaNamal)

  Stop word removal is the process of removing unnecessary words which does not provide any sentiment value to the sentence. There is an existing Sinhala language — based stop words list defined by Lakmal, D., Ranathunga, S., Peramuna, S., & Herath, I. (Lakmal, D et al., 2021b) and it is used as the predefined stop word list as an initial step. During further study, it was identified that some of the words, which are defined as the stop words, have a sentiment value in the "election domain". Example of such words in the predefined list are, "අපොයි", "අයියෝ", "වැඩි", "විශේෂ" and "වඩා". Hence the customized predefined stop word list was created based on the initial list. And also, the performance of the classifiers was evaluated with and without stop word list to

identify the impact of stop words during the performance evaluation. More information regarding this, is available in the evaluation chapter.

සජිත් මැතිඳුනි ඔබට වැලිගමදී ජනතාව අද ඉදිරිපත් කල ප්‍රශ්න වලට ක්ෂනිකව විසදුම් ලබාදුන් ආකාරයෙන් ඔබ වෙත ජනතාවගේ විශ්වාසය තහවුරු විය. **(remove කල, වෙත)**

- Stemming is the process of reducing a word into its stem. In some scenarios, tokenized words may not be available in the predefined sentiment word list, instead its stem may be available in the sentiment word list. It is further explained through the below example.
  මහතාය — This word is not included in the predefined sentiment word list, So, the sentiment value will be provided as 'None'. But the stem of this word 'මහතා' is included in this list with the sentiment value – 0. This behavior would be beneficial during the sentiment value calculation.

  Therefore, words are converted into stems by using a pre developed tool for Sinhala language called 'Sinhala language stemmer' (Yasas Senarath, n.d.) but this stemmer is still in the experimental phase. For this research, a combination of the aforesaid stemmer and Sinhala — based tokenizer was used. As per below figures 5 and 6, it was identified that this stemmer was not successful in stemming some words, hence the stemmer was eliminated.

Figure 5. Example of a Sinhala Stemmer Results for keyword "සජිත් "

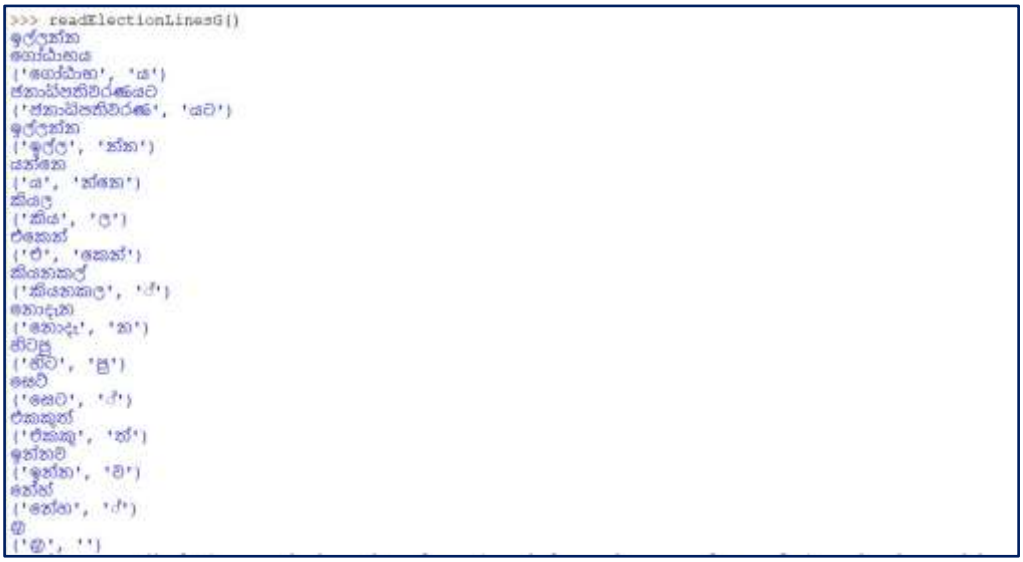


Figure 6. Example of a Sinhala Stemmer Results for keyword "ගොඩ්ඩාහය"

3.1.3 Dataset labelling

Here the main focus is to use supervised learning algorithms, so the labelling data is a prerequisite. Once a particular data set is labeled, it will highlight the features and also it helps to predict the text data behavior. These data sets will be labeled based on the polarity levels (sentiment level).

When compared with the resource rich language like English, it is very difficult to find a Sinhala sentiment lexicon for research purposes. Most of the existing Sinhala sentiment labelling methods are conducted with the help of experienced annotators as a manual approach. Some of the existing researches conducted in different languages use the available labelled sentiment from different resources. Among the different types of labelling approaches such as Inhouse (use existing resources), outsourcing (freelancers), crowdsourcing (third party), data programming, the approach used in this research is "data programming" — automated approach ("5 Approaches to Data Labeling for Machine Learning Projects," n.d.). Initial plan was to use an already developed Python API for Sinhala Sentiment lexicon – "helasentilex" (Karunanayake, n.d.) which provides more than 14000+ sentiment lexicons (figure 7) with polarity level tagged Sinhala words. But during the implementation it was identified that this API modification is not permitted. Therefore, Python code was developed for data labelling process by using the helasentilex API as a baseline.

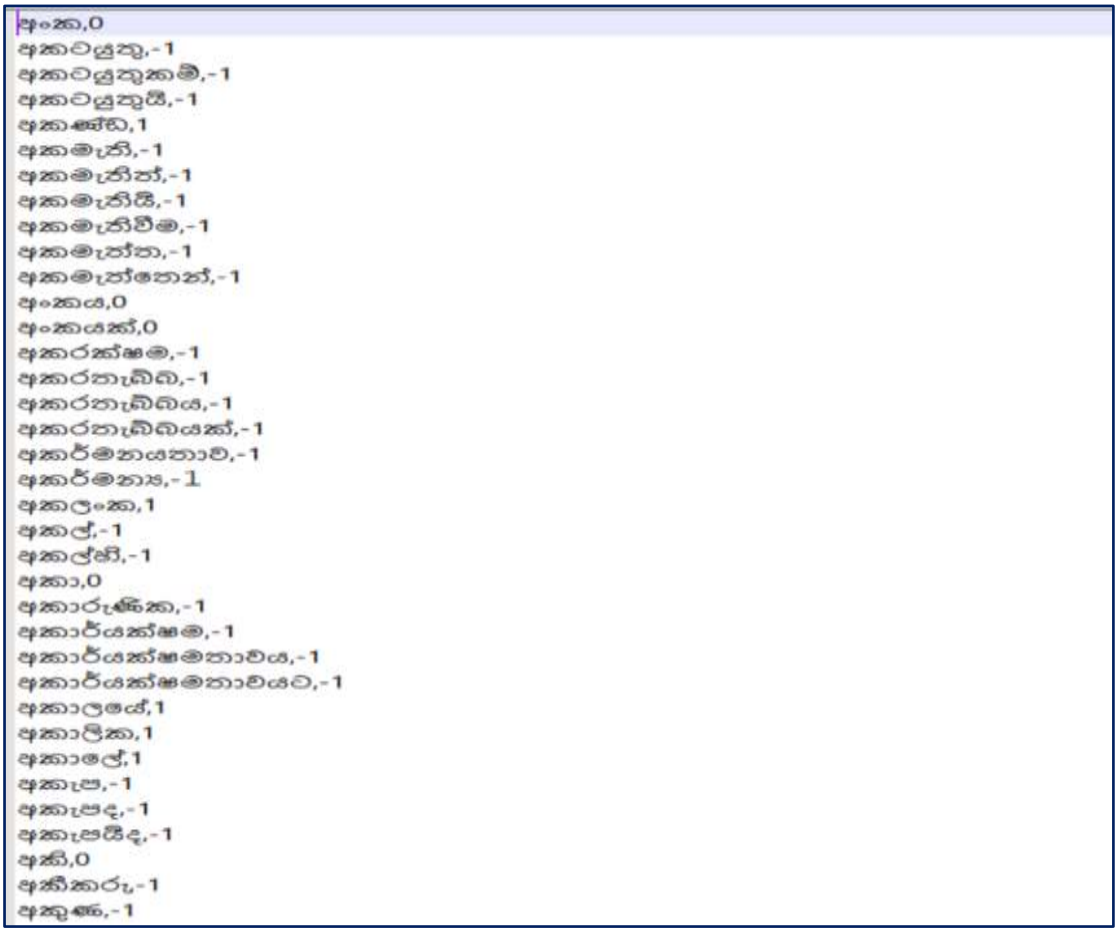


Figure 7. Sentiment level example in “Helasentilex” API

Words in each sentence will be labelled according to one of the polarity levels using the developed python program. Each word sentiment value will be extracted with the help of existing 14,000 list in helasentilex API. If a word exists in the predefined list and the sentiment value is 1(positive sentiment), that word is labelled as a ‘polarity level = 1’ word. If a word exists in the predefined list and the sentiment value is -1 (negative sentiment), that word is labelled as a ‘polarity level = -1’ word. If a word exists in the predefined list and it does not reflect any sentiment value, that word is labelled as ‘0’ (neutral sentiment). If a word does not exist in the helasentilex list, that word is labelled as a ‘polarity level – None’.

The overall sentiment value (Positive, Negative & Neutral) will be calculated by combining the sentiment score for a sentence. (Following examples will be obtained from twitter data set and the polarity will be generated using the Sinhala sentiment lexicon)

e.g.

2020 අපේ එකම පැතුම ගෝඨාභය රාජපක්ෂ මහතාය..LKLK @PodujanaParty @GotabayaR @PresRajapaksa @RajapaksaNamal

සජිත් මැතිදුනි ඔබට වැලිගමදී ජනතාව අද ඉදිරිපත් කල ප්‍රශ්න වලට ක්ෂනිකව විසදුම් ලබාදුන් ආකාරයෙන් ඔබ වෙත ජනතාවගේ විශ්වාසය තහවුරු විය.

Each word sentiment value was calculated and combined to find the overall sentence level sentiment as per the below examples.

අපේ — 0, එකම — 1, පැතුම — 1, ගෝඨාභය — None, රාජපක්ෂ — None, මහතාය — 0

සජිත් — None, මැතිදුනි —0 ඔබට — 0, වැලිගමදී — None, ජනතාව — 0, අද — 0 ඉදිරිපත් — 0, කල — 0 ,ප්‍රශ්න — -1, වලට — 0, ක්ෂනිකව — 1, විසදුම් — 1, ලබාදුන් — 1, ආකාරයෙන් — 0, ඔබ — 0, වෙත — 0, ජනතාවගේ — 0, විශ්වාසය — 1, තහවුරු — 1, විය — 0.

Sum each word sentiment value and find polarity value of the sentence

0+1+1+0 = 2 = Positive

0+0+0+0+0+0-1+0+1+1+1+1+0+0+0+0+1+1+0 = 4 = Positive

Negation handling

Negation handling is an important aspect when it comes to sentiment analysis. As mentioned in the previous sections, the overall sentiment values of the sentences are calculated by considering each word's sentiment value from the predefined sentiment value list. But this process does not include the concept of negation which will change the overall expression of the sentences.

Negations

Negations are words which affects the polarities of the next/previous words in a sentence. Examples of negation words in English language are No, Not, Never, None, don't etc. When these words are included in a sentence, the overall polarity will be changed. The same will apply for Sinhala language as well. "එපා", "නැහැ", "නොමැත", "නොවෙයි", "නැති", "නෑ", "නැත", "නොවේ", "බැරි", "බෑ" and "බැහැ" are the Sinhala language negation list which considered during this study. If the sentence polarities are calculated only considering the sentiment value of a word, the accuracy will be reduced as explained below.

<u>Sentence 1</u>

ගෝඨාභයට + දිනන්න + බෑ

(0) + (+1) + (-1) = Overall sentiment value = 0 (neutral sentiment)

Sentence 2

ගෝඨාභය + පරදන්න + බෑ

(0) + (-1) + (-1) = Overall sentiment value = -2(negative sentiment)

As shown in above examples, sentence 1 is a neutral sentence and sentence 2 is a negative sentence. But the actual sentiment values should be negative for the first sentence and positive for the second sentence. If the negation word was combined with the positive word, overall sentiment value for those two words were considered as negative. If the negation word was combined with the negative word, overall sentiment value for these two words were considered as positive. Refer below example.

