# Detecting Sinhala Language Based Racial and Religious Offensive Statements in Social Media

A dissertation submitted for the Degree of Master of Computer Science

K. A. T. L. Wimalasena
University of Colombo School of Computing

# Declaration of Authorship

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: K.A.T.L. Wimalasena
Registration Number: 2017/MCS/097
Index Number: 17440976

2021-11-27
Date

Signature of the student

This is to certify that this thesis is based on the work of Mr. K. A. T. L. Wimalasena under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,
Supervisor Name: Mr. W.V.Welgama

2021 / 11 / 29
Date

Signature of the supervisor

# Acknowledgements

First and foremost, I'd like to express my heartfelt gratitude to my research supervisor, Mr.W.V Welgama, Senior Lecturer, University of Colombo School of Computing, for giving me the opportunity to do research and providing invaluable guidance throughout this research. His wide knowledge and logical way of thinking have provided me with valuable help, guidance and advice on how to conduct the research successfully.

I whole-heartedly expressed my gratitude to all the scholars who guided me with their findings in literature review, and direct me towards correct path of research methodology. It's impossible to express my gratitude enough to my family, particularly my parents, who provided me with the support I needed throughout this process....

# Abstract

The offensive statements and few people who promoted the violence using Facebook posts are the main reasons for few devastating incidents which took place in Sri Lanka. In March 2018, the Sri Lankan government was forced to impose a one-week social media ban in order to prevent the dissemination of false information and racial ideas that could complicate the situation. However, once the ban lifted, there were no mechanism to moderate the comments and posts in Facebook. Relevant authorities have failed to stop the spread of hate via social media platforms since they don't have capable Sinhala language interpreters to detect racial and religious offensive statements. In this study, a machine learning based model has presented to detect Sinhala language based racial and religious offensive statements. The pre-processed TF-IDF weighted character n-grams was used as features and three prominent machine learning based classifiers as Logistic Regression, Naive Bayes and Support Vector Machines were trained and tested. Naive Bayes classifier recorded F1 Score of 0.741 while SVM records 0.801. The highest accuracy and F1 Score of 0.824 and 0.851 respectively were obtained with Logistic Regression. As per the results, TF-IDF weighted character n-grams features with Logistic Regression is a comprehensive model for detecting sinhala labguage based racial and offensive statements in social media.

Key Words : Natural Language Processing, Machine Learning

# Contents

# Abbreviations

| | |
|---|---|
| NLP | Natural Language Processing |
| LR | Logistic Regression |
| TF-IDF | Term Frequency Inverse Document Frequency |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Problem domain

The Social media is a platform that allows people from all over the world to connect instantaneously and transform the way they communicate and share information with each another. Social media has positively influenced the society in many ways. For an example, there are so many Facebook groups which facilitating online learning on different subjects. However, It has also expanded the chance of harm. The ability to communicate with a large audience at once has completely changed the way people engage with politics, public affairs, and one another. At times, different in opinions among people lead to verbal assaults. With the emergence of new platforms for communicate with one another, it has become a way to spread racial thoughts, hatred and discrimination on several grounds.

## 1.2 The problem

Hate speech which consist offensive statements in internet has been connected to a global upsurge in violence toward minorities, including mass shootings and ethnic cleansing (Council on Foreign Relations,2020). A statement, one has put in social media can be gone viral in worldwide. This has become a severe problem when certain statements and behaviors cause an outbreak of violence among communities. The offensive statements and few people who are promoting the violence using Facebook posts are the main reasons for few devastating incidents which happened in Sri Lanka. That situation occurred solely as a result of a miscommunication between parties on social media, that caused the issue to spread like a virus to neighbouring cities. In March 2018, Sri Lankan government was forced to ban Facebook for almost one week to stop the spreading of false information and racist sentiments that could worsen the situation. Ultimately, after an inquiry concluded that hate speech and rumors circulated on Facebook may have led to violence against Muslims, Facebook has apologised for its involvement in the deadly communal disharmony that shook Sri Lanka two years ago (ALJAZEERA.COM,2020).

According to the research on "Anti-Muslim Sentiments and Violence", Sinhalese nationalists have exploited print and electronic media, especially social media platforms like Facebook and Twitter, to further propagate their anti-Muslim propaganda among Sinhalese. (Sarjoon et al.,2016). However, due to the lack of Sinhala language supported filtering technology, the relevant authorities were unable to control and stop the propagation of hatred and racist comments. Therefore, it is the timely need not to allowing the publishing of offensive statements in social media using Sinhala language. The Sinhala language supported offensive statement detector should be incorporated with social media to achieve this target. An offensive statement can be defined as any statement that attacks a person or group on the grounds of attributes such as race, religion, ethnicity, nationality gender or sexual orientation. Filtering social media posts in Sinhala language plays a vital role to prevent from those devastating ethnic or religious conflicts.

## 1.3 Motivation

As a multi-cultural country, social cohesion and harmony is vital to become a developed nation for Sri Lanka. Regulating social media to get away from spreading religious extremism and racist thoughts should be done as a prioritized task by the government authorities. Due to the lack of language translators which support Sinhala, social media authorities have still failed to control and prevent the spread of hatred and racist sentiments. Simple keyword spotting techniques are not adequate to accurately identify the exact meaning or intent of the statement(Dias et al.,2018). Therefore, it is much needed to develop an efficient and effective model to detect Sinhala language based racial and religious offensive statements in social media.

