Identify Hateful Comments in Sinhala Language on Social Media

W.W.E.N. Fernando
189461H

Faculty of Information Technology
University of Moratuwa

July 2021

Identify Hateful Comments in Sinhala Language on Social Media

W.W.E.N. Fernando
189461H

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

July 2021

Declaration

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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student
W.W.E.N. Fernando

Signature of Student

Date

Supervised by
Name of Supervisor
Mr. S. C. Premaratne

Signature of Supervisor

Date

Acknowledgements

I am so grateful to my supervisor Mr. Saminda Premaratne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for academic guidance, advice, time, encouragement and patience that has seen me through, making this work success.

Also I would like to give a special thanks to the Prof. Rathnasiri Arangala, Senior Professor in Sinhala, University of Sri Jayewardenepura for his guidance, advice, time and encouragement in carrying out this research. As well as special thanks to Mr. K.K. Premarathne, Former Principle, Bandaragama National College for his precious support as an annotator and his encouragement.

Furthermore, my thanks also go out to my beloved family members for their support, love, advice and encouragement throughout the period of my study. I lastly wish to extend my sincere gratitude toward my colleagues who have tirelessly guided and encouraged me at all times through my course.

Abstract

In present, the spread of hate speech through social media has become a very serious problem, both globally and locally. The route cause for this is the increasing use of social media with the rapid expansion of computer science and information technology. Therefore, it is very important to use same to control this kind of situations. Although there is a mechanism in place on social media to automatically control such hate speech in English language, but it is still not seen in Sinhala Language. The reason for this is the lack of knowledge about the native languages such as Sinhala in the social media service providers. Therefore, the identification of hateful contents in Sinhala language is an urgent and vital task that needs to be addressed.

This research propose lexicon based and machine learning based approaches for the automatic identification of hateful speech in Sinhala on social media. With different pre-processing techniques and machine learning algorithms, machine learning algorithm based approach was conducted with four different approaches. These approaches were begun with 3000 comments which is equally divided into hateful and non-hateful. Using these comments, it was able to identify the most appropriate featured groups and model to identify the hateful speech in Sinhala language on social media.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prolegomena

As per the global and local interconnectivity among people is rapidly growing through internet, social networking sites has become more and more popular than expected in the present. In this kind of situation, there can be positive as well as negative impacts on society. So it's really important to control negative impacts of social media as every social media already has their own policies regarding to control this kind of situations. When we go through most popular social networking sites like Facebook, YouTube, Instagram, Twitter and etc. "Commenting" is commonly used to express their opinions in front of others ideas. When users made comments those may be intentionally or unintentionally, hateful, abusive or insulting for another user. As well as it also indirectly influenced to social crimes.

So, growth of this kind of harmful commenting is a common issue in present social networking sites. Therefore, an accurate and efficient method for control this kind of comments in any language is really important and also it will be a great relief to the society.

In that case many researches' attention is focused on this area. Since many of them are based on the English language, it is important to continue to this area of research in our native language, as the use of Sinhala language in social media is now gaining popularity.

Sinhala Unicode is currently very popular among social media users in Sri Lanka, but there is no accurate and efficient mechanism to control comments made in Sinhala language. This is because service providers unable to implement a control mechanism until someone reports on the content of such comments. Therefore, predicting harmful comments in accurate and efficient manner can be identified as a social responsibility in present.

Therefore, this research study focuses on to identify accurate and efficient model to predict harmful comments which are written in the Sinhala language on social media by using comments made in Sinhala Unicode on Facebook.

1.2 Background and Motivation

As per the rapidly growing of technology, online social networks(OSNs) become more and more popular in these days and the virtual civilization become a reality with this wealth. Then the social impact of OSNs usage also increased. As a result, today's social media users are able to publish anything with none management or restrictions on the content, which increases the spread of hate and offensive speech amid the users. At the same time, in many countries the government and the relevant companies or organizations have enacted laws and regulations to prevent the spread of such hatred.

When we narrow down the focus on to the countries that facilitate users to use social media networks with their native language, we can see that some users are embarrassed by the way some others publish their thoughts in their native language without restraint. On the other hand, due to lack of knowledge of indigenous languages, relevant organizations fail to control such situations. Therefore, countries that allow users to use social media networks with their native language need a formal procedure to control this kind of situations.

Considering such incidents, an incident related to Sri Lanka in the year 2018 can be cited [61]. In there, Sri Lanka was informed about the hate speech to the Facebook, but the Facebook has pointed out that they have failed to remove it. This shows that there is a problem with the analysis of the native languages of the institute and that even Sri Lanka does not have a proper mechanism to control such matters. In the discussion that followed the incident, it was identified that the best solution to such problems was to create a dictionary containing hateful Sinhala words. But such a dictionary is still not found in Sri Lanka.

Considering all of the above factors, this research focuses on identifying the unfold of hate speech in Sinhala language on social media. I hope to do this study by creating and using a dictionary of hateful Sinhala words as well as machine learning techniques. Among many of social media platforms, Facebook has become very popular and it has allowed to create big data on society. Therefore, this research will conduct with Facebook data.

1.3 Problem Statement

Expressing one's opinion using Sinhala language is becoming more and more popular on social media as Sinhala language is that the linguistic communication of Sri Lankans. With the support of Sinhala Unicode, it has become very easy in these days. And so on among the popular social media, Facebook has become very popular in Sri Lanka. Also, they hope to share and connect more freely when using social media. But as a result, the rapidly growing use of social media has both positive and negative effects on society. So it is very important to identify such kind of negative impacts on society, more accurately and efficiently.

1.4 Aim and Objectives

1.4.1 Aim

The aim of this research is to identify accurate and efficient model to predict harmful comments expressed in the Sinhala language on social media.

1.4.2 Objectives

- To critically evaluate the underlying literature regarding hate speech identification in Sinhala language and other languages.
- To develop in-depth knowledge on the hate speech detection techniques that are currently using or new concepts for the identification of harmful written contents on social media.
- To preprocess the comments which are made using Sinhala Unicode on Facebook that are gathered from “Kaggle” which is a web community of information scientists and machine learning professionals.
- To create a dictionary containing hateful Sinhala words and build a model to identify harmful ideas on social media using it.
- Validation of built dictionary which is containing hateful Sinhala words by evaluating the performance of model.
- To create a machine learning based models and identify harmful ideas on social media using those.

- To identify best model to predict harmful comments that state in the Sinhala language on social media platforms.

1.5 Proposed Solution

In this research I proposed to use labeled Sinhala Unicode based data set to identify hateful comments in Sinhala language on social media. It may classify social media comments as positive or negative based on classification model. Models will be built using newly created dictionary which contain negative words and classification techniques. Those built-in models can be used to identify hate speech in Sinhala on social media. It will help to minimize the negative impact that social media has on society.

1.6 Structure of the theses

The content of the overall thesis is structured as follows.

The First Chapter includes an introduction of project with the background, problem, aim and objectives and the proposed solution. The second chapter critically reviews the dictionary based approaches and approaches which are based on machine learning based methods to the identification of hateful contents on social media. The Third Chapter is for adopted technologies and the Forth Chapter presents the approach inputs, methodology and outputs with performance measurements. The Fifth Chapter is the design of the solution and the implementation of the salutation presents in the Sixth Chapter. Evaluating the solution is described in Chapter Seven. Finally, Chapter Eight concludes the solution with a note of further work.

Chapter 2

Literature Review

2.1 Introduction

“Social media are interactive computer-mediated technologies that facilitate the creation or sharing of information, ideas, career interests and other forms of expression via virtual communities and networks” [1][2]. As well as a result of availability of social media on mobile devices such as smart phones and tablets it inserted into people's lives a kind of virtuality, resulting in people and social media becoming closer. Since then, people have started using this platform to express their knowledge, thinking, belief, feelings, emotions and opinions. Social media platforms usually used in social communication, such as Facebook, YouTube, Instagram, Twitter and etc. form the basis for this research study.

When we consider in globally these platforms create priceless opportunities to a country. But misuse this kind of common platform can be made some negative impact on society. Therefore, while getting benefits from this kind of media, it is important for a country to have some control over it. The society can have different minded people. So it's really difficult to define which contents are harmful to people with different mind levels. But consider about some definitions such as hate speech, verbal abuse, insulting and etc. we can get some idea about harmful contents.

Cambridge Dictionary was defined hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation" [3]. “American Heritage Dictionary” of the English Language defined verbal abuse as “the act of forcefully criticizing, insulting, or denouncing another person” [4]. And also The Verbally Abusive Relationship it characterized by “underlying anger and hostility, it is a destructive form of communication intended to harm the self-concept of the other person and produce negative emotions” [5]. As well as Erving Goffman stated an insult as “an expression or statement (or sometimes behavior) which is disrespectful or scornful. Insults may be intentional or accidental” [6].

As a result of Unicode support, social networks allow users to share their thoughts among each other in their native language such as Sinhala, Tamil, Hindi and etc. Then belong community becomes larger and most social status got opportunity to connect with this network. And we all know everyone has right to express their own thoughts. But it should not be harmful to a person or group.

Particularly when using a common platform, words or phrases used to express their views should not harm anyone. In other words, the way users use social media to express their opinions in front of others ideas should not include any harmful content. Therefore, according to the rules and regulations regarding human rights in a countries and well defined polices of social medias, we can make an effort to control these.

As well as a social study [9] done in 2014 has been assimilate how important is the automatic identification of harmful speech on social media platforms. Also it dispenses how some persons in Sri Lanka used “Facebook” to unfold hate.

Therefore, in view of all the above, this research study focuses on providing a solution to make the prediction of publishing harmful comments in Sinhala language which are a major problem on social media websites.

2.2 Related work for Social Media Text Mining for Identify Objectionable Content

Social media become very popular communication media in present as it supports several non-international languages with the support of Unicode encoding. “According to the Digital 2020 Report [13], Social media users in Sri Lanka increased by 491 thousand (+8.3%) between April 2019 and January 2020 and there were 6.40 million social media users in Sri Lanka in January 2020”. But the same time, some people use it to spread harmful contents on others in native language. Similarly, Sri Lankans are increasingly biased to use the Sinhala Language on social media to spread objectionable content towards others. This has become a huge social problem today because service provides do not consider media content until someone report it. Even so, there may be times when they do not recognize them as objectionable content due to their lack of Sinhala language proficiency.