Sentence 1

ගෝඨාභයට + දිනන්න + බෑ

(0) + (+1) + (-2) = Overall sentiment value = -1 (negative sentiment)

Sentence 2

ගෝඨාභය + පරදන්න + බෑ

(0) + (-1) + (+2) = Overall sentiment value = +1(positive sentiment)

Sentiment value was calculated using the pre described logic.

Negations could be categorized as Morphological and Syntactic but morphological negations were not included in this level for simplicity. Only syntactic level negation was considered.

Emoticon's sentiments

In social media platforms like Twitter, Facebook, etc. people tend to use emojis for most of their sentences to express their true feelings. When considering the sentiment analysis, sentiment of the emojis should take into consideration for accurate results. Sinhala tweets which extracted from the election period also consist of the emojis as shown in below examples.

සජිත් හිතන් ඉන්නෙ සංවර්ධනය කියන්නේ ගෙවල් හදන එක කියලා 😂 😂 😂 කොහොමද කියපන්කෝ ඉතින් ඡන්දයක් දෙන්නේ 🤦 මූත් කරන්නේ ආප්පකරපු දේම තමයි ගෙවල් හද හද නිධන් හොයනවා 🤷

@daughterislife 2015 ඉඳන් ගෝඨාභය රාජපක්ෂ ආරක්ෂක ලේකම්, මහින්ද රාජපක්ෂගේ ආණ්ඩුවක්නෙ තිබිලා තියෙන්නෙ. ඉතිං උන් ගහන්න එනකං දන්නෑ 😂😂😂

Each emoji sentiment value was observed by using the emoji sentiment ranking (Department of Knowledge Technologies, 2015). Customized list was created as per the figure 8 with the help of the baseline emoji sentiment ranking list (figure 9). If the sentence has the emoji icon in it, emoji icon sentiment value was taken from the emoji ranking list and add its sentiment value to the overall sentence level sentiment.

\U0001f602	0.221
\U00002764	0.746
\U00002665	0.657
\U0001f60d	0.678
\U0001f62d	-0.093
\U0001f618	0.701
\U0001f60a	0.644
\U0001f44c	0.563
\U0001f495	0.632
\U0001f44f	0.52
\U0001f601	0.449
\U0000263a	0.657
\U00002661	0.669
\U0001f44d	0.521
\U0001f629	-0.368
\U0001f64f	0.417

Figure 8. Emoji Sentiment Rankings

Figure 9. Baseline of Emoji Sentiment Ranking

As per the emoji sentiment ranking, emojis were represented in decimal values, not in integer values (as +1 or -1 or 0). But Sinhala sentiment lexicon was constructed based on the integer

values. Due to this behavior, another sentiment calculation was conducted with integer values (+1, -1, 0) for emoji sentiment ranking. Same emoji sentiment ranking was taken into consideration, and the value was set to +1 if the sentiment ranking value is positive, -1 if the sentiment ranking value is negative and 0 if the sentiment ranking value is zero. New sentiment ranking (figure 10) was created based on the above logic.

\U00002665	1
\U0001f60d	1
\U0001f62d	-1
\U0001f618	1
\U0001f60a	1
\U0001f44c	1
\U0001f495	1
\U0001f44f	1
\U0001f601	1
\U0000263a	1
\U00002661	1
\U0001f44d	1

Figure 10. Customized Emoji Sentiment Ranking

Automatic labelling was used to label the dataset in this study. It was tested against the negation handling and emoji ranking (with integer value — based emoji ranking and decimal value — based emoji ranking) and calculated the percentage of each candidate.

Determine the Automatic labelling accuracy

In this research study, automatic labelling approach was used instead of manual labelling process. Accuracy of this process needs to be calculated to understand the reliability of this automatic labelling process. 1000 records relevant to each candidate, were considered for this analysis. News related comments were removed and only the actual comments related to these candidates were considered and processed. These actual comments were manually annotated by researcher with the assumption that my annotation is 100% accurate. A comparison study was conducted with the manually and automatically annotated comments. Evaluation results will be further explained in the Evaluation chapter.

3.1.4 Determine the Election Prediction results

After the automatic labelling process, each sentence is labeled with the positive, negative, and neutral values as shown in below figure 11 and figure 12.

Figure 11. Example of an automatic labelling sentences for keyword "ගෝඨාභය"

Figure 12. Example of an automatic labelling sentences for keyword "සජිත්"

After finding the sentiment level polarities, overall sentiment percentage for each candidate was calculated. This results will be shown as a percentage of their total comments as shown in figure 13 and figure 14.

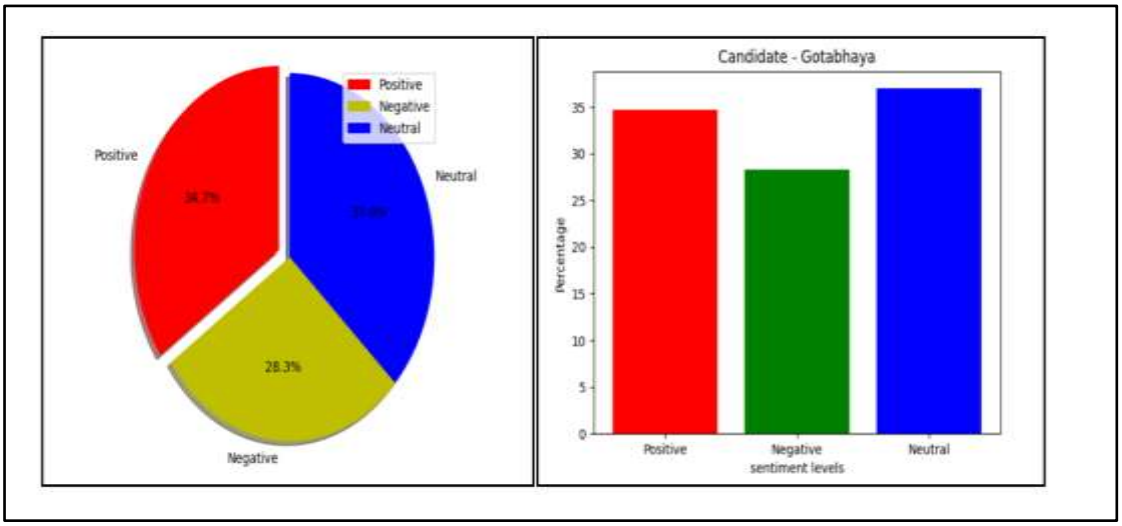


Figure 13. Sentiment level percentages for candidate "ගෝඨාභය"

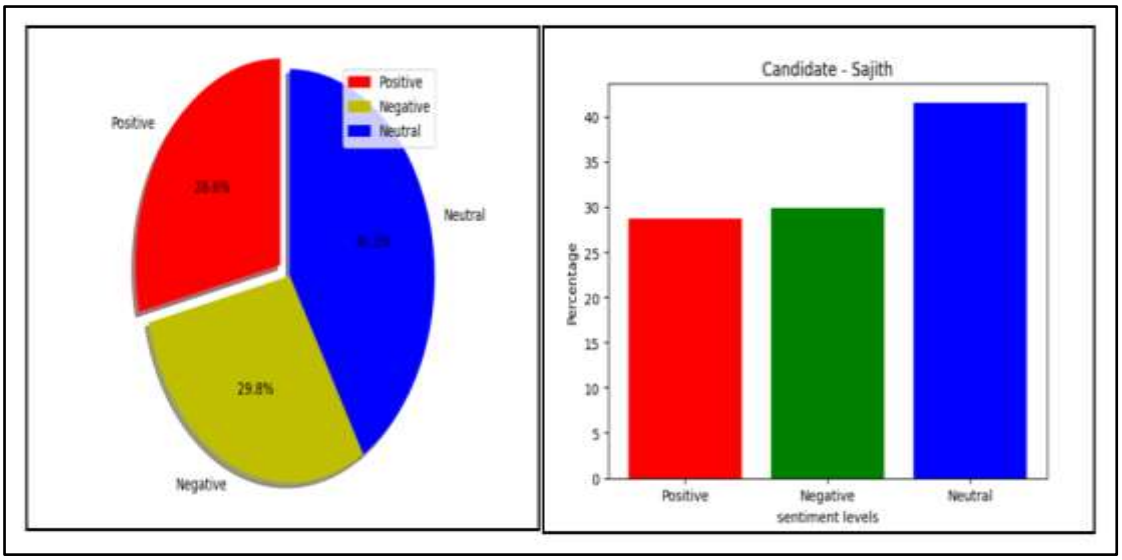


Figure 14. Sentiment level percentages for candidate "සජිත්"

As mentioned in the previous section , two methods were used to increase the accuracy of the automatic labelling by including emoticons sentiment values and negation handling. Emoji sentiments were tested against the two methods; sentiments with decimal values and sentiments with integer values. Each candidates overall sentiment percentage was calculated after applying the emoji sentiment and negation handling.

3.1.5 Construct model

Model dataset will be divided into 2 categories as training and test dataset with 70 and 30 percentages respectively. Training dataset was created by combining dataset of each candidate. Different Supervised learning classifiers such as Naïve Bayes, SVM, decision tree (Tarang Shah, 2017) will be used to train the model. Below is the short description about the classifiers which will be used during this study.

Selection of classifiers

Naïve Bayes classifier

This is considered as the simplest probabilistic machine learning algorithm which was based on Bayes theorem. As per the name 'Naive' suggests, features which use in the model are independent of each other. Change of one feature does not impact the other features in the model. This is a popular classifier due to its simplicity, easiness of coding and scalability (scikit-learn developers, 2007). The below figure 15 describes the calculation which is used for this classification.

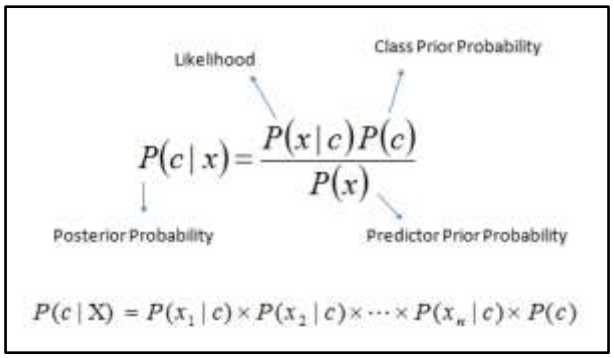


Figure 15. Naïve Bayes Algorithm (Sunil Ray, 2017)

Gaussian, multinomial and Bernoulli are the types of naïve Bayes algorithms. Bernoulli naïve bayes assumes that the features are represented in binary such as True, False/1,0. Multinomial naïve bayes assumes that the features are represented in discrete values such as movie ratings from 1 to 5. Gaussian Naïve Bayes assume that the features are continues such as length, width. For this study, Multinomial naïve Bayes used, since the labels are categorized into discrete levels with Positive, Negative and Neutral.("(4) What is the difference between the Gaussian, Bernoulli, Multinomial and the regular Naive Bayes algorithms? - Quora," n.d.)

Support vector Machine

Support Vector machine which is abbreviated as SVM is a simple classification algorithm. It is popular due to its significant accuracy with the low computational power. SVM supports both regression and classification but widely used in classification (Gandhi, 2018). As mentioned in below figure 16, SVM used to find an optimal hyperplane which classifies the datapoints in a N — dimensional space. N — dimensional space has N number of features there.