## 1.4 Computer science problem

This study is conducted with the aim of developing a model to detect Sinhala language based racial and religious offensive statements in social media. Semantics of statement should be thoroughly considered in order to classify the offensive statements. In the Computer science domain, this is basically a natural language processing problem where machine learning algorithms can be applied to solve the problem. The proposed model should be able to take social media posts or comments as string inputs and detect those inputs with racial and religious offensive language.

## 1.5 Project objectives

This study is conducted with the main aim of developing a model which can be used to filter Sinhala language based racial and religious extremism statements in social media. It will be very useful to relevant authorities to prevent publishing comments or posts which damage ethical and religious harmony in Sri Lanka. To achieve this aim following objectives are identified.

- To propose a model to detect Sinhala language based religious and racial offensive statements in social media.
- To implement the proposed model as a Sinhala language based religious and racial offensive statements detector in social media.

- To detect Sinhala language based religious and racial offensive statements in social media.
- To block Sinhala language based religious and racial offensive statements in social media.

## 1.6 Scope of the study

It is impossible to provide a thorough definition for the phrase “offensive statement”. However, a statement which tends to be defined as an offensive statement that targets groups or individual in a way that could promote violence or social disorder. Furthermore, statements which cause someone to feel hurt, angry or upset on the grounds of race or religion have been considered as offensive statements. Even though a statement can be categorized as ”Offensive” on a ground except race or religion, have been categorized as ”Not Offensive on the grounds of race or religion” in this study. An example is stated below.

Statement A

පලයන් අබ්බගාතයා යන්ඩ. කකුලත් ඇද ඇද එනවා අපිට උගන්නන්ඩ.

Even though the statement A can be categorized as ”Offensive” on disability ground, it has been categorized as ”Not Offensive on the grounds of race or religion” in this study. The reason for that is, statements in relation with race and religion are only considered in this study. Even though platforms such as Facebook, Twitter, Instagram, Viber belong to social media, only Facebook and Twitter are focused here.

## 1.7 Organization of the thesis

The background and existing literature related to this study is reviewed in Chapter two. The third chapter dives deep into the design architecture of the proposed model. The evaluation results are discussed in Chapter four, and the fifth chapter concludes the thesis with a conclusion and a discussion of future work.

# Chapter 2

# Literature review

Over the years, offensive statements and hate speech have been extensively studied in machine learning and natural language processing areas aiming to develop semantic analysis systems. A statement can't be classified as an offensive statement merely looking at the words. The positions of the words in a statement and the intent of the statement should be thoroughly analyzed in order to detect the offensive statements more accurately. This chapter will cover the definitions of offensive language and the work related to its automatic detection using lexicon based models and with machine learning models, as well as the features used for classification, for English and specially for Sinhala language.

## 2.1 Natural language processing (NLP)

Natural Language Processing is a computer science discipline in where computer systems are being used to analyse, understand, classify, interpret or generate natural languages. NLP has been widely applied in wide range of areas such as automatic reasoning, information retrieval, knowledge representation, relationship extraction, semantic web, entity recognition, speech recognition etc. Basically, there are five main components of Natural Language Processing in artificial intelligence as morphological and lexical analysis, semantic analysis, syntactic analysis, discourse integration and pragmatic analysis. Out of five main components, only the literal meaning of words and phrases is considered in semantic analysis. This basically, abstracts the real meaning or the dictionary meaning from a particular context. In this study, semantics of complete statements are analysed since this is a study which proposes a model to detect offensive statements.

## 2.2 NLP use cases

NLP techniques enable computers to understand natural languages as humans do. Basically, Data pre-processing and algorithm development are the two basic aspects of natural language processing. Once data has been pre-processed using one or more preprocessing methods, an algorithm should be developed to process data. Generally, two main types of algorithms can be seen in NLP. One approach is rules-based systems that use carefully constructed linguistic rules. That approach has been applied earlier and it is still being used. Machine learning-based system is the other type of algorithms in NLP. Statistical methods are used in these algorithms. They learn to do tasks based on training data that is provided to them, and when more data is fed, they change their approaches.

NLP has been used for perform various tasks. NLP has powered translation tools that can be used to translate low impact content like regulatory texts, emails etc. NLP can be also used for advertising and brand monitoring. Billions of social media interactions can be analysed in order to find out what customers are expecting and what should be developed to expand the business. Chat box and call center operations are using NLP techniques in world wide. In addition to that sentiment and context analysis is another popular application of NLP. Here, NLP assists in identifying and classifying texts according to the context of what are being discussed. Hate speech recognition and offensive language detection are fallen under the sentiment and context analysis. Even though the freedom of expression is a human right, it is very important to control the expressions of extremist people for the sake of other's rights to ensure the long lasting peace. For achieving that target, detection of offensive language has become a timely need.

## 2.3 Definitions of offensive language

Offensive language and hate speech are umbrella terms that are frequently used to denote offensive content on social media. Offensive language varies greatly, ranging from simple sentences to much more severe types of the language. Therefore, it's very difficult to clearly define the term of "offensive statements". Many scholars have mentioned how difficult it is to define offensive language while annotating data because it is often subjective to individuals(Dias et al.,2018). Someone who does the annotating should have a common cultural and social background and good understanding of the different versions of particular language. Here, we define any statement that attacks a person or group on the grounds of attributes such as religion, race, ethnicity, gender, disability and sexual orientation which cause someone to feel hurt, angry or upset as an offensive statement (Dias et al.,2018). Even though a statement can be categorized as ”Offensive” on a ground except race or religion, they have been categorized as ”Not Offensive on the grounds of race or religion” in this study since offensive statements which related to racism and religious extremism have been considered. An example is stated below.