Along with the above, social studies have also been made on this issue, which discusses the spread of harmful content in Sinhala Language using social media in Sri Lanka as well as importance of identifying them [9].

As describes in “Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning” [15], Text rational model with Support Vector Machine (SVM) was built to identify racialism based social media statements made in Sinhala with the purpose of solving the lack of human interpreters in native languages. They pre-processed data by removing “stop words”, “numbers”, “special characters”, “duplicate characters”, “URLs” and “email addresses” from comments. And then “n-grams features” have been extracted with applying various information matrices. Afterwards performances were measured using “accuracy”, “precision”, “recall”, “F1-score” and “Receiver Operating Characteristic (ROC) curve”. The trained model was classified racist based comments with 70.8% accuracy and 100% precision. It has been shown by paying attention to the curves, dual-class SVM classifier was out perform over classifier with no power. Then the researchers have been thought the data amount was insufficient, so they increased the amount of data but then they got low accuracy value 57.6% while the precision value 100%.

As mentioned in “Sinhala Hate Speech Detection in Social Media using Text Mining and Machine Learning” research [17], Sandaruwan, Lorensuhewa and Kalyani has been conducted the research recently through two main experiments to automatically determine hateful and insulting speeches made in Sinhala language in social media. Those are the lexicon based method and Machine learning algorithms based method. For the lexicon based method they have used translation of “Google bad word list” which contains forbidden English terms by the “Google” as the source term list. Hate speech detection text pre-processing has been done with the sub process of “Removing of non-Sinhala characters”, “Removing of stop words” and “Stemming”. Thereupon the features have extracted considering for features named “Bag of Words”, “Word n-gram”, “Character n-gram” and “skip-gram feature”. After that the extracted features have been vectorized and weighted with methods called “Count Vectoriser” and “TF-IDF transformer”. Then before apply statistical methods, Finally, under supervised machine learning algorithms, classification has been performed with three different algorithms. Those are “SVM”, “Multinomial Naïve Bayes(MNB)” and “Random Forest Tree(RFDT)”. As well as performance has been evaluated with four measurements call “Accuracy”, “Precision”, “Recall” and “F1-score”. Out of two experiments have done Lexicon corpus based experiment given 76.3% of accuracy and out of main three experiments, character trigram with MNB given 92.33% accuracy with 0.84 recall. Therefore, researchers have been concluded that character trigram and

MNB as the best approach to automatically discover hateful and insulting speeches that shared along Social media in Sinhala language.

As well as Sandaruwan, Lorensuhewa and Kalyani [19] recently conducted a research to identify best feature group and classification models that can be used to identify of Sinhala Comments with abusive meaning in Social Media. Since the lack of resources for computer-based natural processing, they conducted a research to automatically identify abusive comments which are made in Sinhala language. They were trained three models with MNB, SVM and RFDT and the features were extracted in four different ways. Out of those trained models MNB with tri-gram and four-gram have been given highest accuracy and recall. Since many insulting contents are written with typos and substitute with similar letters they concluded that character n-gram feature perform well in detection of Sinhala abusive contents. Also they used "cross-lingual lexicon approach" and "corpus-based lexicon approaches" to detect Sinhala insulting contents. Either way cross-lingual lexicon approach gave the highest accuracy.

The study [20] which was conducted by Amali and Jayalal identified that SVM with Radial basis function (RBF) kernel perform better when classified cyberbullying Sinhala language comments on social media. They designed a hybrid text analytic model using rules and machine learning methods. With the expert decision researchers have been produced five different rules for identify cyberbullying Sinhala text. Then the patterns were learnt by feeding feature into machine learning algorithms. They also found that the accuracy of the model depended on the size of the corpus. They realized that by increasing the amount of data set increases the accuracy of the models as well.

Recently, S. W. A. M. D. Samarasinghe et al. [14], proposed a deep learning mechanism with two Convolution Neural Networks (CNN). In there, they first classified a given text corpse as hateful or not and then re-classified it according to its hate level if it contains hate content. They were conducted their research with 8000 user comments written in Sinhala Unicode on gossip sites which are extracted using a web scraper. As same as the above researches they have preprocessed the data set and convert it into vectors. Then with those vectors train the model using an improved CNN and SVM. Finally, they have calculated Accuracy and F1 score. Since their data set was not evenly distributed across classes, "Area Under Curve (AUC)" technique was also used in the experiment. According to the accuracy value, CNN achieved better performance when detecting hate speech and SVM achieved better performance when hate level classification. When considered about F1 scores, CNN achieved better F1 score

compared with SVM. According to the AUC, CNN have been performed better than SVM.

Hajung Sohn and Hyunju Lee [21] proposed a "multi-channel model" with three versions of Bidirectional Encoder Representations from Transformers (BERT), the "English", "Chinese", and "multilingual BERTs" for hateful content detection. Similarly, they have pre-processed labeled data set and they have translated source language into English language and Chinese language by using Google API, with the purpose of creation of multi-channel BERT (MCBERT) model for different languages. Then the results of the models evaluated with accuracy and F1 macro score. They have used three datasets from different languages with three versions of BERT and MCBERT for their study. For their first data set, the best model among different BERT model was the MCBERT. But English and multilingual BERTs were also showed similar performance to the MCBERT. For the second data set, the highest accuracy and F1 score were found with English BERT and MCBERT and lowest accuracy was found with Chinese model. Finally, for the third data set multilingual BERT model showed highest accuracy and F1 score. But other English, Chinese, and MCBERT were also performed similar performance. They concluded that MCBERT, main model with translations as well as state-of-art models performed well over the hate speech detection.

The above facts show that due to the lack of research done on the Sinhala language, it is not possible to come to a definite conclusion from it. As a result, this literary review needs to be further expanded. It is also identified that to detect hate speech by translating native languages into resource-rich English language is very rare. But it was also identified as an area that could be explored. Therefore, it is expected that this study will be done one experiment with the Sinhala Language Corpus and in parallel another experiment with translation of the same database into English Language.

Table 2.1: Summary of Feature Extraction Techniques

Feature	References
Bag-of-words (BoW)	[17], [19], [30], **[38]**, [50], **[55]**
Word N-gram	[17], [19], [27], **[29], [31], [32],** [34], **[44],** [45], **[46]**, [51], [54], [57]
Character N-gram	**[17]**, **[19]**, **[27]**, [29], **[44], [45]**, [46]

Word skip-gram	[17], [19]
Syntactic Feature	[45]
Negative Sentiment Feature	[45]

The above table 1, provides a summary of the feature extraction techniques used in some of the past studies. The bold color indicates the feature extraction techniques that used when it's achieve high accuracy. Most of the research have not been mentioned feature extraction technique that they used when obtaining its high accuracy. But with the existing data, we can identify that most of the time get best accuracy when they used “Word N-gram” or “Character N-gram” feature. Based on that for this research it will be use “Word N-gram” and “Character N-gram” as feature group extraction technique.

Table 2.2: Summary of Feature Vectorization Techniques

Feature	References
TF-IDF	[17], [19], **[25]**, [30], **[31]**, **[32]**, **[34]**, **[35]**, **[36]**, **[37]**, **[38]**, **[39]**, **[43]**, **[44]**, **[45]**, **[47]**, [48], **[49]**, **[50]**, **[51]**, **[56]**
Information Gain	**[47]**
Count Vectorizer	[17], [19], **[25]**, [31], [43], [50]

Before apply a classification methods vectorization is an important step that we have to conduct in text classification process. From the above table 2, can be identified that in most of the cases researchers used TF-IDF method to weighted their term features. Therefore, in general, we can conclude that TF-IDF performs vital role within the process of text classification. The bold color indicates the feature vectorization techniques that used when it's achieve high accuracy. Similarly, when we consider about feature vectorization method, most of the research have not been mentioned feature extraction technique that they used when obtaining its high accuracy. But since a considerable amount of data has been extracted here, TF-IDF will be used as feature vectorization method in this research.

In a related study, K. M. Hana et al. [36] the SVM model out performed over the Convolution Neural Network (CNN) model. In their study, several experiments were conducted to find the better method for multi-labeled categories on Tweets with hateful contents. They have concluded that categorization using SVM model and Classifier

Chains(CC) with the datum without stemming, take off stop word, and translation is the better way to detect hateful speech in Twitter. It further explained, SVM model with Classifier Chains gives best outcome over other models in their study. The Study of R. Martins et al. [41], describe the use of SVM to predict hateful contents contained in a text, using an emotional approach though sentimental analysis. They have used SVM, Naive Bayes and Random Forest for text classification. But they got highest accuracy with SVM.

Similarly, in study, H Sahi et al. [51], designed a Linear SVM classification to identify hateful contents against female in Twitter. They have been collected tweets written in Turkish, and then five matching learning based classification algorithms (SVM (using polynomial and RBF Kernel), J48, Naïve Bayes, Random Forest and Random Tree) were applied with those tweets. Finally, they found that Linear SVM performed significantly better than other models. As in previous studies, they tested the SVM algorithm with other algorithms in [44], [46], [49]. Even then, they got the best results with SVM.

In a related research, Shovon Ahammed et al. [25] design a SVM classification model and NB classification model to detect hateful speeches in Bangla Language. The developed SVM model is a binary classification which is having ability to predict whether a content in Bangla language is hateful or not. In their study they prepared a new data corpus in Bangla language and divided it into two sets and tagged them. Then the features were extracted from their datum to use it for create a model. Then applied the SVM and NB machine learning algorithms and NB gave them highest accuracy. Similarly, in study [26], NB was used to classify hateful tweets with political motive. In their study, the dataset was categorized into three sets: “non-political hate speech”, “hate speech with political motive” and “non-hate speech”. Then it gives high accuracy for hate speech detection with political motive with NB algorithm.