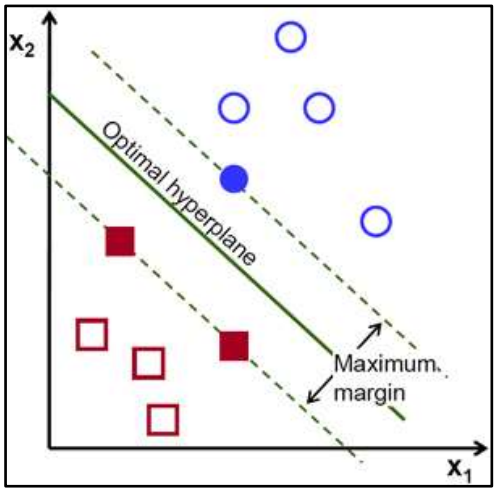


Figure 16. Support Vector Machine – Optimal hyperplane (Gandhi, 2018)

Random Forest classifier

Random Forest classifier is one of the most used classifiers due to its flexibility, simplicity, and easiness. This can be used for both regression and classification. As 'Forest' name suggests, this classifier is based on the ensemble learning concept which is used to combine multiple classifiers to get more accurate, high performance results (Niklas Donges, 2021). Predictions of each decision tree was taken into account and the final result was observed through the majority vote (figure 17).

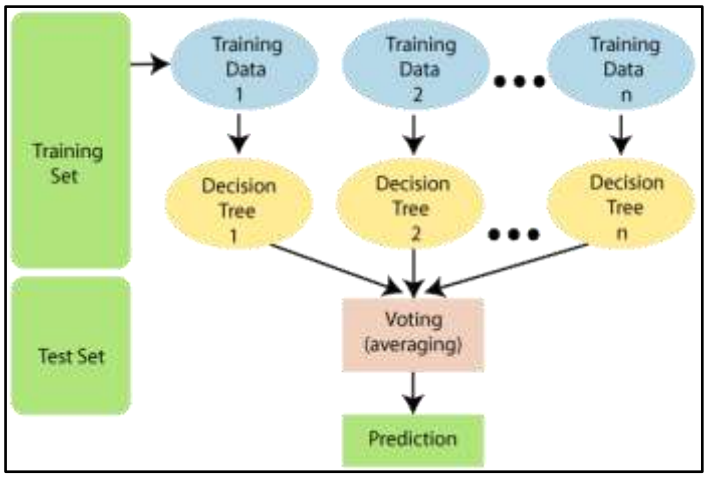


Figure 17. Random Forest classifier ("Machine Learning Random Forest Algorithm - Javatpoint," n.d.)

K — Nearest Neighbor

K — Nearest Neighbor which is abbreviated as KNN is also a simple, widely used algorithm which can be used for both regression and classification. But it is most used for classification. As the name suggests, it looks for the nearest point to predict the class of the new data point (Sai Patwardhan, 2021). KNN algorithm is categorized further into instance — based learning, lazy learning and non — parametric. Training data set

is used to predict the output in instance — based learning. In lazy learning, prediction will be done at the time of the prediction required. Non — parametric learning does not have any predefined function.

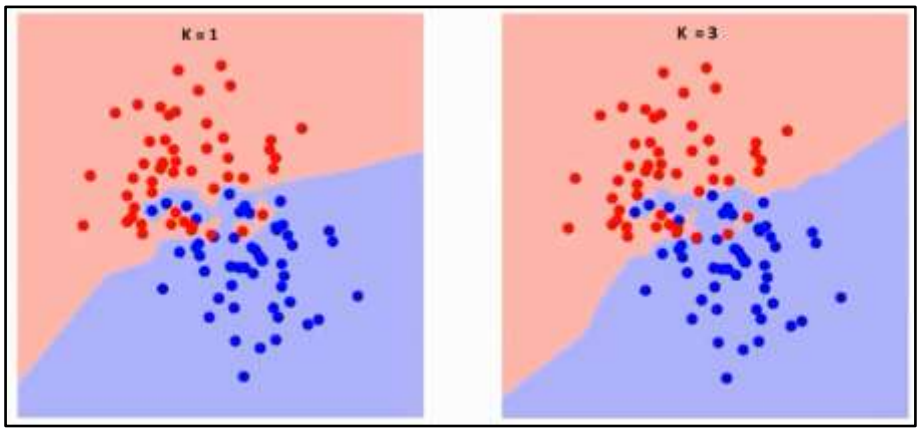


Figure 18. KNN behavior (Sai Patwardhan, 2021)

Decision Tree

Decision tree supports both regression and classification. As the name suggests, this used "tree" like structures to make a decision(Gupta, 2017). Decision tree has three features as nodes, branches, and leaves. All the internal nodes represent the condition, each branch represent the result of the condition and each leaf represent a result/class label as per the below figure 19.

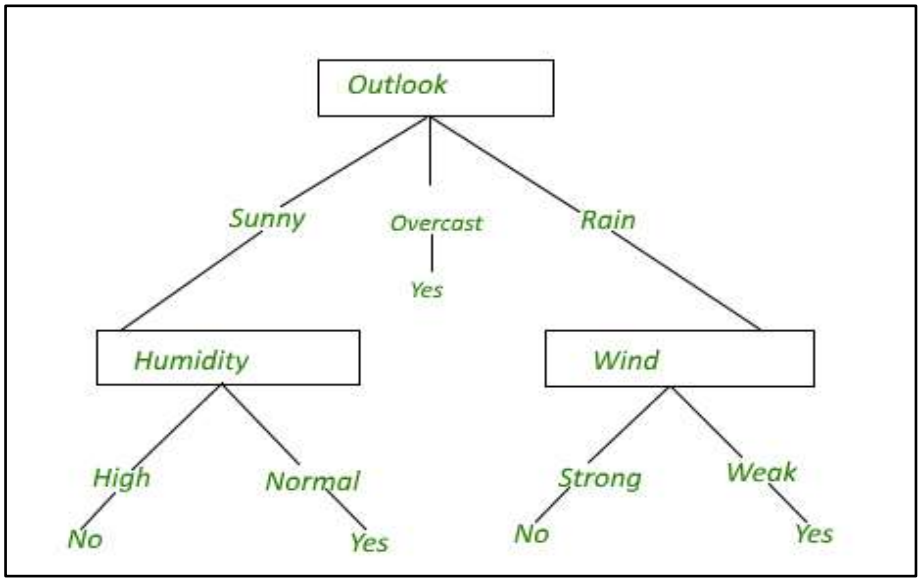


Figure 19. Decision Tree ("Decision Tree," 2017)

Feature extraction

After selecting the supervised learning classifiers, feature extraction methods were applied accordingly. As mentioned in previous sections, most of the language independent methods were applied as feature extraction methods. Hence POS tagging was eliminated but negation handling, which is a language dependent method, was used. During the feature extraction, it was identified that the baseline tokenizer in Scikit learn is not appropriately tokenize the words due to the language differences in English and Sinhala. Even though the python library supports feature extraction, there was a requirement to use a proper tokenizer for Sinhala language due to the aforesaid issue. A Sinhala tokenizer, which was developed by Yasas Senarath(Yasas Senarath, n.d.), used as a baseline tokenizer.

Bag of Words (BoW)

As the name suggests, all the texts are represented as the bag of its words without considering the word order or grammar (eiki, 2019). All the words in a sentence are considered and a feature vector was created based on these words. During the implementation, CountVectorizer by Scikit learn was used with the Sinhala tokenizer to avoid any misinterpretations during tokenization.

```
    # Create our vectorizer
    # Use Sinhala tokenizer as a base tokenizer
    vectorizer = CountVectorizer(tokenizer=SinhalaBaseTokenizer)


def SinhalaBaseTokenizer(sentence):

    try:
        tokenizer = SinhalaTokenizer()

        tokens = tokenizer.tokenize(sentence)

        return tokens

    except Exception as e:
        print("Oops!", e, "occurred.")
```

Figure 20. Use of Sinhala based tokenizer

Term Frequency, Inverse Document Frequency (TF — IDF)

In TF — IDF, both frequency and importance of the words were considered, and this is the main different of this method when compared to BoW. (“Feature Extraction Techniques - NLP,” 2020).

$$tf(w_i, r_j) = \frac{No.\,of\,times\,w_i\,occurs\,in\,r_j}{Total\,no.\,of\,words\,in\,r_j}$$

Figure 21. Term Frequency Equation (“Feature Extraction Techniques - NLP,” 2020)

$$idf(d, D) = \log \frac{|D|}{\{d \epsilon D : t \epsilon D\}}$$

Figure 22. Inverse Document Frequency Equation (“Feature Extraction Techniques - NLP,” 2020)

$$tfidf(t, d, D) = tf(t, d) * idf(d, D)$$

Figure 23. Term Frequency-Inverse Document Frequency (“Feature Extraction Techniques - NLP,” 2020)

N — grams – Unigram/Bigram/Trigram

N number of word sequence is called as N — grams. If the contiguous word sequence is one, it is called as Unigram. If the contiguous word sequence is two, it is called as Bigram. Trigram has the contiguous word sequence as three.

Unigram, Bigrams, n — grams (consecutive 1 — word, 2 — words and n — words)

Unigrams — ගෝඨාභය, අපේක්ෂක, කල, සැණින්

Bigrams — ගෝඨාභය අපේක්ෂක, කල සැණින්

Trigrams — ගෝඨාභය අපේක්ෂක කල

At the initial stage, BoW and TF — IDF were tested with the Unigram features. Another level of testing was conducted with the Bigram and Trigram for both BoW and TF — IDF to check whether the n — gram feature affects the classifiers’ accuracy.

As the next step, comparison of each classifier will be conducted based on the performance metrics and the most suitable classifier for this model will be identified.

Figure 24 shows the proposed architectural design of this research.

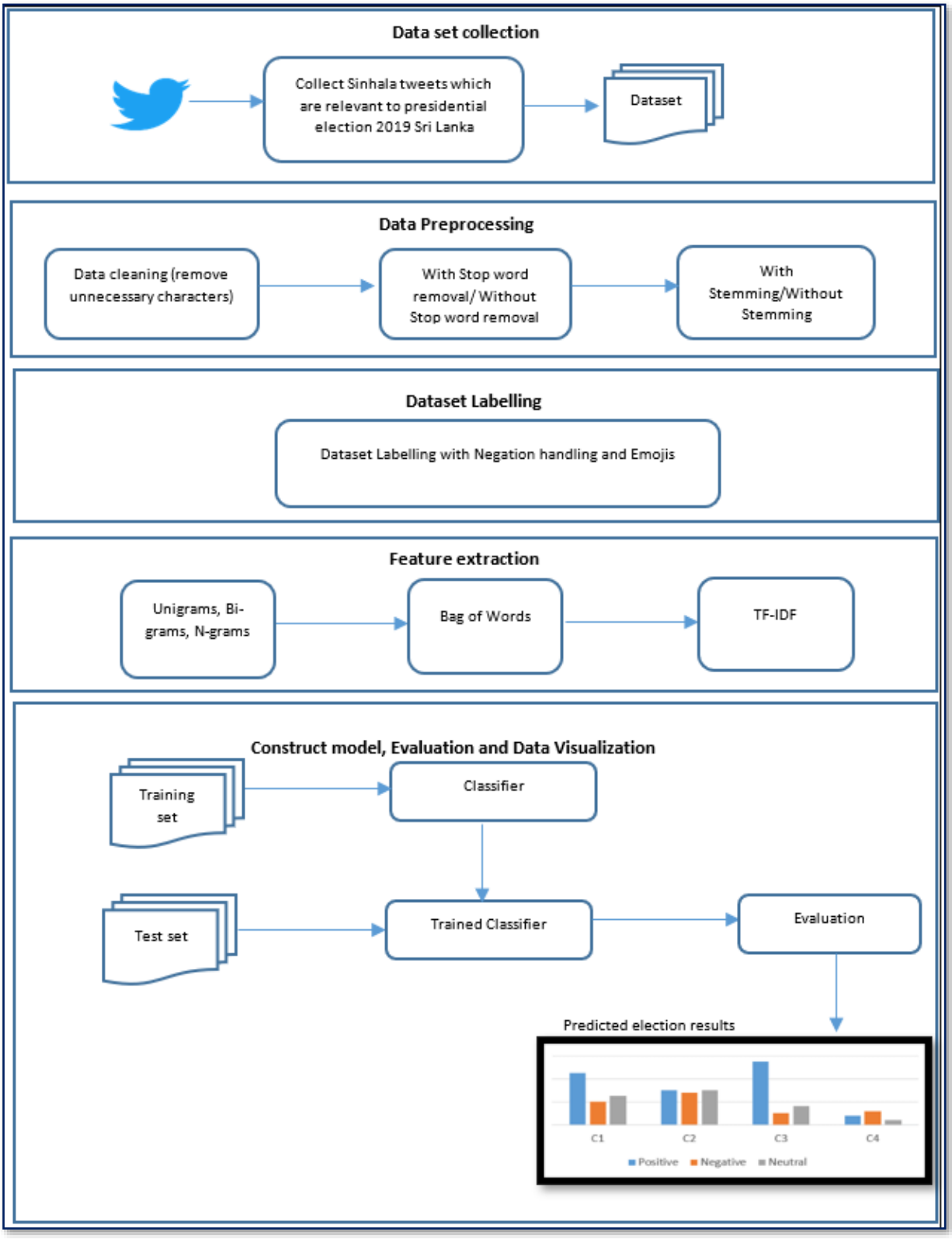


Figure 24. Proposed Architectural Design of the Research

3.1.6 Project Plan

Figure 25. Project Plan

CHAPTER 4 - EVALUATION AND RESULTS

4.1 Introduction

Internet is a rapidly growing and massive network which connects billions of humans in less than seconds. As per the latest statistics, there are 4.66 billion internet users in the world and more than 875,000 new users join this community every day ("Global Digital Overview — DataReportal – Global Digital Insights," n.d.). Surprising fact is that, statistics reveals that the average internet user spends more than 6 hours online per day ("Global Digital Overview — DataReportal – Global Digital Insights," n.d.). According to the latest statistics (Kemp, 2020), social media plays a huge role among these multiple internet usages by attracting various kind of users. They are not only overusing this, but also posting their personal life details, ideas, emotions, innovations etc. Analyzing their opinions through the social media posts, provide a lot of statistics to the society due to this limitless behavior of them.