Statement B

පිරිමි වගේ නෙමෙයි ගැණුන්ට ඕනි කොහෙටහරි වෙලා ඕපාදූප හොයන්ඩ

Even though the statement B can be categorized as ”Offensive” on the ground of gender, it has categorized as ”Not Offensive on the grounds of race or religion” here.

## 2.4 Various approaches to detect offensive statements

Many research have been conducted in the fields of hate speech detection, online harassment detection and offensive statements detection. Here, the strengths and limitations of the existing approaches are discussed. In the majority of the studies, they have applied either lexicon or machine learning approaches. Lexicon approaches rely solely on keywords that contain offensive words that are commonly used in hate speech. If it contains at least one offensive word, the statement is classified as an offensive statement or hate speech. The simplicity and independence of training data, as well as the easy adoption in other languages by providing adequate lexica by experts, are major advantages of these approaches(Bretschneider & Peters,2017). However in the real context, the intention should be “offensive” in order to classify a statement as an offensive statement.

Statement 01

හම්බයො දෙමලු අපි ඔක්කොම ශ්‍රී ලාංකිකයො උන්ට ගහන්න මරන්න යන්න එපා

Traditional lexicon approach may have put above statement 01 with racist label, since it contains few words which can be considered as offensive words. However if the meaning of the statement is analyzed carefully, that statement can’t be classified as an offensive statement. This is the major pitfall in lexicon based approaches.

In contrast, machine learning approaches rely on training data to automatically learn criteria to recognise hate speeches. Features should be extracted from training data as numerical inputs and those inputs should be fed to whatever neural network or support vector machine(SVM) . These features are derived by experts from characteristics of hate speech messages and include, for example, the presence of offending words defined in a lexicon and the presence of words typically referring to persons(Bretschneider & Peters,2017). The performance of these classifiers are slightly better with compared to lexicon approaches. However, gathering a significant amount of training data is a challenge.

In recent past, deep learning approaches have gotten a lot of attention to tackle the problem of offensive language detection to achieve higher efficiency(Alshalan & Al-Khalifa,2020). The majority of the research studies have used various convolutional neural network (CNN) and recurrent neural network architectures (RNN). Park and Fung have proposed a two-step classification approach by combining two classifiers. One classifier has been used to determine whether the text is abusive or not, and another to determine which form of abusive language the text contains provided that the text is abusive. In a nutshell, offensive language detection approaches can be basically divided as lexicon approach and machine learning approach. Machine learning approaches are more suitable for complex problems since lexicon approaches solely rely on words excluding the intent and the context.

### 2.4.1 Related work for English and foreign languages

Offensive language is realized in many different ways, thus its typology needs to be examined carefully. Some studies have been conducted to automatic detection of offensive languages for English, Greek and several languages. The techniques, scientists have used are ranged from rule based methods to deep learning. In 2019, Zeses Pitenis has presented a model to detect offensive posts in Greek social media. The data set produced for that research has been extracted using twitter API. The several models were trained with Term Frequency/Inverted Document Frequency matrices representations of word unigrams and linguistic information such as part-of-speech and dependency relation tags as features. Finally, the classification models were evaluated and it has showed significant results in identifying offensive language in Greek social media. That research addresses misclassifications produced by the traditional machine learning methods and provides an overview of several obstacles imposed on the classifiers for offensive posts detection in Greek(Pitenis,2019).

Zaghi(2019) has introduced a model to detect hate speech in social media. The system was designed too perform a binary task, made by a linear SVC model with unbalanced class weights using various linguistic features. They implemented the system using the Scikit-Learn Python toolkit using default values for the other hyper-parameters. In addition to that, they adopted this model for the size of the distantly supervised datasets and their unbalanced labels distribution. Two groups of surface features namely unigrams and bigrams have been used as features.

Furthermore, a research has been conducted to detect aggression in social media using deep neural networks(Medisetty & Desarkar,2018). They developed an ensemble-based system for labeling input posts into three categories: Overtly Aggressive, Covertly Aggressive, and Non-aggressive. Three deep learning techniques as Convolutional Neural

Networks (CNN) with five layers, Long Short Term Memory networks (LSTM), and Bi directional Long Short Term Memory networks (Bi-LSTM)have been used in this study. Still a majority voting based ensemble method issued to combine these classifiers together. They have trained the method on Facebook comments data set and tested on Facebook comments (in-domain) and other social media posts (cross-domain). This method can be described as a deep learning based method to detect aggressive statements.

A team from University of Antwerp, Belgium presented a model to classify Dutch posts as racist posts and non-racist posts. They have conducted two experiments in which multiple classifiers were trained on the same training set in both. This training set consists of Dutch posts obtained from two publicly accessible Belgian social media pages that are expected to elicit racist responses. The Support Vector Machine technique is used in all of the classification models, however with different sets of linguistic features, such as lexical, stylistic, or dictionary-based features. The best-performing model in both experiments uses a dictionary with various word categories specifically related to racist discourse(Tulkens et al.,2016).

Research studies have been conducted to detect abusive content on social media not only for English language, but also for complex and ambiguous language like Arabic. Haddad et al. (2020) used a Bidirectional Gated Recurrent Unit and Convolutional Neural Network which are deep neural networks to detect offensive language in Arabic social media. They have used Word2Vec Arabic model as the feature representation and proposed model have been tested with various pre-processing and oversampling techniques to enhance the performance. For the task of offensive language detection, they have obtained an F1 score of 0.859, and for hate speech detection, they have achieved an F1 score of 0.75. Those are good numbers considering the fact that On social media, Arabic content is noisy, with a variety of dialects, and most Arabic users are unconcerned about proper language and spelling.