We can found that commonly used event models in Naïve Bayes classification. Those are “Multivariate Bernoulli Event Model” and “Multivariate Event Model (Multinomial Naive Bayes)”. Out of those two Multinomial Naïve Bayes(MNB) is a classification method design for text. And also its generally better and faster than plain NB. The research of Sandaruwan et al. [17] designed a MNB classifier optimized by n-grams features model for hateful content detection in Sinhala language on social media. The evaluation results of the developed MNB based classification model showed better performances than other machine learning approaches.

Similarly, in study of Ruwandika and Weerasinghe [50], they have used five different models to detect hateful contents. The specialty of this study is the use of supervised and non-supervised machine learning methods. They have used SVM, LR, NB algorithm, Decision Tree algorithm as supervised learning model and K-Means clustering algorithm as unsupervised learning model. Finally, by comparing the accuracy values and F-scores, the most appropriate classifier for the function of hateful speech identification was explored. Then the Naïve Bayes model was identified as best out of all models. As well as in studies [33] and [56], NB performed well against other machine learning algorithms.

In the study of I. Alfina et al [29], a model based on RFDT classifier with word n-gram feature representation, was proposed to detect whether the tweets are hatful or not on Twitter. They have conducted their study with manually annotated tweets with balance data set. Contrary to previous researchers' claims that RFDT, BLR and SVM have the similar performance in detecting hateful contents, SVM performance has been found to be significantly lower than RFDT and BLR. Based on the experimental results, they concluded that, higher F-measure was obtained when the word n-gram was used, especially when combined with RFDT.

According to the R. Hendrawan's experiments [37], RFDT classification method with classifier chains as data transformation and TF-IDF as feature extraction gives the best results. That experiment was performed best without translation, without stemming, and without stop word removal. Most of the researchers uses preprocessing techniques for increase accuracy of their research. But here they have found that they provide high accuracy without preprocessing with RFDT. In similar study conducted by K Nugroho et al. [42] also shows that RFDT has a better level of "accuracy", "precision", "recall" and "F1-score" compared to the other methods that they have tested.

The study of M. Sajjad et al. [52] presented a LR with Convolution Neural Network (CNN), Glove Embedding and Baseline Feature. In here Glove means an unsupervised learning algorithm for deriving vector representation for terms. They have first initialized embedding gloves and random weight embedding. Then they trained them on "CNN" and "Long Short-Term Memory (LSTM)". According to its reputation in previous studies, they were chosen CNN and LSTM. After that they have extracted other fundamental features. Finally, they trained and tested a some classifiers using those features with a spotlight on LR, RF and SVM. All the proposed methods have

given good results and out of them "CNN + Glove Embedding + Baseline Features+ Logistic Regression" method has given the best result.

In the research of G. Koushik et al. [38] suggested a LR classification model for the hateful content detection with support of Bag of Word (BoW) and the TF-IDF approaches. In this study, too, the LR achieved considerable high accuracy. But in this research they haven't compared with different classification models. Therefore, it can be given more accurate performance with different classifiers and by incorporating the linguistic features. But in similar study conducted by N. Rai et al. [47] compared the LR model with NB and SVM with feature selection methods i.e. Information gain and TF-IDF. In this study selected features are then passed through GridSearchCV and Pipeline approach and finally machine learning approaches are applied. Then the LR classifier has outperformed over the other classification methods.

Similarly, in study of D. Elisabeth et al. [32] presents a design to detect hateful tweets written in Indonesian with LR, NB and RFDT algorithms. And also they have used "TF-IDF" as feature vectorization method and selected "word bigram" as feature extraction technique. In this study they have used two structures: "hate code detection from hate speech classification" and "hate code detection from hate code classification". From that, they have found that detect hatful contents through "hate code classification" is the best method out of their two scenarios. And also found that the LR with TF-IDF and word bigram feature is outperformed with other classifiers. In study [35] which was conducted by P.S.B. Ginting, found that "Multinomial Logistic Regression (MLR)" is an acceptable categorization way for text classification. MLR is a classifier which generalizes the logistic regression to classification problems where the output can take more than two possible values.

The below table 3, provides a summary of the traditional machine learning techniques that used in past studies for hate speech classification. From, that we can identify that most of them are used SVM, NB, RFDT or LR with their research study and also their accuracy has been varied. But it can be clearly identified that SVM, NB, RFDT and LR out performed over KNN, J48 and K-Means. But it can identify that Multinomial Naïve Bayes (MNB) outperformed over other machine learning techniques in study [17 and [19] which are based on Sinhala language. As well as according to the [53], it can be identified that MNB is generally use for text classification. Therefore, based on that, in this study it will test the accuracy of hate speech detection over dataset with SVM and MNB. As well as it also hopes to study a dictionary-based approach as a new

approach in this study. In here, it will not use translation of Google bad word list as lexicon, as previous done lexicon based studies which are related with Sinhala Language. Instead, I hope to use the training data set to create a dictionary of bad words and use it to identify hate speech.

Table 2.3: Summary of Machine Learning Techniques used in Hate Speech Detection

Machine Learning Approach	**References**
Support Vector Machine (SVM)	[17], [25], [27], [28], [29], [31], [34], [36], [41], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [54], [55], [56]
Naïve Bayes (NB)	[25], [26], [27], [29], [32], [33], [41], [43], [44], [46], [47], [49], [50], [51], [54], [56], [57]
Multinomial Naïve Bayes (MNB)	[17], [19]
K-Nearest Neighbor (KNN)	[27], [43], [49]
Random Forest Decision Tree(RFDT)	[17], [27], [29], [32], [37], [40], [41], [42], [43], [45], [46], [48], [51], [52], [54]
J48 Tree	[27], [51]
Logistic Regression (LR)	[27], [28], [29], [31], [32], [35], [38], [40], [43], [45], [47], [48], [49], [50], [52]
K-Means	[50]

Almost every selected paper has used at least one method of precision, recall, F1-score, accuracy and AUC to evaluate performance. From the below table 3, we can determine that most of the traditional machine learning performed well over txt classification. But based on the data set, it's gives different values for a same machine learning techniques as well. Therefore, in this study will use "accuracy", "precision", "recall", "F1-score" measurements for evaluate the models.

Table 2.4: Performance Evaluation Summary of surveyed Machine Learning Techniques

Reference	Machine Learning Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)	AUC Value
[17]	MNB	92.33%	95%	84%	89%	N/A
[19]	MNB	96.5%	96%	96%	96%	N/A
[25]	NB	72%	70%	74%	72%	N/A
[26]	NB	93.22%	N/A	N/A	N/A	N/A
[28]	RNN	79%	76%	78%	77%	0.84
[29]	RFDT	82.60%	N/A	N/A	N/A	N/A
[32]	LR	94.44%	94.47%	96.94%	94.90%	N/A
[33]	NB	93%	N/A	N/A	N/A	N/A
[35]	LR	87.68%	80.02%	82%	N/A	N/A
[37]	RFDT	76.12%	N/A	N/A	N/A	N/A
[38]	LR	94.11%	N/A	N/A	N/A	N/A
[41]	SVM	80.56%	76.80%	73.60%	N/A	N/A
[42]	RFDT	72.20%	71.10%	72.20%	71.30%	N/A
[43]	SVM	82.50%	N/A	N/A	N/A	N/A
[44]	SVM	90.85%	N/A	N/A	N/A	N/A
[46]	SVM	68.43%	N/A	N/A	N/A	N/A
[47]	LR	83%	N/A	N/A	N/A	N/A
[48]	BERT	74%	64%	66%	63%	N/A
[49]	SVM	87.07%	N/A	N/A	N/A	N/A
[50]	NB	73.90%	75%	73.90%	71.90%	N/A
[51]	SVM	72.50%	100%	45%	62%	N/A
[52]	CNN + Glove + Baseline Features + LR	N/A	98.40%	96.50%	97.40%	N/A
[54]	GRU	N/A	N/A	N/A	75.52%	N/A
[56]	NB	92.20%	89.90%	96.40%	92.40%	N/A

2.3 Objectionable Content Identification

When we narrow down our focus into Facebook, they do not allow objectionable contents. Under objectionable content they define “Hate speech”, “Violent and graphic content”, “Adult nudity and sexual activity”, “Sexual solicitation” and “Cruel and insensitive” [7]. Out of these community standards of Facebook, definition of hate speech and cruel and insensitive content are related when consider harmful contents.

Facebook defines hate speech as a “direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability”. As well as they define cruel and insensitive content as “contents that targets victims of serious physical or emotional harm” [7].

When consider their procedure, they continue their work from defining boundaries of hate speech to removing those. But when they concluding what left on the social media and what to take off, they consider contexture, intent, mistakes, continuing to improve, and etc. based on English language [9].

2.4 Tools Available for Sinhala Language

[10] Describes there are significant number of implementations regarding NLP for Sinhala Language but it becomes a resource-deficient language as they are not available for further use and research of researchers. [11] Further the paper says that out of six most using free software tools (RapidMiner, R, Weka, KNIME, Orange, and scikit-learn) implemented for common data mining tasks there is no specified tool since each tool has its pros and cons. Even so, by considering characteristics for a fully-functional Data Mining, they recommend Rapid Miner, R, Weka, and KNIME for data mining tasks.

Furtherly [12] says out of five open-source Data Mining tools (Weka, RapidMiner, KEEL, Orange and Tanagra) that impart numerous learning methods can be applied for predicting, "Weka" and "RapidMiner" recommended for most of Data Mining tasks, because those who prefer a fully functional and flexible platform have that feature.

2.5 Summary

As reported in the relevant literature, use of Text categorization with traditional techniques can be identified as most performable area to classify text content in social media. Under dictionary based approach we can identify translation of "Google bad word list" was used as base for the dictionary. As well as for a machine learning techniques based text classification process we can mainly identify five different sub process and those are "Text pre-processing", "Feature extraction", "Feature Vectorization", "Construct Model" and "Performance Evaluation".