In this research, predictive sentiment analysis is conducted using Sinhala tweets. Although there are many studies conducted based on predictive analysis in different languages, it is hard to find the Sinhala language based predictive analysis research. This predictive sentiment analysis is based on 2019 presidential election held in Sri Lanka. Same could be applied for future election results even though the research is based on 2019 election. As per the current situation, we can assume that this dataset will be much larger due to the higher social media usage.

Research questions of this research are listed below.

1. How do the sentiment of Sinhala tweets regarding Election can be used to predict election results?

- How to use automatic labelling over manual labelling for Sinhala language?

2. What are the best classifiers for predictive sentiment analysis in Sinhala language?

Evaluation approach in this research is, Experiment based.

4.2 Determine the automatic labelling accuracy

Sinhala tweets were extracted from the twitter dataset during the year – 2019 since presidential election was conducted in November 2019. Tweets related to the keywords 'ගෝඨාභය', 'සජිත්' were extracted assuming that the extracted data was based on the presidential election. There were 3928 sentences which includes 'සජිත්' as a keyword and 3893 sentences which includes

‘ගෝඨාභය’ as a keyword. After the data collection process, text preprocessing techniques were used for the initial data cleaning. After the data cleaning process, data set was labelled with the polarities Positive, Negative and Neutral. In this research, automatic labelling approach is followed since the manual labelling process is very time consuming and required additional efforts from annotators.

In this study, automatic labelling process was used to label the dataset. So, the accuracy of this automatic labelling process needs to be evaluated. Collected dataset sample count was 7821. Samples of 2000 data was extracted from the initial dataset to measure the accuracy of the automatic labelling over manual labelling. Even though there were 2000 sentences, it was identified that this collection includes both news and comments as shown below.

Example Comments

ගෝඨාභය ජනාධිපතිවරණයට ඉල්ලන්න යන්නෙ කියල Aljazeera එකෙන් කියනකල් නොදැන හිටපු සෙට් එකකුත් ඉන්නව නේහ්... 🤭

සජිත් මැතිඳුනි ඔබට වැලිගමදී ජනතාව අද ඉදිරිපත් කල ප්‍රශ්න වලට ක්ෂනිකව විසදුම් ලබාදුන් ආකාරයෙන් ඔබ වෙත ජනතාවගේ විශ්වාසය තහවුරු විය.

Example News

නියෝජ්‍ය නායකකමෙන් මං සෑහීමකට පත් වෙන්නේ නෑ..- සජිත් අගමැතිගේ යෝජනාවෙන් පසු කියයි...[Video] https://t.co/4VG21vqhAd

මම දැන් ඇමරිකානු පුරවැසියෙක් නෙමෙයි - ගෝඨාභය රාජපක්ෂ (දෙරණ 360) #SriLanka

Automatic labelling accuracy was evaluated by comparing the manual labelling results with automatic labelling results. Only the comments were extracted from the sample dataset to measure the automatic labelling accuracy. There were 653 comments included in the sample dataset of 2000 records. Out of these 653 comments, correctly classified sentences count was 219 and incorrectly classified sentences count was 434. This manual labelling was done by the researcher assuming that the annotation was conducted accurately.

33% of dataset is correctly classified and 67% of dataset is incorrectly classified using automatic labelling. Cohen’s kappa coefficient calculation was conducted to find the accuracy of the automatic labelling process. Cohen’s kappa coefficient (k) statistic measures the inter — rater reliability (Audrey Schnell, 2020) for categorical variables.

Cohen's kappa coefficient was measured based on the positive and negative sentence count with the assumption that neutral sentences do not carry any sentiment information about the candidates.

Table 1. Confusion matrix for actual and predicted values

	Predict Positive	*Predict Negative*
Actual Positive	91	94
Actual Negative	24	68

Step 1: calculate observed proportional agreement

91 sentences were rated as positive sentences by both manual labelling and automatic labelling

68 sentences were rated as negative sentences by both manual labelling and automatic labelling

Observed percentage agreement = (91 + 68) / 277 = 0.57

Step 2: Calculate probability both randomly shows the sentences are positive

115 sentences were rated as randomly positive by automatic labelling

185 sentences were rated as randomly positive by manual labelling

(115 / 277) × (185 / 277) = 0.27%

Step 3: Calculate probability both randomly shows the sentences are negative

162 sentences were rated as randomly negative by automatic labelling

92 sentences were rated as randomly negative by manual labelling

(162 / 277) × (92 / 277) = 0.19%

Step 4: Get the overall probability

0.27 + 0.19 = 0.46%

Step 5: Calculate Cohen's Kappa value

(Stephanie, 2014) Cohen's Kappa value = $\frac{(\mathrm{Po} - \mathrm{Pe})}{(1 - \mathrm{Pe})}$

Po= relative observed agreement

Pe = hypothetical probability of chance agreement

Cohen's Kappa value = (0.57 - 0.46) / (1 - 0.46)

= 0.20

Kappa statistic interpreted as below figure 26. (Audrey Schnell, 2020)

Kappa	*Agreement*
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

Figure 26. Kappa statistics interpretation ("Table 2," n.d.)

As per the statistics interpretation (figure 26), there is a slight agreement between automatic labelling and manual labelling. One of the aims of this research is to increase the accuracy of this automatic labelling process. This will reduce the huge labor time which incurred during the manual labelling process. Below methods will be included during the labelling process to further improve the accuracy of this automatic process.

- Inclusion of emoticons sentiment value during the labelling process
  - Emoticon's sentiment value with Decimal values
  - Emoticon's sentiment value with Integer values
- Inclusion of negation handling for sentiment calculation during the labelling process

4.2.1 Emoticon's sentiment values with Decimal sentiments and Inclusion of negation handling

After repeating the same process for emoji "decimal" sentiments with negation handling, it was identified that the correctly classified sentence count was 283 and incorrectly classified sentence count was 370. These results clearly indicate that the accuracy of the automatic labelling process was increased up to 43%.

Table 2. Confusion matrix for actual and predicted values

	Predict Positive	*Predict Negative*
Actual Positive	130	72

Actual Negative	38	60

Step 1: calculate observed proportional agreement

130 sentences were rated as positive sentences by both manual labelling and automatic labelling

60 sentences were rated as negative sentences by both manual labelling and automatic labelling

Observed percentage agreement = (130 + 60) / 300 = 0.633

Step 2: Calculate probability both randomly shows the sentences are positive

168 sentences were rated as randomly positive by automatic labelling

202 sentences were rated as randomly positive by manual labelling

(168 / 300) × (202 / 300) = 0.377%

Step 3: Calculate probability both randomly shows the sentences are negative

132 sentences were rated as randomly negative by automatic labelling

98 sentences were rated as randomly negative by manual labelling

(132 / 300) × (98 / 300) = 0.14%

Step 4: Get the overall probability

0.377 + 0.14 = 0.517%

Step 5: Calculate Cohen's Kappa value

(Stephanie, 2014) Cohen's Kappa value = $\frac{(\mathrm{Po} - \mathrm{Pe})}{(1 - \mathrm{Pe})}$

Po= relative observed agreement

Pe = hypothetical probability of chance agreement

Cohen's Kappa value = (0.633 - 0.517) / (1 - 0.517)

= 0.24

As per the Kappa statistics interpretation, there is a fair agreement between automatic labelling and manual labelling, and this implies better results than the previously observed results.

4.2.2 Emoticon's sentiment values with Integer sentiments and Inclusion of negation handling

After repeating the same process for emoji "Integer" sentiments with negation handling, it was identified that out of these 653 comments, correctly classified sentence count was 283 and incorrectly classified sentence count was 370. It resulted the same percentage value as received from the previous section.

So, the conclusion is, there is no significant difference in accuracy levels of emoji sentiment with "decimal" values and "integer" values.

As per the research done by Demotte (Demotte et al., 2020), use of the binary values (Positive and Negative) for labelling has increased the Cohens kappa value from 0.52 to 0.92. Hence the binary classification is used for this study to test whether the automatic labelling accuracy could be increased. Dataset with 1000 records which were extracted earlier consists of 253 positive labels and 129 negative labels. Out of 253 records, 130 records were correctly classified as positive and out of 129 records, 60 records were correctly classified as negative. So, after the binary classification was applied, automatic labelling accuracy was increased up to 49.7% ~ 50% which is a significant amount when considering limitation of resources for Sinhala language.

4.3 Determine Election prediction results accuracy

After completing the automatic labelling of the dataset, the calculation was carried out to check whether the individual candidate's results are based on Positive, Negative or Neutral sentiments. Graphical representation will be used to display prediction results as figure 27 and 28.

Table 3. Evaluation of Polarity measures for each candidate

Candidate Name	*Sajith Premadasa*	*Gotabaya Rajapaksa*
No of Positive comments	1116	1337
No of Negative comments	1162	1091
No of Neutral comments	1619	1426
% of Positive comments	28.63%	34.69%