### 2.4.2 Related work for Sinhala language

A research was conducted to analyze racist and non-racist comments in Sri Lankan context. They have used a two class support vector machine mechanism to train the network with two sets as racist and non-racist comments obtained from Facebook(Dias et al.2018). The model's main objective is to predict whether or not a given statement is racist, which turns into a two-class classification problem. SVMs are used since they are universal learners. The proposed model has produced experiment results with accuracy of 70.8% and a precision of 100%. However, they have only considered about racial comments in social media.

Smith and Thayasivam(2019) has conducted a research to detect Sinhala and English words code-mixed data that could be the first research in code-mixed data concern. The model has been developed only to identify the language not to deal with meanings or semantics. For an example, the statement "අද මම university යනවා" is a code-mixed data with English and Sinhala Unicode languages.Word-level n-grams, character-level n-grams and BOG have been used as features to train deep neural networks, recurrent neural networks, CNN, XGB, and LSTM machine learning models. It's noteworthy to state that XGB outperformed all the other models with the accuracy of 92.1% with bi-gram features.

Nanayakkara(2018) has developed a model to classify Sinhala texts based on n-grams. The model has been trained using n-grams and tfidf vectorization method. The model has performed the task with 76% training accuracy with 70% testing accuracy. It can be used to classify news lines to two classes as "International' and "Local". However author hasn't trained the model with Sinhala fonts and phonetic representation used to represent Sinhala words. For an example, a local news line has represented like "me adupadu niveradi kara ganimin idiri metivaranayata ya yutu bava e mahata vediduratat prakasa kaleya". After pre-processing the data, the model was trained using an n-grams-based technique to predict Sinhala news whether they are local or international.

Malaviarachchi and Jayalal(2020) created a model for classifying Cyber bullying comments made in Sinhala language on social media. They have collected Twitter comments with offensive words. Outliers have been removed from the Twitter comments and the remaining tweets have been pre-processed. Five rules have employed to extract features from the text. The proposed model was trained and tested with Support Vector Machine, K-nearest neighbor and Naïve Bayes algorithms. It has achieved the F1 score of 91% when they are applying SVM with a RBF kernel. Even though they achieved higher F1 score, there are few limitations in that model. Tweets with words less than 6 and Tweets with words more than 23, have been considered as outliers. However, Tweets with lesser words and Tweets like paragraphs can be seen in real world. There were 292 tweets left after removing outliers. It can not be achieved higher results with a model which have been trained using a small data set in today's context.

## 2.5 Summary

In a nutshell, the most of the research have applied either lexicon or machine learning approaches. Lexicon-based approaches rely solely on a lexicon of offensive words commonly used in hate speech. If a text contains at least one offensive word, those models classify it as hate speech. The simplicity and independence of training data, as well as the ease of adoption in other languages by providing suitable lexica by experts, are important advantages of these systems. (Bretschneider & Peters,2017). Machine learning algorithms, on the other hand, rely on training data to develop rules for classifying hate speech messages automatically.

Having observed the literature, it's clearly observed that even though there are so many offensive language classifiers for foreign languages(Zaghi,2019;Medisetty & Desarkar, 2018;Tulkens et al.,2016) lack of classifiers which supports Sinhala language where statements are in Sinhala Unicode characters is a real problem. To address the shortage of trained human resources in the form of language interpreters in Sinhala language, there should be a language classifier that can be recognized offensive statements with respect to racism and religious extremism. The precision and accuracy of the model which developed by a team from University of Sri Jayawardanapura has dropped once the data corpus size was increased(Dias et al.,2018). When considering the model develop by

Malaviarachchi et at(2020), Tweets with words in between 6 and 23 have been taken into training and other have been eliminated as outliers. In addition to that, they have tested the model with only 292 Twitter comments which quite low. Hence, the basic aim of this study is to producing a comprehensive model which can be used to the offensive statements with respect to racism and religious extremism with higher accuracy and precision.

# Chapter 3

# Methodology and design

In order to address the research problem discussed, a machine learning based model is proposed. Under this chapter, we describe the methodology that utilized to solve the problem of detecting racial and religious offensive statements in Sinhala language on social media. The steps which have been followed in order to build and train the model are preparation of data sets, data pre-processing, feature extraction and training the model with three prominent machine learning classifiers namely Support Vector Machines, Logistic Regression and Naive Bayes. The mentioned steps are thoroughly explained under this section.

## 3.1 Preparation of data sets

A database of statements was developed by extracting Sinhala language based comments from popular Facebook pages and Twitter profiles that can be easily recognised as places from where racism and religious extremism thoughts arise. Majority of the pages which were focused here owned by Sinhala extremist groups. It was observed that majority of offensive comments have been targeted the Muslim people rather than the Tamil people since people may have thought it would be useless to publish a comment targeting Tamil people from whom most of them can’t understand or express their ideas in Sinhala.

The considerable portion of the statements which were annotated as ”Not Offensive on the grounds of race or religion” also have been selected from those pages. The main reason for that is more or less same vocabulary have been used for both kinds of statements. As mentioned earlier it was quite harder to label the data manually into two classes as ‘Offensive’ and ‘Not Offensive on the grounds of race or religion’. The data was annotated by the author with the assistance of an expert in Sinhala language. It is obvious that it can’t be merely used traditional keyword based approach when assigning labels to statements since statements which consists some racial based keywords might not be categorised as offensive if the intention and semantics of the statement has been considered. The obtained data set has total of 1250 entries, which are separated into two sets as training and validation, with an 80:20 split.