As per concluded earlier on relevant literature, this will be conduct with two approaches. One will be lexicon based method and the other will be machine learning techniques based method. Therefore, there will be two main experiments as "Experiment A" and "Experiment B" in this study. For the Experiment A, bad word dictionary will be created by using training data set and under Experiment B, it will be conduct pre-processing, feature extraction, feature vectorization and construct model. Performance will be evaluated with each model in both experiments.

For the feature extraction of this study "Word N-gram" and "Character N-gram" technique will be used as concluded according to the above literature. Similarly, TF-IDF will be used as the feature vectorization method. And then dictionary based model and SVM and MNB machine learning techniques based models will be built. Finally built models will evaluate with accuracy, precision, recall, F1-score values. According to those values best fit model will be identified.

Chapter 3

Adopted Technologies

3.1 Introduction

In the previous chapters we discussed about different findings related to the hate speech identification area and defined the research problem. And also identified different stages which are going to be perform to identify hateful comments which are made in Sinhala language on social media. This chapter explains how to use selected technologies that are differentiated from the technologies used in the available writings.

3.2 Text Mining Techniques

It is a process that used to identify new information and predict future trends by extracting and discovering patterns and relationships in a large dataset. There are several techniques used in text mining. Some of those techniques are classification, Clustering, Association rules, Regression and etc. In this research classification will be used in one experiment as data mining technique for identify hateful comments. "Classification is a procedure of detecting a model that separates data classes based on previously categorized data. When we classify documents into pre-defined categories based on their content, it is called text classification. MNB and SVM will be used as text categorization algorithms in this study.

3.3 Multinomial Naïve Bayes

According to the Ayo et al. [62], Naïve Bayes defined as “probabilistic machine learning method that model conditional dependences of events in the form of a directed edges graph using Bayes theorem”. Considering the uses of NB, it can be used for a variety of purposes such as “prediction”, “classification”, and “diagnosis”. Although it is a simple algorithm, it performs well in most text classification problems. Other than that less training time and less training data also can be identified as advantages of NB classifier. In other words, this approach’s CPU and Memory consumption is less than

other approaches. As with other machine learning methods, we must have a labeled database before performing the classification task.

We can found that commonly used event models in Naïve Bayes classification. Those are "Multivariate Bernoulli Event Model" and "Multivariate Event Model (Multinomial Naive Bayes)". Out of those two Multinomial Naïve Bayes(MNB) is a classification method design for text. And also its generally better and faster than plain NB.

3.4 Support Vector Machine

According to the Ayo et al. [62], SVM defined as "supervised machine learning method to solve problems of binary classification in the absence of a suitable statistical solution". That method was proposed by proposed by Cortes and Vapnik. When we think about text classification task, SVM classify new text in to pre-defined classes. In the case of hate speech identification, the SVM needs both hateful and not hateful training set to explore the decision surface. Since this algorithm have higher speed and better performance with a limited number of samples, it is more commonly used in text classification than complex methods.

3.5 Rapid Miner Studio

There are lot of machine tools such as Rapid Miner, Weka and etc. available for data classifications. Since Rapid Miner Studio software provides user friendly GUI environment [12] with many statistical graphs and summarize data, it was selected as the tool of this research with Weka extension.

According to the [24], Rapid Miner defined as "a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics". Also it provides a Graphical User Interface to designing and implementing analytic process. Rapid Miner is written in Java Programming Language and can be extended using "R" and "Python" programming languages. It provides more extensions for "data preparation", "machine learning", "deep learning", "text mining", and "predictive analytics". Those extensions can be download and share for Rapid Miner from the Rapid Miner Marketplace.

3.5.1 Text Processing

The Text Processing extension includes all operators required for statistical text analysis and Natural Language Processing (NLP). It provides standard operators for “tokenization”, “stemming”, s”top word filtering”, and “n-gram generation” to text preparing and analysis [58]. In this study for Experiment A, it will use Tokenize, Filter Tokens (by content), and Dictionary Based Sentiment (Documents) and for Experiment B, it will use Process Document from data, Tokenize, Filter Tokens (by content), Filter Stop words (Dictionary), Generate n-Grams (Character), and Generate n-grams (Term) for text classification.

3.5.2 Weka

With this extension, all modelling methods and attribute evaluation methods rom the Weka machine learning library are available within Rapid Miner [58]. In other words, with the Weka extension Rapid Miner can extend with everything possible with Weka while keeping the full analysis, preprocessing, and visualization power of Rapid Miner [58]. In this study “W-NaïveBayesMultinomial” class will use for built models in Experiment B.

3.5.3 Operator Toolbox

This extension brings together some useful additional operators such as Attribute Generation, Merge Attributes, Append (Superset), Group Into Collection, and etc. They range from utility operators to improve the flexibility and usability of the process design, over additional outlier detection algorithm and additional performance criteria to advanced analysis methods [58]. In this study, “Stem Tokens using Example Set” will use for built Sinhala tem dictionary for Sinhala Unicode based text classification process.

3.6 Summary

This section represent the technologies and tools used. There are many other Technologies that can be used for do the same. In next chapter, it describes the approach that used to identify hateful comments in Sinhala Language on social media.

Chapter 4

Research Methodology

4.1 Introduction

Third chapter described about the techniques that used in this study. This chapter present the approach that applied with problem of identifying hateful speech in Sinhala language on social media. Here it presents the approach by highlighting the hypothesis, input, output, and process.

4.2 Hypothesis

We hypothesis that the identification of hateful speech spread on social media in Sinhala language, which is a problem without a proper methodology, can be achieved through the use of classification analysis. For that we are going to use dictionary based classification and machine learning techniques based classification methods. Under machine learning based classification NB and SVM will be used as algorithms and also it is expected to compare the effect of stop words removal and stemming, when identifying hate speech. Finally, we are going to pick up the most accurate classifying model.

4.3 Input

In this approach, here it is going to be implement a decision support system for social media comments analysis to identify hate related comments which are made in Sinhala Language. For this, From the “Kaggle” data science company’s database, I collected social media comments made in Sinhala language, in the year 2020 which was labeled as Negative or Positive [18]. Then the comments saved in excel file with utf-8 formatting which was in .csv format.

4.4 Output

The major output of this study is to provide a classification model to identify whether the user's comment which are made in Sinhala language, are negative or positive. It included main components as below,

- Classification model based on dictionary which contain Negative words.
- Classification model based on machine learning techniques.

4.5 Process

Commonly for the both experiments, selected data corpus has been splitted into two groups for training purposes and testing purposes. In the experiment A, a list of negative words was extracted to construct a dictionary through the training data set. Then the dictionary based classification model was created by using Rapid Miner software. While creating the model, pre-processing of data also done with same software. For the Experiment B, machine learning techniques based model was created by using Rapid Miner software as same as experiment A. Within that text pre-processing, feature extraction and vectorization also done with the same software. Then, several models were created, by using several algorithms and modifying the feature extraction methods. After that the accuracy of all above models were evaluated and finally, identify the best fit model for hate speech identification.

4.6 Summary

This chapter describes lexicon based method and machine learning based method for conducting this research. Here we identify the possible research process to solve the identified research problem. The following chapter present the proposed solution's design

Chapter 5

Analysis and Design

5.1 Introduction

In forth chapter, we described the approach to develop the hate speech classification model using negative word dictionary and machine learning algorithms. In here, we present the analysis and design of the proposed solution design.

5.2 High level Architecture of System

In this research we have developed two main approaches. Those are dictionary based method and machine learning techniques based method. High level architecture of the proposed approaches is illustrated by Figure 5.1.

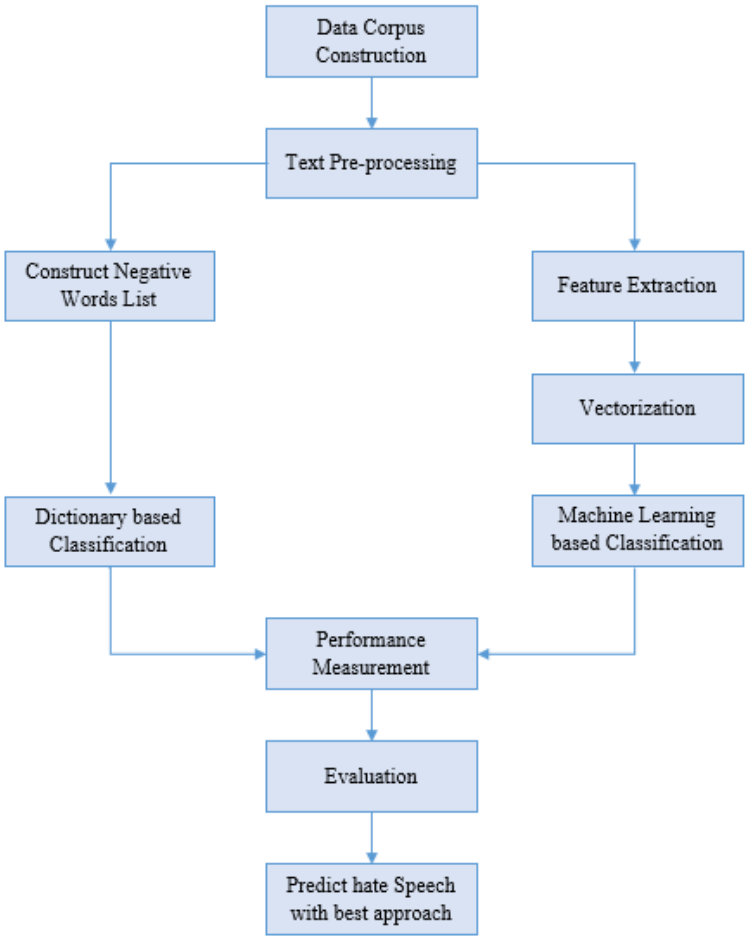

Figure 5.1: High Level Architecture of the Design

In the first phase, data corpus was constructed with retrieved data set from “Kaggle”. For that, it was divided into two thirds for one part and one third for the remaining part. The first two thirds were used as a training dataset and the other part as a test part. After that the training data set were tokenized under pre-processed section and remove all the non-Sinhala words. The process was then carried out as two experiments, A and B, and the B experiment was further divided into two parts.