% of Negative comments	29.81%	28.30%
% of Neutral comments	41.54%	37.00%

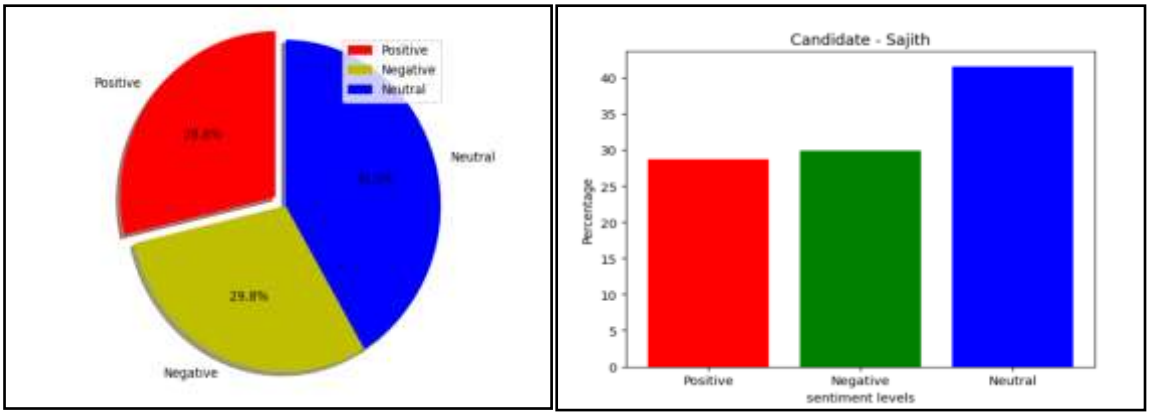


Figure 27. Percentage of Overall comments for Candidate - Sajith

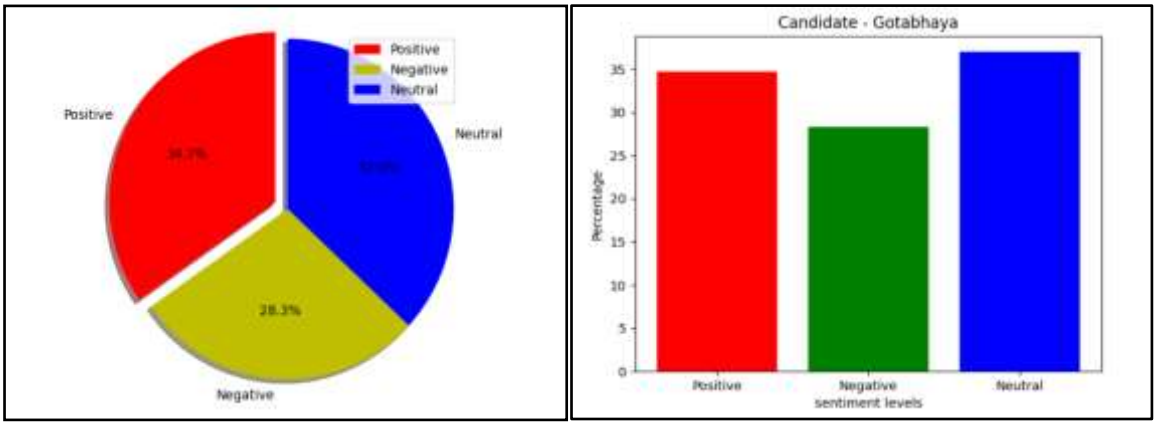


Figure 28. Percentage of overall comments for Candidate - Gotabaya

4.3.1 Election prediction results using negation handling and emoji sentiment

As mentioned in the methodology section, election results were calculated by applying negation handling and emoji sentiment to improve the automatic labelling process. Emoji sentiment impact for automatic labelling was tested against two methods; sentiments with "decimal" values and sentiments with "integer" values.

Emoticons sentiment — with decimal values

Table 4. Evaluation of Polarity measures for each candidate

Candidate Name	*Sajith Premadasa*	*Gotabaya Rajapaksa*

No of Positive comments	1364	1593
No of Negative comments	958	895
No of Neutral comments	1575	1366
% of Positive comments	35.0%	41.3%
% of Negative comments	24.6%	23.2%
% of Neutral comments	40.4%	35.4%

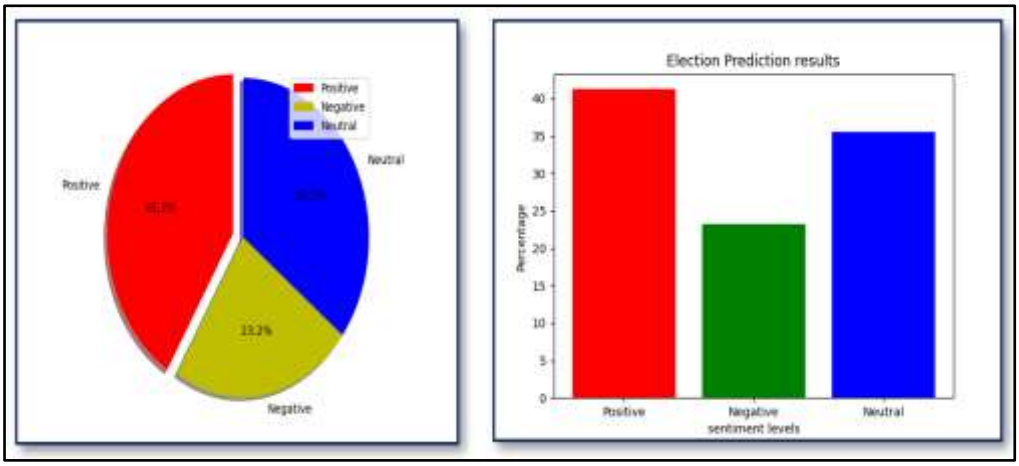


Figure 29. Sentiment level percentages for candidate "ගෝඨාභය"

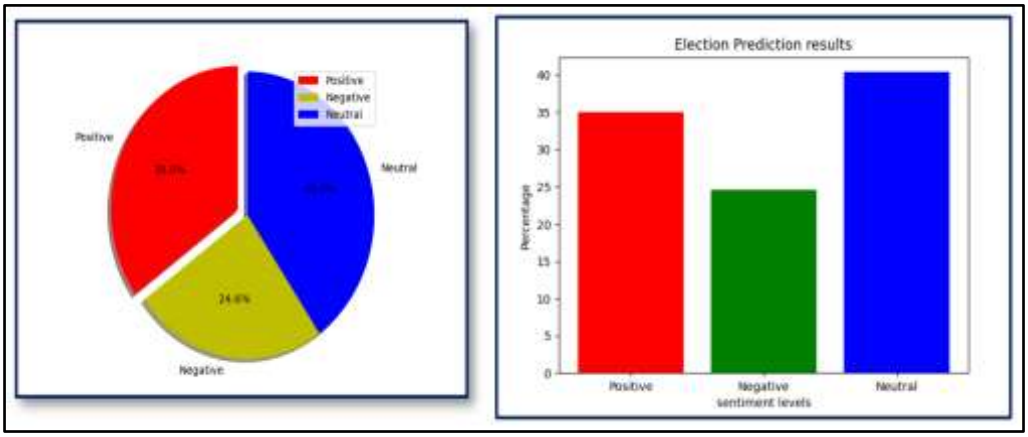


Figure 30. Sentiment level percentages for candidate "සජිත්"

Emoticons sentiment — with integer values(+1,-1,0)

Table 5. Evaluation of Polarity measures for each candidate

Candidate Name	*Sajith Premadasa*	*Gotabaya Rajapaksa*
No of Positive comments	1368	1590
No of Negative comments	906	851
No of Neutral comments	1623	1413

% of Positive comments	35.1%	41.3%
% of Negative comments	23.2%	22.1%
% of Neutral comments	41.6%	36.7%

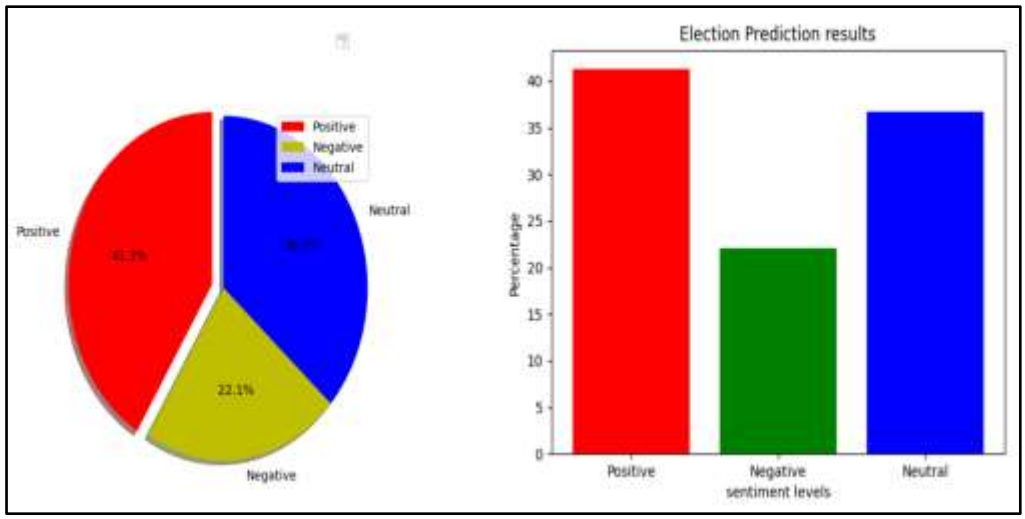


Figure 31. Sentiment level percentages for candidate "ගෝඨාභය"

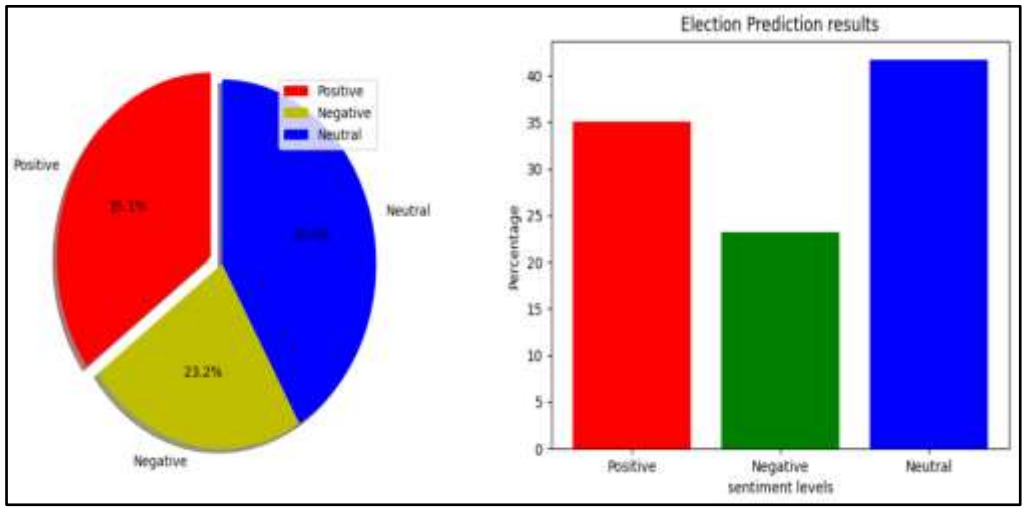


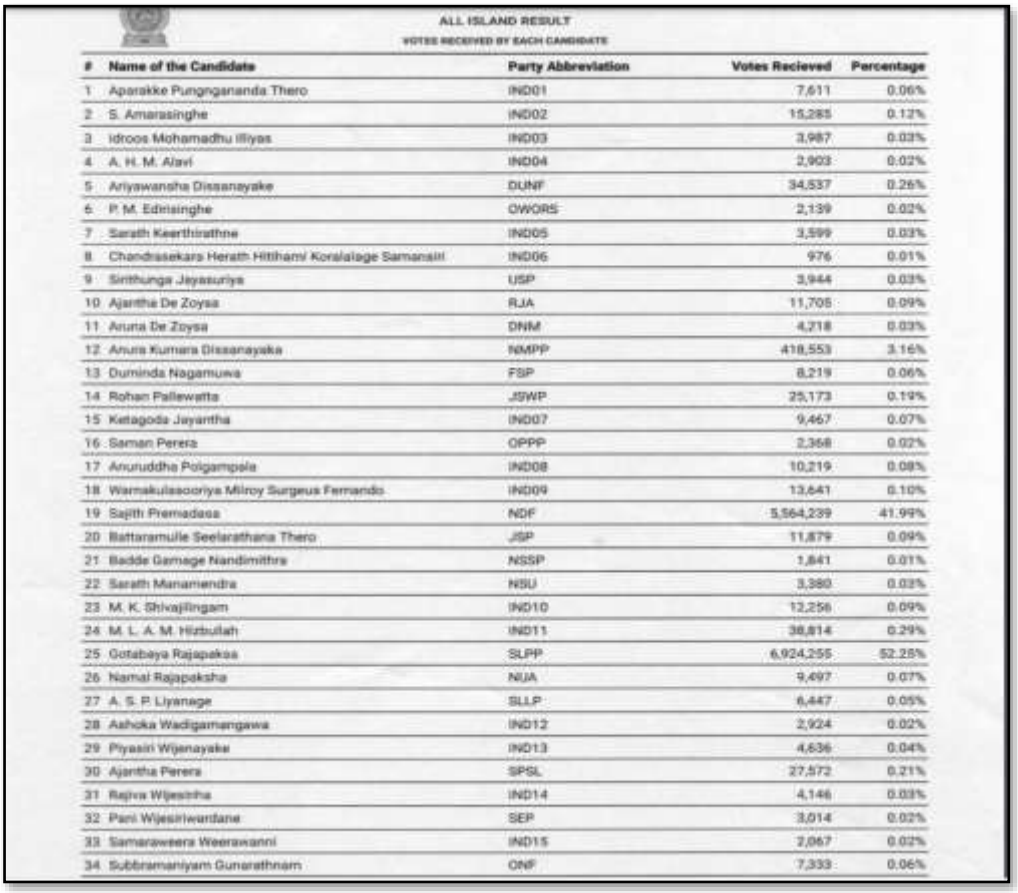
Figure 32. Sentiment level percentages for candidate "සජිත්"

As per above figures (figure 29, 30, 31 & 32), there is not any significant difference on the accuracy of the automatic labelling when using the two methods identified for emoji sentiments. Therefore emoticons polarity values which were defined by the Department of Knowledge Technologies (Department of Knowledge Technologies, 2015) will be used for this study.