Statement 01

හම්බයො දෙමලු අපි ඔක්කොම ශ්‍රී ලාංකිකයො උන්ට ගහන්න මරන්න යන්න එපා

Key word spotting techniques may have put above statement 01 with offensive label, since it contains few keywords which can be considered as offensive words. However if the semantics of the statement and intention are concerned carefully, that statement can’t be classified as an offensive statement. Having understood this kind of issues, the database of statements was annotated carefully and part of that annotated database has shown below.

| ID | Statement | Label |
|---|---|---|
| 1 | මේ රට කරන්න මට ඉඩ දෙන්න | Not offensive on race,religion |
| 2 | මූ තාම සිංහලයට කෙළින්න එනවා | Offensive |
| 3 | සැබෑවටම ජාතිවාදය ප්‍රතික්ෂේප කරන්නන් අවශ්‍යයි | Not offensive on race,religion |
| 4 | අල්ලා කිව්ව නිසා කිසිම අප්පිරියාවක් නැතුව ඔටු මුත්‍රා බොනවා, හොඳ වෙලාවට ඔටු ගූ කන්න කියලා නැහැ නැත්නම් ඒකත් කයි | Offensive |
| 5 | අපේ ලංකාවෙ පර තම්බින්ට රඟන්න දෙන්න බැ | Offensive |

Table 3.1: Label annotation of statements

## 3.2 Pre-processing

When working with text in Natural Language Processing, text pre-processing is always a mandatory step. There are a variety of words with incorrect spelling, special characters, numerals, emojis, and other items in real life human writable text data. Cleaning this type of noisy text input before feeding it to a machine learning model is critical. Sinhala is a morphologically rich language which is being used by more than 21 million speakers and it constitutionally recognized as a one of official language in Sri Lanka. It has taken a long time to evolve into its current form, with influences from a number of languages such as Pali, Tamil, Portuguese, and English. Therefore, the same word can be written in different forms unlike in English, even though the spellings are wrong. The comment “බුදු සරණ වේවා” can be written as “බුදු සරන වෙවා” without losing its meaning. That is a special thing to be considered when the pre – processing is done which is not required in English. Having considered the nature of the Sinhala language pre-processing stage has been performed with three steps.

### 3.2.1 Simplifying Sinhalese characters

Social media comments and posts are frequently informal, unstructured, and sometimes misspelled. It is easy to identify same word with different misspelled words by simplifying Sinhalese characters. Without having this step the same word might be recognized as different words merely because of its spelling. Simplifying characters dictionary has been used for achieving this purpose.

```
(
”ඛ”: ”ක”,
”ඝ”: ”ග”,
”ඟ”: ”ග”,
”ඡ”: ”ච”,
”ඣ”: ”ජ”,
”ඦ”: ”ජ”,
"ඣ": "ජ",
"ඨ": "ට",
"ඪ": "ඩ",

"ණ": "න",
"ඳ": "ද",
"ඵ": "ප",
"භ": "බ",
"ඹ": "බ",
"ශ": "ෂ",
"ළ": "ල"


"ඈ": "ඇ",
"ඊ": "ඉ",
"ඌ": "උ",
"ඓ": "එ",
"ඖ": "ඔ",


"ෑ": "ැ",
"ී": "ි",
"ූ": "ු",
"ේ": "ෙ",
"ෝ": "ො",
"◌aa": "◌a"
)
```

### 3.2.2 Removal of special characters

Special characters including punctuations don't make any sense when we are classifying comments. As a result, we must carefully select the list of punctuation that is going to be discarded from the data set. Here, the other special characters which should be needed eliminate must be included in the list as well. Few of the special characters which have been removed from the comments have been stated below.

[ ! , . , !!! , @ , , , ?, ∗]

The statement “අද පල්ලිය ගාව @මහනුවර. ගෙරි තම්බියා මක බැවිලා පල!” is converted to “අද පල්ලිය ගාව මහනුවර ගෙරි තම්බියා මක බැවිලා පල” after simplifying and removal of special characters step.

### 3.2.3 Tokenization

Before going into feature engineering, this is the process of extracting words as tokens from comments. Generally, tokens are separated by whitespace characters such as space and line brake or by punctuation characters.

| Statement | After pre-processing(As separate words) |
|---|---|
| ජාතිවාදය පිළිකුල්. එන්න අපි යහපත් ලොවක් හදමු. බුදු සරණයි!!! | ජාතිවාදය පිලිකුල් එන්න අපි යහපත් ලොවක් හදමු බුදු සරනයි |
| අපේ ලංකාවේ පර තම්බින්ට රඟන්න දෙන්න බෑ. | අපෙ ලන්කාවෙ පර තම්බින්ට රගන්න දෙන්න බැ |
| හම්බයින්ගෙ වැඩිවීම 150%. සිංහලයිනි තොපි බුදිද? | හම්බයින්ගෙ වැඩිවිම සිංහලයිනි තොපි බුදිද |

Table 3.2: Statements after pre-processing

## 3.3 Feature engineering

Basically, all machine learning algorithms including logistic regression or SVM need input data to generate outputs. The features in this input data are usually in the form of structured columns. Those features could be represented as numerical values that can be fed into the machine learning model. The goal of the feature engineering is transforming text data into feature vectors and creating new features using existing data set. A feature is a distinct measurable quality or characteristic of the phenomenon being investigated.

A vital step in developing good pattern recognition and classification algorithms is selecting distinctive, informative, discriminating, and independent features.Despite the fact that most features are quantitative, structural features such as strings and graphs can be utilized to recognize syntactic patterns.