Under Experimental A, the list of negative words was extracted from the training data set for dictionary-based hate speech identification approach. A dictionary-based classification model was then developed and the classification was performed using that model and the test data set.

Under Experiment B1, text data was pre-processed without performing stop words removal and stemming and under Experiment B2, those are performed under text pre-processing part. Then under both experiments, features of words were extracted before vectorization. Modeling was then performed using two learning algorithms. After that, with the test data corpus, classification was performed with two approaches.

For all of the above experiments (A, B1 and B2), the performance was measured and finally those accuracies were evaluated with each other and the best fit model was found. Then found best fit model can be used to identify hateful content in Sinhala language on social media.

5.3 Summary

This chapter discussed the high level architecture of social media comments classification. And also steps followed to achieve identification of hateful content in Sinhala Language on social media was described briefly. The next chapter explains the overall implementation.

Chapter 6

Implementation

6.1 Introduction

Earlier chapter, we considered the high-level architecture of the proposed solution. In this chapter, the overall implementation, which was briefly stated in the above chapter, will be described in comprehensive.

6.2 Data Corpus Construction

Datum was collected by using “Kaggle” data science company’s database. It includes more than 3,000 Sinhala Unicode based labeled data corpus. But in this research only 3,000 labeled comments were used with the purpose of selecting a balanced data corpus and considering the limited resources available for research. Within that balanced data corpus contained 1,500 negative comment and 1,500 positive comments. Then the comments were saved in excel files for training and testing under utf-8 formatting. 2,000 comments in other words two thirds of total comments were used for training scenario and 1,000 comments or one third of total comments were used for testing scenario.

Table 6.1: Annotated Data Corpus

	Data Corpus	
Comment Type	**Testing Corpus**	**Training Corpus**
Hate Comments	500	1,000
Non-hate Comments	500	1,000
Total	1,000	2,000

6.3 Text Pre-processing

Before use the above data corpus for analysis it was pre-processed with following steps.

- **Removal Duplicates**

  Removing duplicates from training data set in Experiment B, was done as a first step in text pre-processing. This step was done to remove any similar phrases from selected corpus. These steps were used to further increase the accuracy as well as the efficiency of this research.

- **Tokenization**

  With this step, words will be extracted from the paragraphs. For that firstly paragraphs will tokenize into sentences, and then the it can be tokenized into words. Rest of the processing is usually done after the text is tokenized appropriately. Tokenization is also known as text segmentation. In the tokenization process, some characters, such as punctuation, are discarded. Then the tokens become the input for next step of pre-processing. In this research, before pre-processed furtherly in both approaches this step was done.

- **Filter Tokens (by content)**

  Since the aim of this research is identify hateful contents that are expressed using Sinhala Unicode, non-Sinhala characters were removed from the data set by using "Filter Tokens by Content". In this step, it used below regular expression to strainer non Sinhala letters.

  Regular expression: [\u0D80-\u0DFF]+

- **Removal of Stop words**

  To preserve the structure of the sentences "Stop Words" are used. And they are not much contributing to its meaning. The words like "සහ", "ද" and "ක්" are belongs to this group. Those words are eliminated from the example set with the purpose of reducing the dimensionality of feature vector. In this research it used publically available stop word list [59] for Experiment B.

- **Stemming**

Stemming can be identified as one step in normalization. In this step it is mainly focused to remove affixes (suffixed, prefixes, infixes, circumfixes) from a word to get a word stem. In the English language, we have suffixes like "-ed" and "-ing" which may be useful to cut off in order to map the words "cook," "cooking," and "cooked" all to the same stem of "cook." In Sinhala Language, "අංශටත්", "අංශය", "අංශයක්", "අංශයටත්", අංශයත්" can be map to word "අංශ". As like this for both approaches stemming was used. It is really important to determine stems of words because that process reduce the largeness of the feature vector. When consider about text classification with some large datasets it is very important to reduce the size of the feature vector. In here for Experiment B, it used publically available stemmer dictionary [60].

6.4 Construct Negative Word List

Experiment A, of this research, was carried out using a dictionary containing the negative words. Therefore, a list of negative words was constructed using training data set.

For that first of all, a list of words was generated with information about the number of times words were used in negative documents and positive documents by using Rapid Miner process as shown in below Figure 6.1.

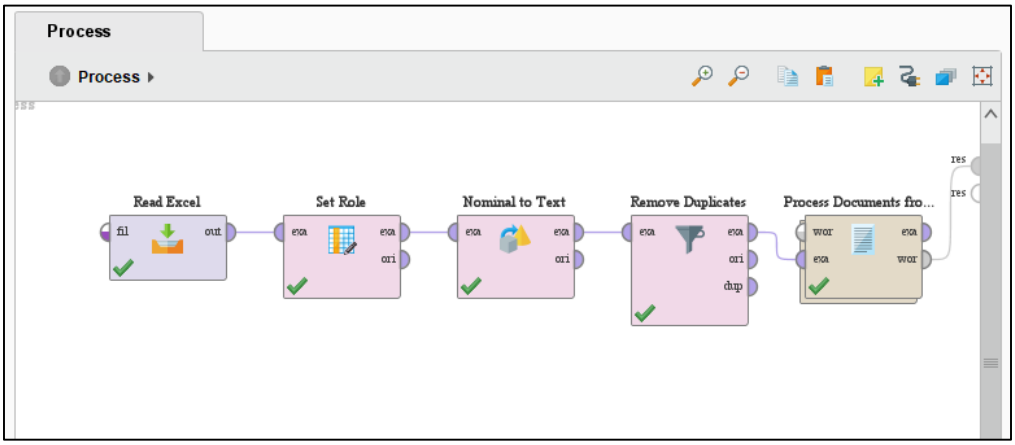

Figure 6.1: Procedure to obtain a detailed list of words from training data set

A detailed list of words generated by the above process is shown in Figure 6.2.

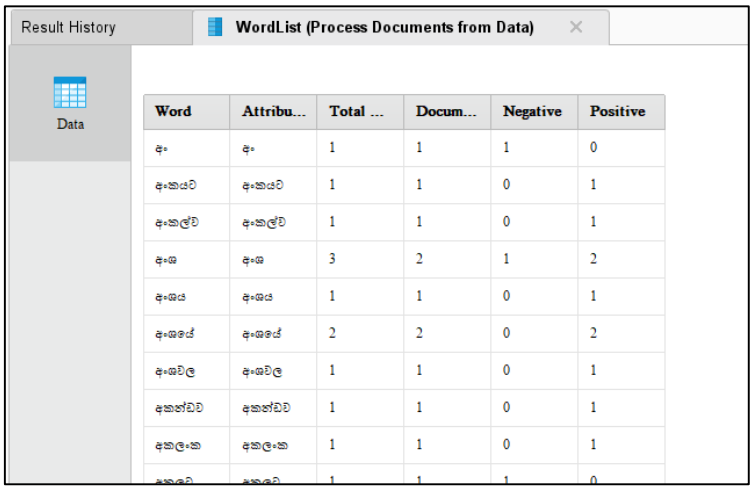

Figure 6.2: Word List of Training Data Set

There was a list of more than 7,500 words. And then a list of words that were more likely to be given a negative meaning by studying those words was separated from it. Of these, nearly 900 words were classified as negative words as shown in below Figure 6.3.

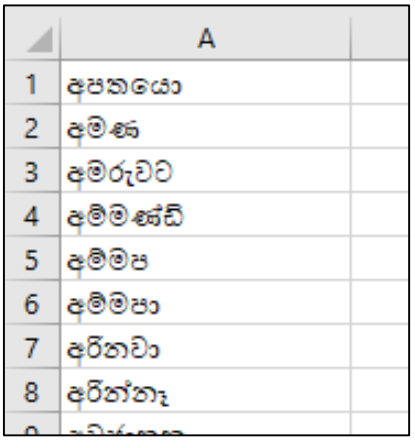

Figure 6.3: Negative word List

6.5 Dictionary based Classification

As stated previously, Dictionary based classification (Experiment A) was implemented with the Rapid Miner Studio software. As a first step of this experiment, we added a score (-1) for each negative word which was extracted from training data set. Then the Dictionary Based Sentiment (Documents) tool which is available in Rapid Miner was used to build a dictionary based classification model.

Then the testing data set was pre-processed with the steps tokenization and Filter Tokens (by Content) to feed for the previously built model. After that the built model was applied to the test datum. From that it was able to find out the total score for each comment relative to our negative words dictionary. According to that total score, comments were predicted as “Negative” or “Positive”. If a comment had a negative score, we labeled it as a negative comment, otherwise it was labeled as positive comment.

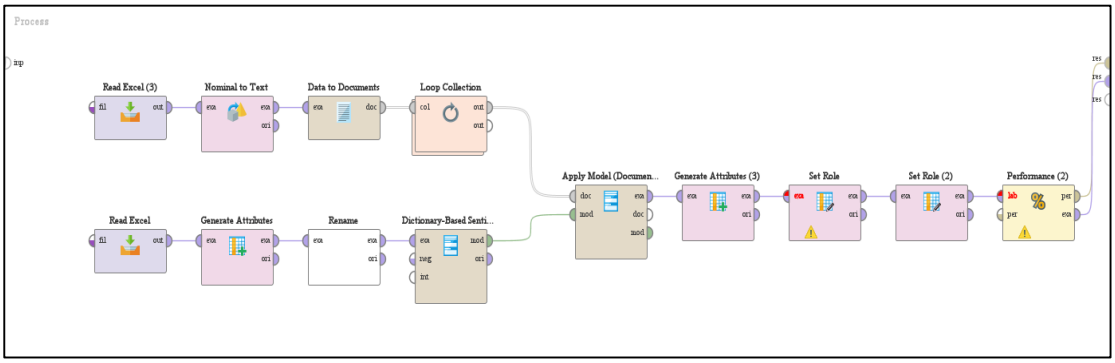

Figure 6.4: Dictionary based Classification Model

6.6 Feature Extraction

An important step in machine learning based text classification, after pre-processing, is the selection of features to create vector space, which enhance the exactitude of text classifier, efficiency, and scalability. The basic idea of feature selection is to select a set of features from the initial document. Selection of feature is done by placing the words with the highest score according to a predetermined measure of the significance of the word. The main problem with text classification is the large dimensionality of the feature vector. According to the literature relevant with Sinhala language based text classification, word n-gram feature and character n-gram feature were selected as feature extraction techniques for this study.