Even though the emoji sentiment methods(decimal/integer) did not have any impact in the accuracy, combination of both methods (negation handling and emoji sentiment) has a impact on the process which leads to the higher accuracy level. as shown in above figures (figure 29, 30, 31 & 32). Positive sentiment count was increased for both candidates significantly.

This clearly indicates that the candidate ‘Gotabaya’ has higher positive comments than the candidate ‘Sajith’ in twitter when using the combination method. And also, there is a slight difference between negative comments for both candidates and the negative comment percentage is lesser for candidate Gotabaya.

This result will be evaluated with the actual presidential election results – 2019 in Sri Lanka. Actual results will be extracted from the government election results site in Sri Lanka (Election Commission of Sri Lanka, n.d.) as per the below figure 33.



ALL ISLAND RESULT

VOTES RECEIVED BY EACH CANDIDATE

#	Name of the Candidate	Party Abbreviation	Votes Recieved	Percentage
1	Aparakke Pungngananda Thero	IND01	7,611	0.06%
2	S. Amarasinghe	IND02	15,285	0.12%
3	Idroos Mohamadhu Illiyas	IND03	3,987	0.03%
4	A. H. M. Azavi	IND04	2,903	0.02%
5	Ariyawansha Dissanayake	DUNF	34,537	0.26%
6	P. M. Edirisinghe	OWORS	2,139	0.02%
7	Sarath Keerthirathne	IND05	3,599	0.03%
8	Chandrasekara Herath Hitthami Koralalage Samansiri	IND06	976	0.01%
9	Sinthunga Jayasuriya	USP	3,944	0.03%
10	Ajantha De Zoysa	RJA	11,705	0.09%
11	Aruna De Zoysa	DNM	4,218	0.03%
12	Anura Kumara Dissanayaka	NMPP	418,553	3.16%
13	Duminda Nagamuwa	FSP	8,219	0.06%
14	Rohan Pallewatta	JSWP	25,173	0.19%
15	Ketagoda Jayantha	IND07	9,467	0.07%
16	Saman Perera	OPPP	2,368	0.02%
17	Anuruddha Polgampala	IND08	10,219	0.08%
18	Warnakulasooriya Milroy Surgeus Fernando	IND09	13,641	0.10%
19	Sajith Premadasa	NDF	5,564,239	41.99%
20	Battaramulle Seelarathana Thero	JSP	11,879	0.09%
21	Badde Gamage Nandimithra	NSSP	1,841	0.01%
22	Sarath Manamendra	NSU	3,380	0.03%
23	M. K. Shivajilingam	IND10	12,256	0.09%
24	M. L. A. M. Hizbullah	IND11	38,814	0.29%
25	Gotabaya Rajapaksa	SLPP	6,924,255	52.25%
26	Namal Rajapaksha	NUA	9,497	0.07%
27	A. S. P. Liyanage	SLLP	6,447	0.05%
28	Ashoka Wadigamangawa	IND12	2,924	0.02%
29	Piyasiri Wijenayaka	IND13	4,636	0.04%
30	Ajantha Perera	SPSL	27,572	0.21%
31	Rajiva Wijesinha	IND14	4,146	0.03%
32	Pani Wijesiriwardane	SEP	3,014	0.02%
33	Samaraweera Weerawanni	IND15	2,067	0.02%
34	Subbramaniyam Gunarathnam	ONF	7,333	0.06%

Figure 33. Actual Presidential Election Results 2019 ((Election Commission of Sri Lanka, n.d.)

2019 actual election results were divided mostly between two candidates: Sajith Premadasa (41.99%) and Gotabaya Rajapaksa (52.25%). Results which were found in this research study is divided between the polarities; positive, negative, and neutral. For the actual and predicted results evaluation, only the positive data was considered since there is no confirmation whether the negative votes in one candidate will become a vote in another candidate. Even though there are negative sentences in twitter, people might not vote or vote to a completely different candidate. So, only considering the positive, Sajith had 46% and Gotabhaya had 54% in the predicted results.

Total Number of Positive comments = 1364 +1593 = 2957

Gotabhaya comments = 1593

Sajith comments = 1364

Gotabhaya winning percentage = 1593/2957 = 54%

Sajith winning percentage = 1364/2957 = 46%

it is evident from the results that the proposed methodology gave a near prediction to the actual result from analyzing tweets in Sinhala language.

4.4 Measuring classifiers accuracy

Sinhala tweets were divided into training and test data set with the ratio of 0.67 and 0.33 respectively. Since the focus of this research is on supervised learning classifiers, developed model is trained with the labelled dataset. Test dataset results will be evaluated using the trained model. During dataset evaluation, multiple supervised learning classifiers such as Random Forest, Naïve Bayes, Support Vector Machine, K Nearest Neighbor and Decision tree were used. These classifiers' performance was measured using the performance metrics such as Accuracy, Precision, Recall and F1 — score. (Mohammed Sunasra, 2017)

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$
$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$
$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$
$$\text{F}_1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Figure 34. Performance metrics ("machine learning - Classification report in scikit learn," n.d.)

During the testing it was identified that classifiers performance was varied based on the below listed criteria.

4.4.1 Stop words removal/non removal in preprocessing

Stop words contain few information and does not add much meaning to the sentence (Ganesan, 2019). Example of stop words in Sinhala language are "හා", "නිසා", "බවට", "සිට", "සහ" and "සමග". There is an existing Sinhala language — based stop word list defined by Lakmal (Lakmal, D et al., 2021b) and it is used as the baseline stop word list. When further examining this stop word list, it was identified that some of the words mentioned in this list contain sentiment values which provide value addition to the sentences. Example of such words in the

predefined list are, “අපොයි”, “අයියෝ”, “වැඩි”, “විශේෂ” and “වඩා”. Hence the customized predefined stop word list was created based on the available list by Lakmal (Lakmal, D et al., 2021b). Then the classifiers’ performance was measured by two scenarios as with stop word removal and without stop word removal. Below figures (figure 35, 36) show the results of the classifiers during these two methods.

With stop word removal

Figure 35. Evaluation metrics for classifiers with stop word removal

Without stop word removal

Figure 36. Evaluation metrics for classifiers without stop word removal

Above figures 35 and 36 clearly imply that stop word removal did not have positive impact on the classifier’s performance.

Conclusion: Stop word removal process was eliminated from the preprocessing stage, since it doesn’t affect the classifier’s performance during sentiment analysis

4.4.2 Question marks removal/non removal in preprocessing

People use question marks in a sentence to emphasize the uncertainty. Preprocessing step was carried out, with and without question marks to identify the effect of question marks removal.

Below figures (figure 37, 38) show the results of the classifiers during these two methods: with question marks removal and without question marks removal.

With question marks removal

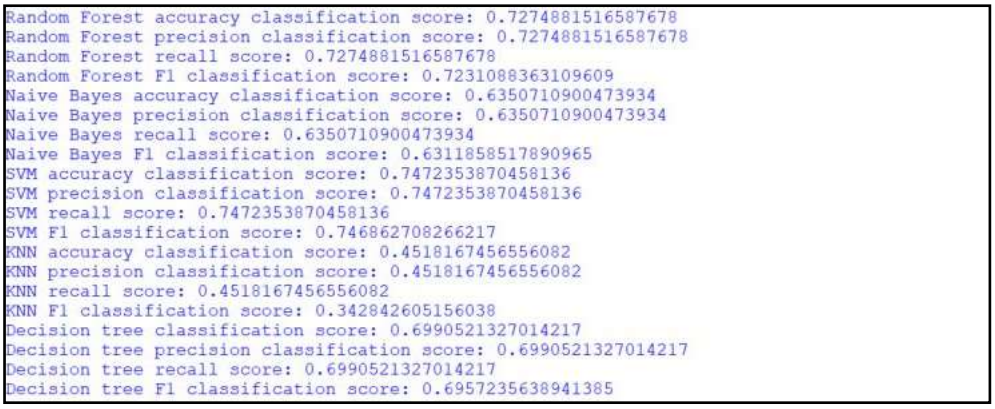


Figure 37. Evaluation metrics for classifiers with question marks removal

Without question marks removal

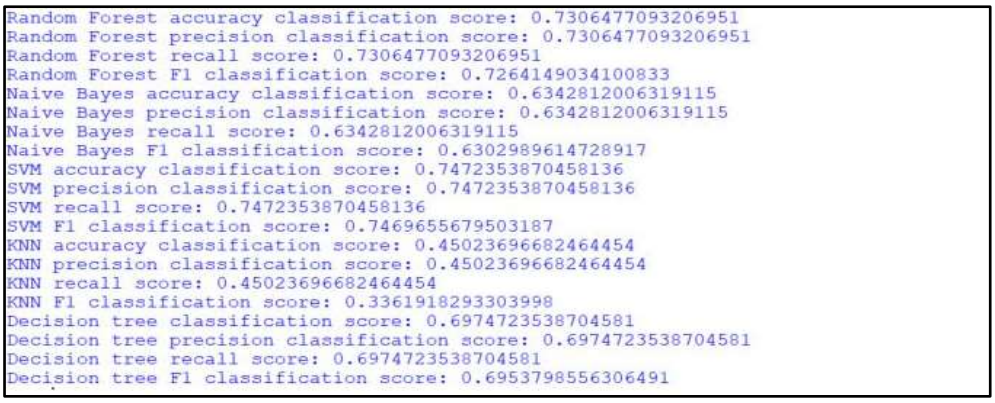


Figure 38. Evaluation metrics for classifiers without question marks removal

Above figures 37 and 38 clearly imply that question marks removal did not have any positive impact on the classifiers' performance. It has a negative impact on the classifier's performance. Classifier's performance is slightly higher when a sentence has question marks.

Conclusion: question marks removal process is eliminated from the preprocessing stage since question mark carried out valuable sentiment data in it as per the experimental results.

As per the observed results, Support Vector Machine classifier provides higher accuracy when compared to other classifiers, Random forest, Naïve Bayes, K Nearest and Decision tree.

4.4.3 Feature extraction

For feature extraction, different types of techniques can be used, and feature extraction methods can be divided as language dependent and language independent features. POS tagging,

Negation handling, syntactic dependency and opinion words are examples for language dependent features. Since this study is based on Sinhala language and due to the limited resources available for Sinhala language, only negation handling was considered during this study as a language dependent feature. Bag of words, N — gram, Term frequency and word embedding are the techniques which can be used as language independent features. Bag of Words, N — gram and term frequency are used as the feature extraction techniques in this study. As mentioned in the methodology section, Bow and TF — IDF tested against the n — gram features. Below figures (figure 39, 40, 41, 42, 43, 44) show the results which observed by applying different feature extraction methods.