### 3.3.1 Different feature extraction methods in NLP

To call a feature as a useful feature, it must have a relationship to the target that the model is able to learn. Linear models, for instance, are only able to learn linear relationships. So, when using a linear model, our goal should be to transform the features to make their relationship to the target linear. There are so many mechanisms which have been used by researchers to extract features from text data. Feature extraction from text data is very popular since so many people are being dealt with text data classifications in wide area such as temporal trend classification, risk management and cyber crime protection. Generally, the feature extraction techniques are ranged starting with some basic techniques which will lead into advanced NLP techniques.

The basic techniques can be used even without the domain knowledge and sufficient knowledge of NLP. The number of words in a text is one of the most basic features that can be easily extracted. Number of words feature can't be used a useful feature when we are doing a classification of texts related to semantics. However it is suitable for a liner classification of tweets assuming The negative sentiments have fewer words than the positive sentiments. Here is an example below.

| Tweets | Word count | Label |
|---|---|---|
| I love you very much mom | 06 | Good |
| This is a fantastic city, love to be here | 09 | Good |
| Kill you | 02 | Bad |
| Oh get out | 03 | Bad |

Table 3.3: Number of words in a statement as a feature

There are many basic feature extraction techniques apart from number of words such as number of characters, average word length, number of stop words, number of special characters, number of numerals, number of uppercase words and number of lowercase words.

In addition to basic techniques, there are advanced feature extraction techniques such as Count Vectors, N-grams, Bag of Words, Sentiment Analysis Term Frequency, Inverse Document Frequency, Term Frequency-Inverse Document Frequency (TF-IDF) and Word Embedding. A contiguous sequence of n items from a given sample of text or speech is called an n-gram. If the items which are concerned are characters they are called as character n - grams while if the items are words then they are called as word n-grams. A ”unigram” is an n-gram of size one while a ”bigram” is an n-gram of size two.

Character n-grams are the most successful sort of feature in authorship identification and are frequently utilized in the text classification arena(Sapkota et al.,2015). Character n-grams have the advantage of being language independent, as they may be adapted to a new language with no additional effort. In addition to that, character n-grams reflect information about their content and context. It is very much applicable to a research like this because of we are trying to detect offensive statements in social media in Sinhala language which is less known and morphologically rich language. Having the knowledge of relationship between adjacent characters or words is much needed to find the semantics of a statement because we are using machine learning approach to classify statements rather than traditional lexicon approach which is working based on keywords. N-grams are more suitable to achieve that target. A study conducted on Language Detection in Sinhala-English Code-mixed Data(Nanayakkara,2018) has shown that For most models, the character n-gram ensures the best accuracy, and after testing different n-grams, bigram proved to be the most accurate for all models.

Statement 01<br>
හම්බයො රටින් පන්නපල්ලා අපි හෙළයො බොලව්

Word Unigrams - හම්බයො | රටින් |පන්නපල්ලා | අපි | හෙළයො | බොලව්

Bigrams - හම්බයො රටින් | රටින් පන්නපල්ලා |පන්නපල්ලා අපි | අපි හෙළයො | හෙළයො බොලව්

Trigrams - හම්බයො රටින් පන්නපල්ලා | රටින් පන්නපල්ලා අපි | පන්නපල්ලා අපි හෙළයො

| අපි හෙළයො බොලව්

Statement 02

අපි යක්කු

Character Unigrams - අ | ප | ි |ය | ක | ් | ක | ු

Bigrams - අප | ප ි | ිය | යක | ක ් | ්ක | ක ු

Trigrams - අප ි | ප ි ය | ියක | යක ් | ක ්ක | ්ක ු

### 3.3.2 TF-IDF vectors as features

A technique called Term Frequency — Inverse Document Frequency (TF-IDF) is used to quantify a word in a document. In general, each word is assigned a weight that represents its importance in the document and corpus. This method is well-known in the fields of information retrieval and text mining(Scott,2019). Even though, human can understand a simple sentence for example “He is my brother, since they know the semantics of the language well, computers cannot understand a sentence by looking at their words. The data should be in numerical type, and then only computer can understand the data. Therefore, for this reason all of the text data should be converted into numerical vectors in order to facilitate the computer to understand the text better.

One of the main applications of TF-IDF is Google search engine. Google has already been using this technique to rank our search content for a long time, as the search engine seems to focus more on term frequency rather than on counting keywords. Search engines employ the TF-IDF to better understand content that is undervalued. As an example, when we search for “America” on Google search engine, Google may use TF-IDF to figure out if a page has a title as “America. The main logic behind the TF-IDF algorithm is described below.

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

#### 3.3.2.1 Term Frequency

The term frequency refers to the number of times a word appears in a document. This is greatly dependent on the document's length and the generality of the words used; for example, a fairly common word like "is" can appear many times in a document. As an example, when a 100-word document contains the term "we" 20 times, the TF for the word 'we' is,
TF(t,d) = count of t in d / number of words in d
TFwe = 20/100 = 0.2

#### 3.3.2.2 Inverse Document Frequency(IDF)

The IDF is the inverse of the document frequency, which assesses the informativeness of term t. When we compute IDF, the most often occurring words, such as stop words, will have a very low value. The most important thing that should be noted is the IDF of a word is the measure of how significant that term is in the whole corpus. If the size of the corpus is 1,000,000 million documents and there are 0.2 million documents that contain the term "we", then the IDF is given by the total number of documents divided by the number of documents containing the term "we".