- **Word-n-gram features**

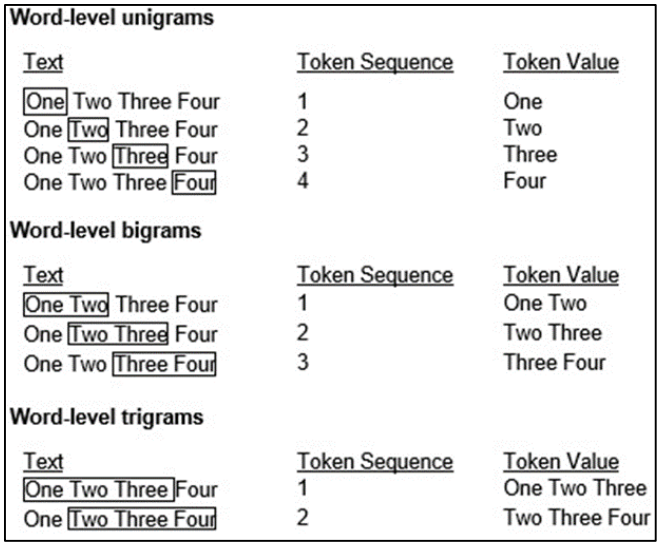

Figure 6.5: Example for Word N-gram

Word n-gram is a term form that grasp the structure of a clause. These n-gram features can be either a "unigram", "bigram", "trigram", etc. or combination of these. According to the literature, many researches in different languages shows that extracting word n-gram features to identify hateful speeches makes better results. Therefore, in this research several groups of word n-gram feature were used. According to the Sinhala Language related hate speech detection studies, [17] [19] selected word n-gram feature groups are as below, Table 6.2.

Table 6.2: Word N-gram Feature Groups

Group	**Feature Group**
WG01	Unigram (UG)
WG02	Bigram (BG)
WG03	Trigram (TG)
WG04	UG + BG
WG05	UG + BG + TG

- **Character-n-gram features**

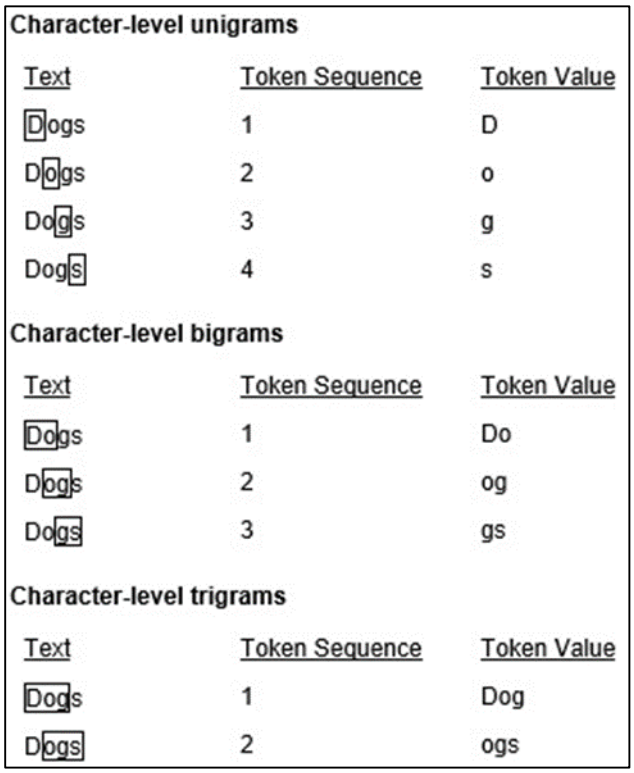

Figure 6.6: Example for Character N-gram

In some languages, such as English, written contents consists of letters, numbers, punctuation, and space. It can be seen that many words are misspelled when commenting on social media. In this type of cases, character n-grams are especially potent at determining patterns in such things than previously mentioned word n-gram features. These n-gram features can be either a “unigram”, “bigram”, “trigram”, etc. or “combination of these features”. According to the Sinhala Language related hate speech detection studies, [17] [19] selected word n-gram feature groups are as below, Table 6.3.

Table 6.3: Character N-gram Feature Groups

Group	**Feature Group**
CG01	Bigram (BG)
CG02	Trigram (TG)
CG03	4 - gram
CG04	BG + TG
CG05	BG + TG + 4 - gram

6.7 Feature Vectorization

Vectorization is the procedure of converting written language text into vector or in other words numbers for further processing steps of text classification. Since software are not able to work with written languages, this step was done and statistical methods that are used to classification need vectorized input data. Therefore, feature vectorization is a major step of machine learning based text classification. As clearly identified according to the relevant literature, in this used "Term Frequency – Inverse Document Frequency" method for feature vectorization.

- **Term Frequency – Inverse Document Frequency (TF-IDF)**

  TF-IDF is used traditional unsupervised term weighting method. What usually happens here is a normalization of the time frequency produced earlier. Inevitably, this weighing system consists of two factors. "Term Frequency (TF)" is the first factor, and it provides the "frequency of term j in the i^{th} document". The next factor is the "inverse document frequency (IDF)". This implies that a small weight should be placed on a term that took place in several documents.

  For a term i in document j;

$$w_{ij} = tf_{ij} \times \frac{\log N}{df_i}$$

  Where tf_{ij} = Number of occurrences of i in j

  dfi = Number of documents contain i

  N = Total number of documents

  Before apply a classification methods vectorization is an important step that we have to conduct in text classification process. From the relevant literature, it can be identified that in most of the cases researchers used TF-IDF method to weighted their term features. Therefore, in general, we can conclude that TF-IDF performs vital role within the process of text classification.

6.8 Machine Learning based Classification

Text classification is the automatic classification of documents into pre-defined categories according to their content. Documents can be classified according to three methods. They are supervised, unsupervised and semi-supervised classification. In this research it focused only on supervised machine learning algorithms. The function of automated text classification has been extensively studied over the past few years, and speedy grown is being made in this area, including machine learning based methods such as the NB, SVM, and etc. From the relevant literature specially based on Sinhala Language related studies, expanded version of Naïve Bayes, Multinomial Naive Bayes (MNB), and SVM were selected to conduct this research study.

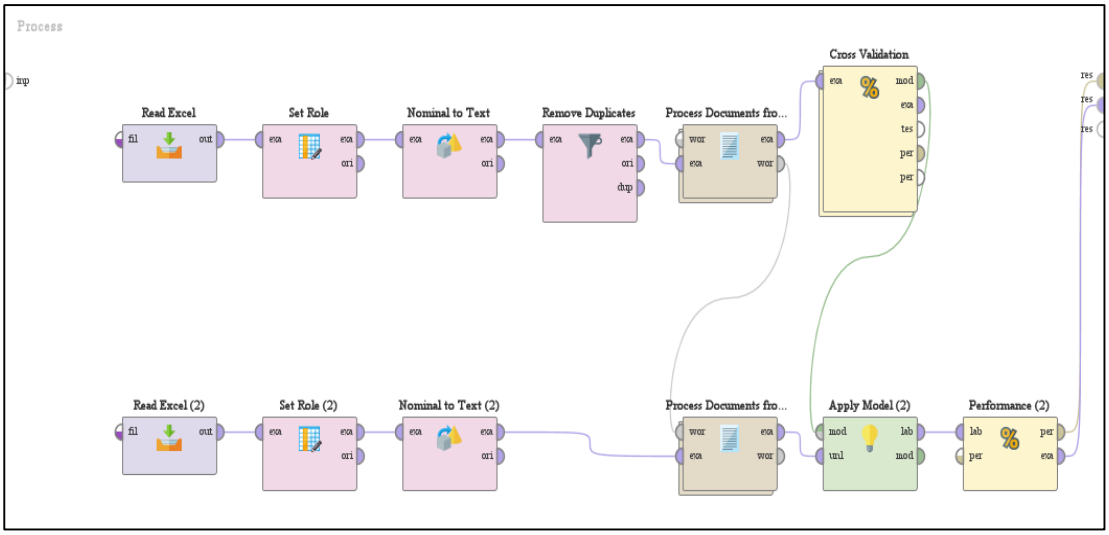

Figure 6.7: Machine Learning based Classification Model

6.9 Performance Measurements

This is the final stage of the text classification in both experiments, and the evaluation of the text classification is usually done experimentally, rather than analytically. Experimental evaluation of the classification seeks to assess the effectiveness of the classification in general, rather than focusing on efficiency issues. A crucial issue in text categorization is how to measure the functionality of the classification. In general, many measurements are used such as accuracy, precision, recall, and F1-score and etc.

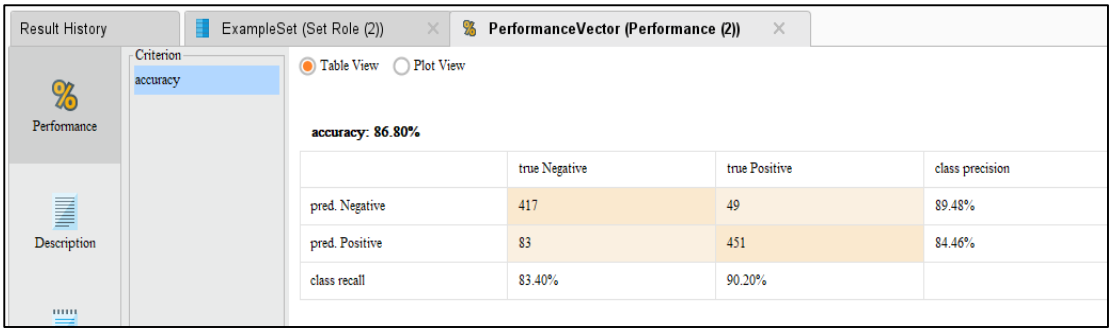

Figure 6.8: Performance Measurements

In this research, accuracy, precision, recall, and F1 score were calculated in order to evaluate all experiments. The formulas used to calculate the measurements are as follows,

- **Accuracy** – "It is the percentage of correctly classified hate tweets over the total number of hate comments. It is shown in below equation" [63];

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FN} + \textbf{FP}}$$

where TP = True Positives
FP = False Positives
TN = True Negative and
FN = False Negative.