BoW (with unigram)

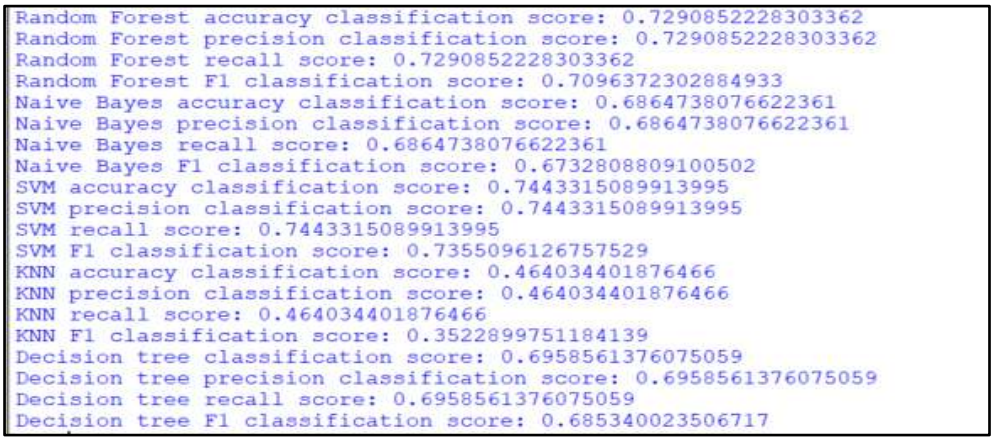


Figure 39. Bow (with unigram)- classifiers performance

BoW(with bigram)

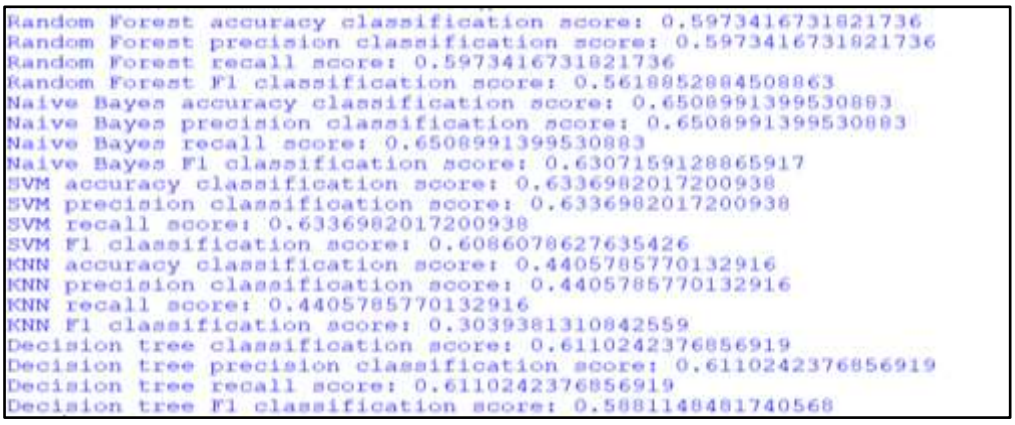


Figure 40. Bow (with bigram)- classifiers performance

BoW (with trigram)

```
Random Forest accuracy classification score: 0.584831899921814
Random Forest precision classification score: 0.584831899921814
Random Forest recall score: 0.584831899921814
Random Forest F1 classification score: 0.5423826108837818
Naive Bayes accuracy classification score: 0.6278342455043002
Naive Bayes precision classification score: 0.6278342455043002
Naive Bayes recall score: 0.6278342455043002
Naive Bayes F1 classification score: 0.6057752125024257
SVM accuracy classification score: 0.599687255668491
SVM precision classification score: 0.599687255668491
SVM recall score: 0.599687255668491
SVM F1 classification score: 0.562906255722599
KNN accuracy classification score: 0.43979671618451915
KNN precision classification score: 0.43979671618451915
KNN recall score: 0.43979671618451915
KNN F1 classification score: 0.3021488770381599
Decision tree classification score: 0.6051602814698983
Decision tree precision classification score: 0.6051602814698983
Decision tree recall score: 0.6051602814698983
Decision tree F1 classification score: 0.5756299707858569
```

Figure 41. Bow (with trigram)- classifiers performance

TF — IDF vectorizer (with unigram)

```
Random Forest accuracy classification score: 0.6989835809225958
Random Forest precision classification score: 0.6989835809225958
Random Forest recall score: 0.6989835809225958
Random Forest F1 classification score: 0.6777672939394156
Naive Bayes accuracy classification score: 0.6422986708365911
Naive Bayes precision classification score: 0.6422986708365911
Naive Bayes recall score: 0.6422986708365911
Naive Bayes F1 classification score: 0.5839993228522297
SVM accuracy classification score: 0.7404222048475372
SVM precision classification score: 0.7404222048475372
SVM recall score: 0.7404222048475372
SVM F1 classification score: 0.7300196967135477
KNN accuracy classification score: 0.5762314308053167
KNN precision classification score: 0.5762314308053167
KNN recall score: 0.5762314308053167
KNN F1 classification score: 0.5547136506277047
Decision tree classification score: 0.6293979671618452
Decision tree precision classification score: 0.6293979671618452
Decision tree recall score: 0.6293979671618452
Decision tree F1 classification score: 0.6154480990520123
```

Figure 42. TF-IDF (with unigram)- classifiers performance

TF — IDF vectorizer (with bigram)

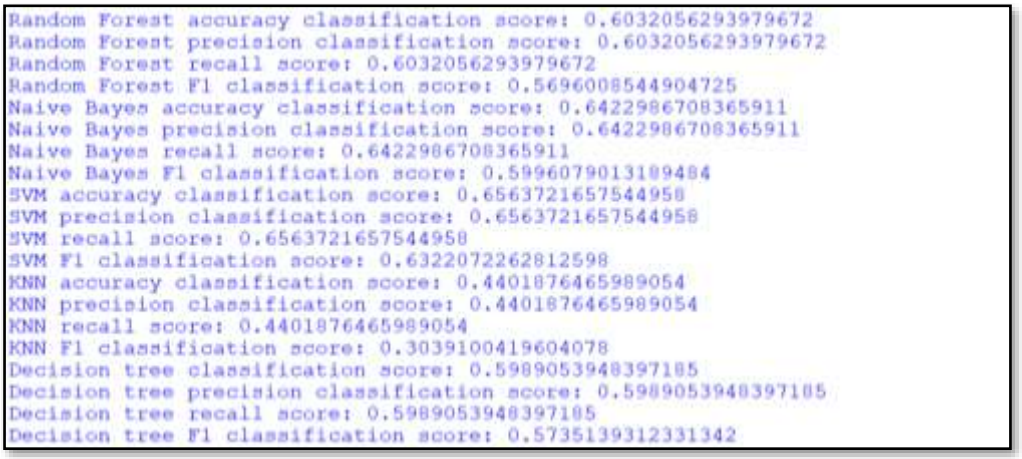


Figure 43. TF-IDF (with Bigram)- classifiers performance

TF — IDF vectorizer (with trigram)

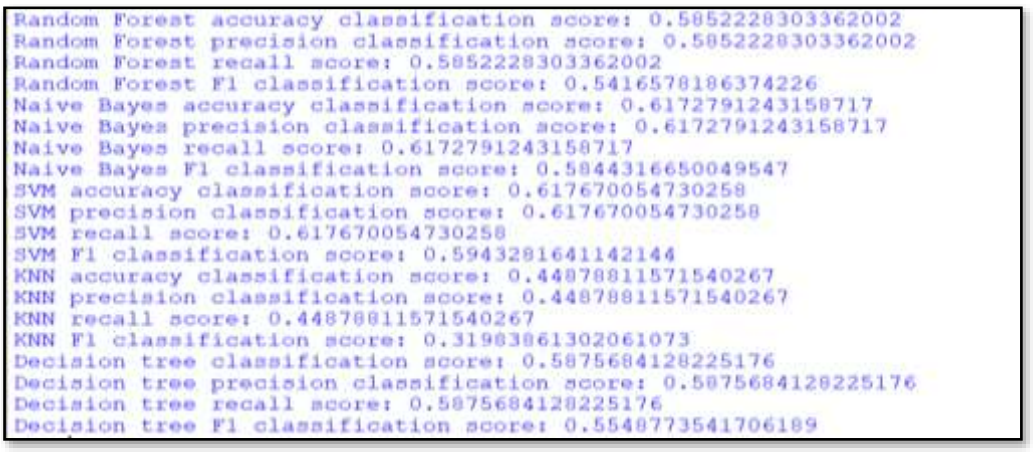


Figure 44. TF-IDF (with Trigram)- classifiers performance

Table 6. Classifier's performance metrics based on feature extraction methods

		BoW			*TF-IDF*		
		Unigram	*Bigram*	*Trigram*	*Unigram*	*Bigram*	*Trigram*
Random Forest	accuracy	0.7290	0.5973	0.5848	0.6989	0.6032	0.5852
	precision	0.7290	0.5973	0.5848	0.6989	0.6032	0.5852
	recall	0.7290	0.5973	0.5848	0.6989	0.6032	0.5852
	F1-score	0.7096	0.5618	0.5423	0.6777	0.5696	0.5416
Naïve Bayes	accuracy	0.6864	0.6508	0.6278	0.6422	0.6422	0.6172
	precision	0.6864	0.6508	0.6278	0.6422	0.6422	0.6172
	recall	0.6864	0.6508	0.6278	0.6422	0.6422	0.6172
	F1-score	0.6732	0.6307	0.6057	0.5839	0.5996	0.5844
SVM	accuracy	0.7443	0.6336	0.5996	0.7404	0.6563	0.6176
	precision	0.7443	0.6336	0.5996	0.7404	0.6563	0.6176
	recall	0.7443	0.6336	0.5996	0.7404	0.6563	0.6176
	F1-score	0.7355	0.6086	0.5629	0.7300	0.6322	0.5943
KNN	accuracy	0.4640	0.4405	0.4397	0.5762	0.4401	0.4487
	precision	0.4640	0.4405	0.4397	0.5762	0.4401	0.4487
	recall	0.4640	0.4405	0.4397	0.5762	0.4401	0.4487
	F1-score	0.3522	0.3039	0.3021	0.5547	0.3039	0.3198
Decision Tree	accuracy	0.6958	0.6110	0.6051	0.6293	0.5989	0.5875
	precision	0.6958	0.6110	0.6051	0.6293	0.5989	0.5875
	recall	0.6958	0.6110	0.6051	0.6293	0.5989	0.5875
	F1-score	0.6853	0.5881	0.5756	0.6154	0.5735	0.5548

During the observations of the classifiers' performance metrics, it was identified that each classifiers performance was reduced with the n — gram features. Classifier's performance was higher when it was calculated with the unigrams. This behavior remains unchanged for both Bag of Words and TF — IDF feature extraction methods. And, when comparing the two feature extraction methods, accuracy of the Bag of words method was higher than TF — IDF method. There were five classifiers which were tested against the dataset. Out of these five classifiers, Support Vector Machine gives the best results with the accuracy level 0.744.

CHAPTER 5 - CONCLUSION AND FUTURE WORK

5.1 Introduction

This research is focused on below categories.

- Use of Automatic labelling over manual labelling for Sentiment prediction of Sinhala language
- Find the best supervised learning classifier for Sinhala language to predict election results
- Use Sinhala tweets to predict Presidential election results in Sri Lanka

7821 Sinhala datasets which were related to the main candidates in Sri Lankan presidential election was extracted from twitter. These tweets were labelled as Positive, Negative and Neutral by using the automatic approach with the help of already defined Sinhala sentiment lexicon. Labelled data was calculated to get an overall idea about each candidate's positive and negative tweets. Finally, each candidate's results were shown as a graphical representation. These results were evaluated and compared with the actual presidential election results in Sri Lanka which held on 2019. 54% for Gotabhaya and 46% for Sajith was found as the result of this research. It was concluded that the proposed approach provides almost accurate results when compared with the actual presidential election results.

Manual labelling was considered as a time consuming, effortful process. Due to these difficulties, automatic labelling process was introduced for Sinhala language during this study. Most of the existing researches (Chathuranga et al., 2019; Demotte et al., 2020; Iu, 2018) related to Sinhala language, were conducted with the help of annotators. So, there is a requirement to do a comparison study for manual labelling and automatic labelling. Sample of the original dataset was extracted and conducted the manual labelling for the selected sample dataset. Same sample dataset was considered for automatic labelling and did the comparison study with manual and automatic labelling. It was identified that the accuracy of the automatic labelling process was 33% and there was a necessity to improve the value. Emoji sentiment and negation handing methods were used to further improve the accuracy level of the automatic labelling. This has increased the accuracy level up to 50%.