IDF(t) = N/DF
IDF (we) = log (1,000,000/200,000) = 0.69
(TF*IDF) we = 0.2 * 0.69 = 0.138

Once the extracted social media statements pre-processed, character n-grams are generated. The tf-idf values are generated based on the n-grams. The correspondence tf-idf vectors can be fed into the machine learning models. The key benefits of tf-idf are that it is simple to compute, that it is a basic metric for extracting the most descriptive terms in a document, and that it can be used to quickly compare the similarity of two documents. However, the basic disadvantage is that it can be only used as lexical feature. Even though this study is based on semantics of statements, tf-idf can be used to feed as feature vectors to three machine learning algorithms since three separate experiments are conducted for three classifiers. Based on the experiment results, the best classifier to detect the Sinhala language based racial and religious extremist statements can be selected. Logistic regression, support vector machine and Naive Bayes classifier have been used as classifiers.

## 3.4 Logistic Regression(LR)

One of the most often used machine learning methods is logistic regression (LR) which has been used for regression analysis and solving problems of classification over the past decades. Basically, Logistic Regression predicts the output of a categorical dependent variable(100). Therefore the outcome must be a categorical or discrete value which can be such as true or false. However, instead of giving exact values such as 0 and 1, the LR algorithm gives probabilistic values that fall between 0 and 1. Except for how they are applied, Logistic Regression is very similar to Linear Regression.For regression problems, Linear Regression is applied, while for classification problems, Logistic Regression is used.

Although the linear regression model is effective for regression, it is ineffective for classification. As a result, logistic regression is the better approach for classification problems. The logistic regression model, rather than fitting a straight line, employs the logistic function to compress the output of a linear equation between 0 and 1. LR based models can be used to create an email classifier to detect spam emails, to predict whether a tumor is benign or malignant using radiological images and to predict whether a customers will either default on their loan repayments or repay the loan using their bank data from the past.

### 3.4.1 Logistic regression classifier

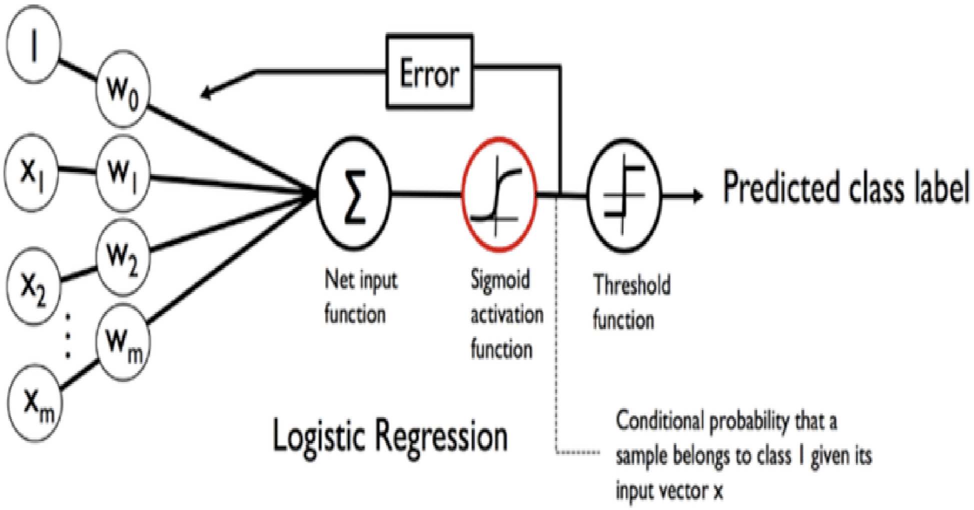

Figure 3.1: Logistic regression model

Basically, logistic regression can be considered as a one layer neural network.Logistic regression is a well-known and often used statistical model in addition to being a machine learning model. The logistic regression model takes the form of a logistic regression equation once it has been trained.

$$y = \frac{1}{1 + e^{-(w_0 + w_1 x)}} \tag{3.1}$$

In this equation, y is the predicted probability of belonging to the default class. The default class is marked with 1 while the the other class with 0 in binary classification. As an example, y=0.99 would mean that the model predicts the example belonging to class 1. During training, y is called as the target variable in machine learning. It represent the predefined classes in classification. The logistic function is employed to transform the predictions, despite the fact that logistic regression is a linear method. Hence unlike with linear regression, the predictions can no longer be understood as a linear combination of the inputs.

### 3.4.1.1 Sigmoid function

The Logistic Regression applies a more complicated cost function known as the Sigmoid function or the logistic function instead of a linear function. To convert predicted values to probabilities, the Sigmoid function is employed. Using this function, any real number can be converted to a number between 0 and 1. In machine learning, the Sigmoid function is utilized to transform predictions to probabilities(Pant, 2019).

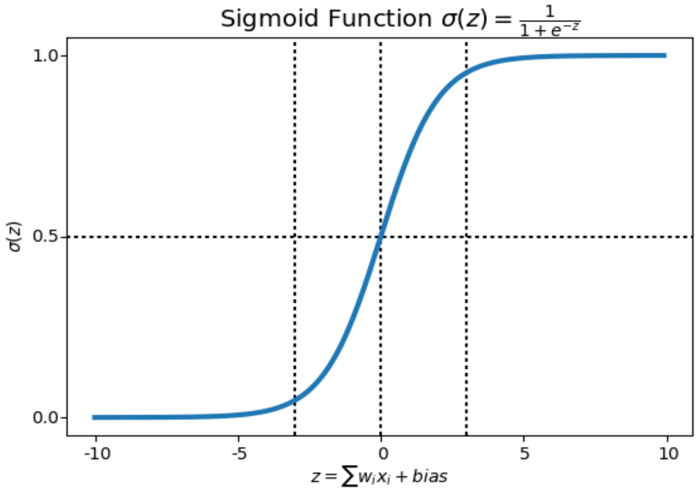

Figure 3.2: Sigmoid function

Logistic regression is very fast at classifying unknown records. That is one of the most fundamental machine learning algorithms. LR is simple to build and, in some situations, delivers excellent training efficiency.Because of these characteristics, this algorithm does not require a lot of computing resources to train a model. Over-fitting is less likely with logistic regression, but it can happen in high-dimensional data sets. In these cases, L1 and L2 regularization techniques can be used to avoid over-fitting.