- **Precision – "**It quantifies the fraction of detected hate tweets that are matches to labeled hate tweets database. It can be mathematically denoted as shown in below equation" [63];

$$\textbf{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

- **Recall – "**It quantifies the percentage of labeled hate tweets that are correctly detected as shown in below equation" [63];

$$\textbf{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

- **F1-score** – "It is the harmonic mean of precision and recall as shown in below equation. The F1-score is a measure of a test accuracy. F1-score reaches its best value at 1 and worst value at 0" [63].

$$\textbf{F1 score} = \frac{\mathbf{2} \times \boldsymbol{P} \times \boldsymbol{R}}{\boldsymbol{P} + \boldsymbol{R}}$$

where P = Precision and

R = Recall

6.10 Summary

This chapter described an overall implementation of the proposed solution. It also describes how software and text classification techniques were used to develop the model. In the following section, all the models executed in the solution will evaluated.

Chapter 7

Evaluation

7.1 Introduction

Sixth chapter described about the overall implementation of the proposed solution. In here rationalize and evaluates the proposed solution.

7.2 Evaluation of Classification Techniques

As described earlier, collected data set was trained under following experiments with the help of Rapid Miner tools,

Table 7.1: Performed Experiments

<table>
<tr><td>Experiment A</td><td colspan="2">Dictionary Based Hate Speech Detection</td></tr>
<tr><td rowspan="4">Experiment B</td><td rowspan="2">Experiment B1 - Machine Learning based Hate Speech Detection (without Removing Stop Words and Stemming)</td><td>Experiment B1.1 - with MNB Classification Technique</td></tr>
<tr><td>Experiment B1.2 - with SVM Classification Technique</td></tr>
<tr><td rowspan="2">Experiment B2 - Machine Learning based Approach (with Removing Stop Words and Stemming)</td><td>Experiment B2.1 - with MNB Classification Technique</td></tr>
<tr><td>Experiment B2.2 - with SVM Classification Technique</td></tr>
</table>

As mentioned under Performance Measurements section in chapter 6, for each experiment accuracy, precision, recall, and F1-score were calculated. Under machine learning based experiments, we used cross-validation techniques with ten folds. Following tables shows the results we got in deferent experiments.

Table 7.2: Results of Experiment A - Dictionary Based Hate Speech Detection

Class	Precision	Recall	F1-Score	Accuracy
Hateful (H)	89.47	81.60	85.35	**86.00**
Non-Hateful (N)	83.09	90.40	86.59	

Table 7.3: Results of Experiment B1.1 - with MNB Classification Technique

MNB Classifier	Class	Vectorization Method - TF-IDF									
		Feature Extraction									
		Character n-gram feature group					Word n-gram feature group				
		CG01	**CG02**	**CG03**	**CG04**	**CG05**	**WG01**	**WG02**	**WG03**	**WG04**	**WG05**
Precision	H	83.12	84.49	84.74	84.62	84.75	83.02	83.52	83.46	83.21	83.24
	N	89.67	91.81	91.45	91.63	92.65	88.15	88.41	87.13	88.36	88.55
Recall	H	90.60	92.60	92.20	92.40	93.40	89.00	89.20	87.80	89.20	89.40
	N	81.60	83.00	83.40	83.20	83.20	81.80	82.40	82.60	82.00	82.00
F1-Score	H	86.70	88.36	88.31	88.34	88.87	85.91	86.27	85.58	86.10	86.21
	N	85.44	87.18	87.24	87.21	87.67	84.86	85.30	84.80	85.06	85.15
Accuracy		86.10	87.80	87.80	87.80	**88.30**	85.40	85.80	85.20	85.60	85.70

Table 7.4: Results of Experiment B1.2 - with SVM Classification Technique

SVM Classifier	Class	Vectorization Method - TF-IDF									
		Feature Extraction									
		Character n-gram feature group					Word n-gram feature group				
		CG01	**CG02**	**CG03**	**CG04**	**CG05**	**WG01**	**WG02**	**WG03**	**WG04**	**WG05**
Precision	H	96.21	96.78	93.27	95.71	94.91	92.88	94.31	94.27	95.32	98.18
	N	74.12	77.83	80.82	77.19	79.08	74.65	77.28	76.90	73.56	59.52
Recall	H	66.00	72.20	77.60	71.40	74.06	67.80	71.87	71.27	65.20	32.40
	N	97.40	97.60	94.40	96.80	96.00	94.80	95.67	95.67	96.80	99.40
F1-Score	H	78.29	82.70	84.72	81.79	83.20	78.38	81.57	81.17	77.43	48.72
	N	84.18	86.60	87.08	85.89	86.72	83.53	85.50	85.26	83.59	74.46
Accuracy		81.70	84.90	**86.00**	84.10	85.30	81.30	83.77	83.47	81.00	65.90

Table 7.5: Results of Experiment B2.1 - with MNB Classification Technique

MNB Classifier	Class	Vectorization Method - TF-IDF									
		Feature Extraction									
		Character n-gram feature group					Word n-gram feature group				
		CG01	**CG02**	**CG03**	**CG04**	**CG05**	**WG01**	**WG02**	**WG03**	**WG04**	**WG05**
Precision	H	83.27	85.16	84.13	84.44	84.25	81.95	82.05	82.12	81.60	81.72
	N	89.69	91.11	90.39	90.43	91.19	87.96	88.55	87.26	88.47	88.30
Recall	H	90.60	91.80	91.20	91.20	92.00	89.00	89.60	88.20	89.60	89.40
	N	81.80	84.00	82.80	83.20	82.80	80.40	80.40	80.80	79.80	80.00
F1-Score	H	86.78	88.36	87.52	87.69	87.95	85.33	85.66	85.05	85.41	85.39
	N	85.56	87.41	86.43	86.66	86.79	84.01	84.28	83.91	83.91	83.95
Accuracy		86.20	**87.90**	87.00	87.20	87.40	84.70	85.00	84.50	84.70	84.70

Table 7.6: Results of Experiment B2.2 - with SVM Classification Technique

SVM Classifier	Class	Vectorization Method - TF-IDF									
		Feature Extraction									
		Character n-gram feature group					Word n-gram feature group				
		CG01	**CG02**	**CG03**	**CG04**	**CG05**	**WG01**	**WG02**	**WG03**	**WG04**	**WG05**
Precision	H	95.68	96.82	92.38	96.25	94.22	93.13	94.88	97.22	96.75	97.75
	N	74.27	78.33	80.69	77.51	79.24	77.92	76.47	69.10	67.91	60.34
Recall	H	66.40	73.00	77.60	71.80	75.00	73.20	70.40	56.00	53.60	34.80
	N	97.00	97.60	93.60	97.20	95.40	94.60	96.20	98.40	98.20	99.20
F1-Score	H	78.40	83.24	84.35	82.25	83.52	81.97	80.83	71.07	68.98	51.33
	N	84.13	86.91	86.67	86.25	86.57	85.45	85.21	81.19	80.29	75.04
Accuracy		81.70	85.30	**85.60**	84.50	85.20	83.90	83.30	77.20	75.90	67.00

According to the above results, Experiment B gave higher accuracy value than Experiment A. Furtherly, the "Experiment B1.1" showed greater accuracy from all experiments. So we can say that Multinomial Naïve Bayes produced the best results in predicting hateful comments for this data set. If this best result is further described,

Table 7.7: Details of methods used and the results of best fit model

Details of methods used		
Pre-processed Techniques		- Removal Duplicates - Tokenization - Filter Tokens (by content)
Feature Extracted Method		Character-n-gram feature
Character N-gram Feature Groups		CG05 – Bigram(BG) + Trigram(TG) + 4-gram
Feature Vectorized Method		Term Frequency – Inverse Document Frequency (TF-IDF)
Classification Method		Machine Learning based Classification
Used machine learning algorithm		Multinomial Naïve Bayes
Details of the results		
Accuracy		**88.30%**
Precision	Hateful	84.75%
	Non - Hateful	92.65%
Recall	Hateful	93.40%
	Non - Hateful	83.20%
F1 - Score	Hateful	84.75%
	Non - Hateful	92.65%

7.3 Summary

This chapter evaluated results of all experiments which was described in the implementation chapter. Next chapter will state the conclusion and also discuss about the limitations and future expansions.

Chapter 8

Conclusion and Future Work

8.1 Introduction

The earlier chapter described about the result of all experiments and this chapter will state the achievements of the objectives, limitations and further developments.

8.2 Conclusion

In this research, the problem of identifying hateful comments in Sinhala language on social media was addressed with the data set which contain Facebook comments in Sinhala language. We have been able to identify hateful ideas by analyzing the words contained in the ideas.

The main goal of this research to identify accurate and efficient model to predict harmful contents which are expressed in the Sinhala language on social media. There were two main experiments in this approach. One is creating a dictionary of bad words and use it to identify hateful comments. And the other one is identifying hateful comments using machine learning algorithms. In the second experiment, it was divided into two parts based on the sequence of action followed under text pre-processing step and also, different machine learning algorithms were used.

By evaluating all models with its accuracy value, Multinomial Naïve Bayes algorithm without removing stop words and stemming under text pre-processing were selected with highest accuracy. Within that model “Removal Duplicates”, “Tokenization”, and “Filter Tokens (by content)” techniques were used under text pre-processing. Since comments published in Sinhala language contain more spelling mistakes, “Character n-gram” feature carry out well in hateful speech identification. Combination of “Unigram”, “Bigram”, and “Trigram” was outperformed as character n-gram under feature extraction techniques. Other than that “Term Frequency – Inverse Document Frequency” technique was used to vectorized extracted features.