Multiple supervised learning classifiers were used to find the best classifier for Sinhala language which can be used to predict election results. Model creation was done with the preprocessing, labelling, feature extraction and training dataset. Preprocessing was done by

removing unnecessary characters, non — Sinhala characters etc. Labelled data was used for the classification. Bag of Words, TF — IDF and N — gram feature extraction methods were applied during the classification and each classifiers' accuracy was measured. Support Vector Machine had better accuracy over other classifiers with a rate of 0.744.

5.2 Future Work

During this research, more focus was on the language independent feature extraction methods rather than language dependent methods due to the limitation of available tools for text classification in Sinhala language. In future, language dependent features such as POS tagging, syntactic dependency, stemming, opinion words, lemmatization, Word2Vec will be included.

It is better to find more methodologies to increase the accuracy of automatic labelling process. During this research, only syntactic level negation was handled but did not consider the morphological negations. So, another future focus of this research study will be to find more methodologies thus to increase the accuracy of automatic labelling process. It will be very useful for future researchers to use this automatic labelling process as the baseline of their study.

Supervised learning classifiers were taken into consideration in this study. Other classifiers such as unsupervised learning and lexicon based can also be used to do the comparison study with others and to find more appropriate classifier for Sinhala language.

In this study, dataset was limited due to the specific date range and the language. But many new features can be enabled during the text classification by increasing the dataset size. So, increasing the dataset will be taken into consideration during future implementations. And this dataset has both comments and news information related to candidates. This accuracy can be further increased by only considering the comments related to candidates.

During this study, only Sinhala data was considered for simplicity. But in Sri Lanka, people use Tamil and English languages to express their ideas through twitter. So, the same methodologies can be tested with the use of English and Tamil languages.

There are many social media platforms such as YouTube, Facebook, Instagram etc. In this research, only the twitter data was considered and extracted. As a further implementation, other platform can be also considered and applied the same methodologies to evaluate the accuracy of the developed model.

REFERENCES

1. Naive Bayes — scikit-learn 0.24.2 documentation [WWW Document], n.d. URL https://scikit-learn.org/stable/modules/naive_bayes.html (accessed 8.28.21).

2. What is the difference between the the Gaussian, Bernoulli, Multinomial and the regular Naive Bayes algorithms? - Quora [WWW Document], n.d. URL https://www.quora.com/What-is-the-difference-between-the-the-Gaussian-Bernoulli-Multinomial-and-the-regular-Naive-Bayes-algorithms (accessed 8.28.21).

3. Approaches to Data Labeling for Machine Learning Projects [WWW Document], n.d. URL https://lionbridge.ai/articles/5-approaches-to-data-labeling-for-machine-learning-projects/ (accessed 1.31.21).

4. A Brief Introduction to Supervised Learning | by Aidan Wilson | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590 (accessed 2.6.21).

5. Niklas Donges, 2021. A Complete Guide to the Random Forest Algorithm [WWW Document]. Built In. URL https://builtin.com/data-science/random-forest-algorithm (accessed 8.28.21).

6. A language processing tool for Sinhalese (සිංහල). | sinling [WWW Document], n.d. URL https://sinling.ysenarath.com/ (accessed 5.16.21).

7. A Simple Introduction to Natural Language Processing | by Dr. Michael J. Garbade | Becoming Human: Artificial Intelligence Magazine [WWW Document], n.d. URL https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32 (accessed 5.16.21).

8. About Train, Validation and Test Sets in Machine Learning | by Tarang Shah | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 (accessed 12.7.20).

9. Amali, H.M.A.I., Jayalal, S., 2020. Classification of Cyberbullying Sinhala Language Comments on Social Media, in: 2020 Moratuwa Engineering Research Conference (MERCon). Presented at the 2020 Moratuwa Engineering Research Conference (MERCon), pp. 266–271. https://doi.org/10.1109/MERCon50084.2020.9185209

10. Andrews, J., 2021. VaderSharp. The best sentiment analysis tool. In C#.

11. Sharma, A., 2020. Applications Of Natural Language Processing (NLP) [WWW Document]. Top 10 Applications of Natural Language Processing. URL https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/ (accessed 5.16.21).

12. Brownlee, J., 2020. Train-Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery. URL https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ (accessed 1.31.21).

13. Chathuranga, P.D.T., Lorensuhewa, S.A.S., Kalyani, M.A.L., 2019. Sinhala Sentiment Analysis using Corpus based Sentiment Lexicon, in: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer). Presented at the 2019

19th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 1–7. https://doi.org/10.1109/ICTer48817.2019.9023671

14. de Silva, N., 2020. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv:1906.02358 [cs].

15. Demotte, P., Senevirathne, L., Karunanayake, B., Munasinghe, U., Ranathunga, S., 2020. Sentiment Analysis of Sinhala News Comments using Sentence-State LSTM Networks, in: 2020 Moratuwa Engineering Research Conference (MERCon). Presented at the 2020 Moratuwa Engineering Research Conference (MERCon), pp. 283–288. https://doi.org/10.1109/MERCon50084.2020.9185327

16. Kemp, S., 2020. Digital 2020: Sri Lanka — DataReportal – Global Digital Insights [WWW Document]. URL https://datareportal.com/reports/digital-2020-sri-lanka (accessed 12.5.20).

17. Digital 2020: Sri Lanka [WWW Document], n.d. . DataReportal – Global Digital Insights. URL https://datareportal.com/reports/digital-2020-sri-lanka (accessed 6.26.21).

18. eiki, 2019. Feature Extraction in Natural Language Processing with Python. Medium. URL https://medium.com/@eiki1212/feature-extraction-in-natural-language-processing-with-python-59c7cdcaf064 (accessed 8.28.21).

19. Department of Knowledge Technologies, 2015. Emoji Sentiment Ranking v1.0 [WWW Document]. URL http://kt.ijs.si/data/Emoji_sentiment_ranking/ (accessed 8.23.21).

20. Everything There Is to Know about Sentiment Analysis [WWW Document], n.d. URL https://monkeylearn.com/sentiment-analysis/ (accessed 2.6.21).

21. Gandhi, R., 2018. Support Vector Machine — Introduction to Machine Learning Algorithms [WWW Document]. Medium. URL https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed 8.28.21).

22. Ganesan, K., 2019. What are Stop Words? Opinosis Analytics. URL https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/ (accessed 6.26.21).

23. Global Digital Overview — DataReportal – Global Digital Insights [WWW Document], n.d. URL https://datareportal.com/global-digital-overview (accessed 11.27.20).

24. Gupta, P., 2017. Decision Trees in Machine Learning [WWW Document]. Medium. URL https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 (accessed 8.28.21).

25. Gupta, Y., Kumar, P., 2019. Real-Time Sentiment Analysis of Tweets: A Case Study of Punjab Elections, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). Presented at the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–12. https://doi.org/10.1109/ICECCT.2019.8869203

27. How to Scrape Twitter Using Twitter Scraper | ScrapeHero Cloud, 2018. . ScrapeHero. URL https://www.scrapehero.com/how-to-scrape-historical-search-data-from-twitter/ (accessed 1.31.21).

28. Ingedata, n.d. Why Supervised Learning still often beats Unsupervised Learning? [WWW Document]. URL https://www.ingedata.net/blog/supervised-learning-vs-unsupervised-learning (accessed 12.7.20).

29. Introduction to Bag of Words, N-Gram and TF-IDF | Python, 2019. . AI ASPIRANT. URL https://aiaspirant.com/bag-of-words/ (accessed 8.28.21).

30. Iu, L., 2018. Sentiment analysis of Sinhala news comments.

31. Jayasuriya, P., Kumarasinghe, B., Ekanayake, S., Munasinghe, R., Thelijjagoda, S., Weerasinghe, I., 2020. Sentiment classification of Sinhala content in social media.

32. Jenarthanan, R., Senarath, Y., Thayasivam, U., 2019. ACTSEA: Annotated Corpus for Tamil Sinhala Emotion Analysis, in: 2019 Moratuwa Engineering Research Conference (MERCon). Presented at the 2019 Moratuwa Engineering Research Conference (MERCon), pp. 49–53. https://doi.org/10.1109/MERCon.2019.8818760

33. Jose, R., Chooralil, V.S., 2016. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach, in: 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). Presented at the 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 64–67. https://doi.org/10.1109/SAPIENCE.2016.7684133

34. Karunanayake, B., n.d. helasentilex: API for Sinhala Sentiment Lexicon.

35. Sai Patwardhan, 2021. KNN Algorithm. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/ (accessed 8.28.21).

36. Medagoda, N., Shanmuganathan, S., Whalley, J., 2015. Sentiment lexicon construction using SentiWordNet 3.0, in: 2015 11th International Conference on Natural Computation (ICNC). Presented at the 2015 11th International Conference on Natural Computation (ICNC), pp. 802–807. https://doi.org/10.1109/ICNC.2015.7378094

37. Nausheen, F., Begum, S.H., 2018. Sentiment analysis to predict election results using Python, in: 2018 2nd International Conference on Inventive Systems and Control (ICISC). Presented at the 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1259–1262. https://doi.org/10.1109/ICISC.2018.8399007

38. Lakmal, D, Ranathunga, S, Peramuna, S, Herath, I, 2021b. nlpcuom/Sinhala-Stopword-list. NLP Centre, University of Moratuwa.

39. Overview [WWW Document], n.d. URL https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview (accessed 5.16.21).

40. (PDF) ACTSEA: Annotated Corpus for Tamil & Sinhala Emotion Analysis [WWW Document], n.d. URL https://www.researchgate.net/publication/335494938_ACTSEA_Annotated_Corpus_for_Tamil_Sinhala_Emotion_Analysis (accessed 12.10.20).

41. (PDF) Sinhala Sentiment Analysis using Corpus based Sentiment Lexicon [WWW Document], n.d. URL https://www.researchgate.net/publication/337590795_Sinhala_Sentiment_Analysis_using_Corpus_based_Sentiment_Lexicon (accessed 12.10.20).

42. Mohammed Sunasra, 2017. Performance Metrics for Classification problems in Machine Learning [WWW Document]. URL https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b (accessed 6.26.21).

43. Sentiment Analysis: Concept, Analysis and Applications | by Shashank Gupta | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17 (accessed 2.6.21).

44. Sharma, P., Moh, T., 2016. Prediction of Indian election using sentiment analysis on Hindi Twitter, in: 2016 IEEE International Conference on Big Data (Big Data). Presented at the 2016 IEEE International Conference on Big Data (Big Data), pp. 1966–1971. https://doi.org/10.1109/BigData.2016.7840818

45. Sri Lanka Demographics Profile [WWW Document], n.d. URL https://www.indexmundi.com/sri_lanka/demographics_profile.html (accessed 12.5.20).

46. Stephanie, 2014. Cohen's Kappa Statistic [WWW Document]. Statistics How To. URL https://www.statisticshowto.com/cohens-kappa-statistic/ (accessed 6.26.21).

47. Brownlee, J., 2016. Supervised and Unsupervised Machine Learning Algorithms [WWW Document]. URL https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ (accessed 2.6.21).

48. Supervised vs Unsupervised Learning: Key Differences [WWW Document], n.d. URL https://www.guru99.com/supervised-vs-unsupervised-learning.html (accessed 12.5.20).

49. John Joseph, F.J., 2019. Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree. pp. 50–53. https://doi.org/10.1109/INCIT.2019.8911975

50. Audrey Schnell, 2020. What is Kappa and How Does It Measure Inter-rater Reliability? The Analysis Factor. URL https://www.theanalysisfactor.com/kappa-measures-inter-rater-reliability/ (accessed 6.26.21).

51. What is Supervised Learning? [WWW Document], n.d. URL https://searchenterpriseai.techtarget.com/definition/supervised-learning (accessed 2.6.21).

52. IBM Cloud Education, 2020. What is Supervised Learning? | IBM [WWW Document]. URL https://www.ibm.com/cloud/learn/supervised-learning (accessed 9.7.21).

53. Election Commission of Sri Lanka, n.d. Election Commission [WWW Document]. URL https://elections.gov.lk/en/elections/results_pre_E.html (accessed 6.26.21).

Emoji Sentiment Ranking v1.0 [WWW Document], n.d. URL

54. Shubham Jain, 2018. NLP For Beginners | Text Classification Using TextBlob. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/ (accessed 9.12.21).