8.3 Limitations

The subtleties of the Sinhala language, the different interpretations of hateful speech, and the difficulty of obtaining data for training and testing of these systems, as well as the deficiency of the annotators to label the data, can be identified as main limitations in this research. Apart from that common factors such as spelling mistakes and replacement of similar characters can be identified as a limitation to achieve very high accuracy.

8.4 Future Developments

As a future work, the ability to increase the level of accuracy by increasing the size of the database can be tested. Since the level of accuracy of this type of research is highly dependent on the accuracy of the labels, we can use group of annotators to increase label accuracy. In addition, it is an untouched zone to apply unsupervised machine learning algorithms to identify Sinhala hateful contents.

In here we consider only the comments published using Sinhala Unicode but we can expand this research with comments which are made in Singlish (Sinhala words written in English language) that contain hate. This will be really challenging area since it will contain both Singlish and English words.

8.5 Summary

This final chapter discussed the conclusion of this research, limitations and future works.

References

[1] Kietzmann, Jan H.; Kristopher Hermkens (2011). "Social media? Get serious! Understanding the functional building blocks of social media". Business Horizons (Submitted manuscript). 54 (3): 241–251. doi:10.1016/j.bushor.2011.01.005.

[2] Obar, Jonathan A.; Wildman, Steve (2015). "Social media definition and the governance challenge: An introduction to the special issue". Telecommunications Policy. 39 (9): 745–750. doi:10.1016/j.telpol.2015.07.014. SSRN 2647377.

[3] “Hate Speech | Define Hate Speech at dictionary.cambridge.org.” [Online]. Available: https://dictionary.cambridge.org/us/dictionary/english/hate-speech

[4] " Verbal abuse | Define verbal abuse at American Heritage Dictionary of the English Language (Sixth ed.).” [Online]. Available: https://www.ahdictionary.com/word/search.html?q=abuse

[5] The Verbally Abusive Relationship, Patricia Evans. Adams Media Corp 1992, 1996, 2010

[6] Erving Goffman, Relations in Public (Penguin 1972) p. 214

[7] Facebook community standards [Online]. Available: https://www.facebook.com/communitystandards/objectionable_content

[8] Facebook: Hard Questions - Who Should Decide What Is Hate Speech in an Online Global Community?[Online]: https://about.fb.com/news/2017/06/hard-questions-hate-speech/

[9] Liking violence: A study of hate speech on Facebook in Sri Lanka‖ [Online] Retrieved from https://www.cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf

[10] Nisansa de Silva (2019), “Survey on Publicly Available Sinhala Natural Language Processing Tools and Research” [Online]. Available: https://arxiv.org/pdf/1906.02358.pdf

[11] Jovic A., Brkić K. and Bogunovic N. (2014). “An overview of free software tools for general data mining”. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)

[12] Hasim N. and Haris A.A. (2015). “A study of open-source data mining tools for forecasting”. ACM IMCOM (ICUIMC) 2015 The 9th International Conference on Ubiquitous Information Management and Communication

[13] Digital 2020 report for Sri Lanka [Online]. Available: https://datareportal.com/reports/digital-2020-sri-lanka

[14] Samarasinghe S. W. A. M. D., Meegama R. G. N. and Punchimudiyanse M. (2020). “Machine Learning Approach for the Detection of Hate Speech in Sinhala Unicode Text”. 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)

[15] Dulan S. Dias; Madhushi D. Welikala; Naomal G.J. Dias (2018). “Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning”. 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)

[16] Hissah Saif and Hmood Al-Dossari (2018). “Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques”. January 2018, International Journal of Advanced Computer Science and Applications 9(10)

[17] Sandaruwan H.M.S.T, Lorensuhewa S.A.S, and Kalyani M.A.L(2019). “Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning”. 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)

[18] Jayasuriya S. “Sinhala Unicode Hate Speech - Based on Sinhala Unicode based comments on Facebook V1.” April 12, 2020. Distributed by Kaggle Data Science Company. https://www.kaggle.com/sahanjayasuriya/sinhala-unicode-hate-speech

[19] Sandaruwan H.M.S.T., Lorensuhewa S.A.S. and Kalyani M.A.L., “Identification of Abusive Sinhala Comments in Social Media using Text Mining and Machine Learning Techniques”, International Journal on Advances in ICT for Emerging Regions 2020 13 (1)

[20] Amali H. M. A. I. and Jayalal S., “Classification of Cyberbullying Sinhala Language Comments on Social Media”, Moratuwa Engineering Research Conference (MERCon) 2020

[21] Sohn H. and Lee H., “Hate Speech Detection using Multi-channel BERT for Different Languages and Translations”, 2019 International Conference on Data Mining Workshops (ICDMW)

[22] Google Translate | Languages [Online]. Available: https://translate.google.com/intl/en/about/languages/

[23] Ulatus, “Translations Made Simple: The Usefulness of Translation Apps” [Online]. Available: https://www.ulatus.com/translation-blog/most-globally-used-translated-apps/

[24] “RapidMiner | Best Data Science & Machine Learning Platform” [Online]. Available: https://rapidminer.com/

[25] Ahammed, S., et al. (2019). Implementation of Machine Learning to Detect Hate Speech in Bangla Language. 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART).

[26] Akbar, R. R. e., et al. (2019). The Implementation of Naïve Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motive. 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC).

[27] Akhter, M. P., et al. (2020). "Automatic Detection of Offensive Language for Urdu and Roman Urdu." IEEE Access **8**: 91213-91226.

[28] Albadi, N., et al. (2018). Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[29] Alfina, I., et al. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

[30] Alrehili, A. (2019). Automatic Hate Speech Detection on Social Media: A Brief Survey. 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA).

[31] Chavan, V. S. and S. S. Shylaja (2015). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[32] Elisabeth, D., et al. (2020). Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study. 2020 8th International Conference on Information and Communication Technology (ICoICT).

[33] Fatahillah, N. R., et al. (2017). Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. 2017 International Conference on Sustainable Information Engineering and Technology (SIET).

[34] Fernquist, J., et al. (2019). A Study on the Feasibility to Detect Hate Speech in Swedish. 2019 IEEE International Conference on Big Data (Big Data).

[35] Ginting, P. S. B., et al. (2019). Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method. 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS).

[36] Hana, K. M., et al. (2020). Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines. 2020 International Conference on Data Science and Its Applications (ICoDSA).

[37] Hendrawan, R., et al. (2020). Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media. 2020 International Conference on Data Science and Its Applications (ICoDSA).

[38] Koushik, G., et al. (2019). Automated Hate Speech Detection on Twitter. 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA).

[39] Luu, S. T., et al. (2020). Comparison Between Traditional Machine Learning Models And Neural Network Models For Vietnamese Hate Speech Detection. 2020 RIVF International Conference on Computing and Communication Technologies (RIVF).

[40] Lynn, T., et al. (2019). A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary. 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA).

[41] Martins, R., et al. (2018). Hate Speech Classification in Social Media Using Emotional Analysis. 2018 7th Brazilian Conference on Intelligent Systems (BRACIS).

[42] Nugroho, K., et al. (2019). Improving Random Forest Method to Detect Hatespeech and Offensive Word. 2019 International Conference on Information and Communications Technology (ICOIACT).

[43] Ombui, E., et al. (2019). Hate Speech Detection in Code-switched Text Messages. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).

[44] Oriola, O. and E. Kotzé (2019). Automatic Detection of Toxic South African Tweets Using Support Vector Machines with N-Gram Features. 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI).

[45] Oriola, O. and E. Kotzé (2020). "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets." IEEE Access **8**: 21496-21509.

[46] Prabowo, F. A., et al. (2019). Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter. 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE).

[47] Rai, N., et al. (2020). Improving the hate speech analysis through dimensionality reduction approach. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS).

[48] Rodríguez-Sánchez, F., et al. (2020). "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data." IEEE Access **8**: 219563-219576.

[49] Rohmawati, U. A. N., et al. (2018). SEMAR: An Interface for Indonesian Hate Speech Detection Using Machine Learning. 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI).

[50] Ruwandika, N. D. T. and A. R. Weerasinghe (2018). Identification of Hate Speech in Social Media. 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer).

[51] Şahi, H., et al. (2018). Automated Detection of Hate Speech towards Woman on Twitter. 2018 3rd International Conference on Computer Science and Engineering (UBMK).

[52] Sajjad, M., et al. (2019). Hate Speech Detection using Fusion Approach. 2019 International Conference on Applied and Engineering Mathematics (ICAEM).

[53] S.-B. Kim, K.-S. Han, H.-C. Rim, S.H. Myaeng, (2006) “Some effective techniques for Naive Bayes text classification.” IEEE Transactions on Knowledge and Data Engineering (Volume: 18, Issue: 11, Nov. 2006)

[54] Sazany, E. and I. Budi (2019). Hate Speech Identification in Text Written in Indonesian with Recurrent Neural Network. 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS).

[55] Senarath, Y. and H. Purohit (2020). Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media. 2020 IEEE 14th International Conference on Semantic Computing (ICSC).

[56] Souza, G. A. D. and M. D. Costa-Abreu (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. 2020 International Joint Conference on Neural Networks (IJCNN).

[57] Yadav, S. H. and P. M. Manwatkar (2015). An approach for offensive text detection and prevention in Social Networks. 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

[58] RapidMiner Marketplace (Online). Available: https://marketplace.rapidminer.com/UpdateServer/faces/index.xhtml

[59] Sinhala Stopword List (Online). Available: https://uom.lk/nlp/tools

[60] Sinhala Stemmer Dictionary (Online). Available: https://github.com/rksk/sinhala-news-analysis

[61] News Related to "Hate Speech in Sinhala Language on Social Media [Online]. Available: https://sinhala.srilankamirror.com/news/16219-senior-minister-of-state-for-law-and-health-edwin-tong-at-an-international-hearing-on-fake-news-and-disinformation-in-london

[62] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions, Computer Science Review, 2020

[63] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga (2020). Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks, International Journal of Intelligent Computing and Cybernetics, 2020