## 3.5 Support Vector Machines(SVM)

The Support Vector Machine (SVM) is a widely used supervised learning model for classification and regression tasks. It is, however, mostly used in machine learning to solve classification problems. The main idea behind the SVM is finding the best line or decision boundary for categorizing n-dimensional space into classes so that new data points can be easily placed in the correct category in the future. The best decision boundary is called as hyper-plane. SVM selects the extreme points/vectors that help to build the hyper-plane. Since these extreme cases are called as support vectors, the algorithm is termed as Support Vector Machine. For linearly separable data, a linear SVM is employed, whereas for non-linearly separated data, a non-linear SVM is utilized. Non-linearly separated data is a data set that cannot be categorised using a straight line.

Suppose a SVM is used to classify Sinhala lannguage based comments in to two classes. Here, the the model should be able to accurately identify whether a given comment is an offensive or not. First, the model should be trained with lots of statements of both offensive and not offensive statements which facilitates SVM to learn about different features. Finally SVM creates decision boundary between these two classes and chooses extreme cases of offensive and not offensive data. Those extreme cases are known as support vectors. SVM will classify comments as offensive or not offensive based on the support vectors.

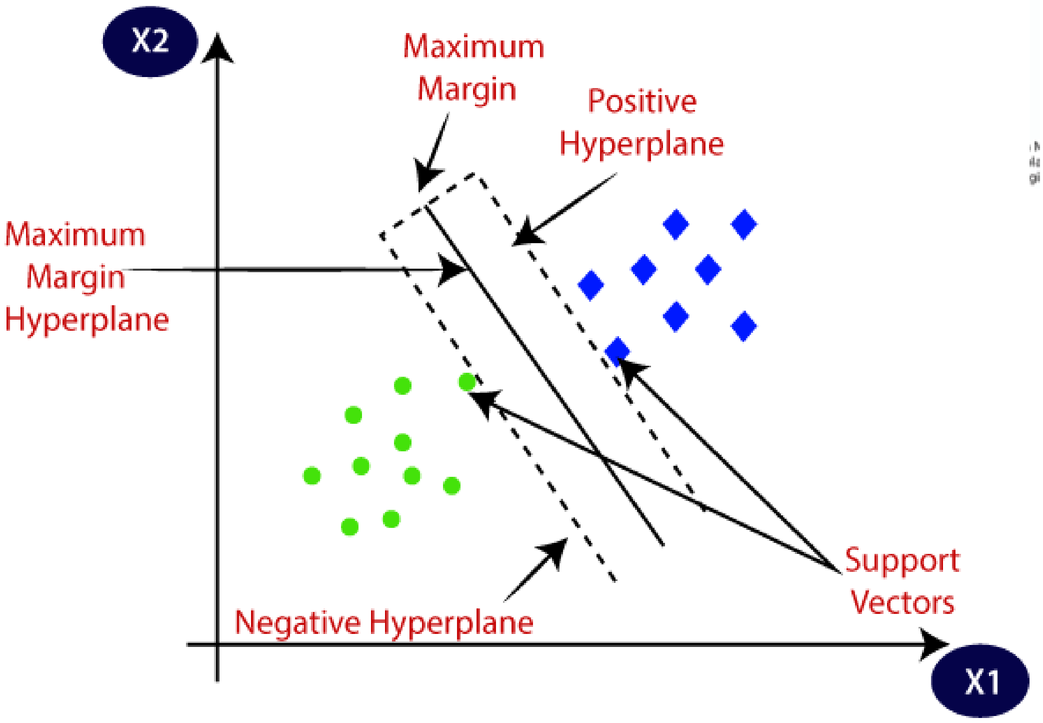

Figure 3.3: Support Vector Machine classification

Generally, the main aim of SVM is to correctly classify data that hasn't been seen before. SVM is a supervised machine learning algorithm that may be used for both classification and regression. It is powerful yet versatile. In real time, SVMs have been used in variety of areas. One of the common application where SVM is used is face detection. SVMs divide the image into two areas: face and non-face, and draw a square border around the face. In addition to that, the use of SVMs improves picture classification search accuracy. It outperforms typical query-based searching techniques in terms of accuracy. SVM is used to detect gene classification, patient classification based on genes, and is applied for protein remote homology detection Bioinformatics(Gour, 2019). Specially, SVM has been used for handwriting recognition and text categorization. That is the main reason SVM is focused in this study.

## 3.6 Naive Bayes classifier

The Naive Bayes classifier is a Bayes theorem-based supervised learning algorithm. It is particularly useful in Natural Language Processing areas such as identifying spam emails, sentiment analysis and categorising news articles into different fields. Naive Bayes classifier requires training data for classification as other supervised learning algorithms. Since the quality of the training data affects the accuracy of classification, it's important to collect qualified training data set. Otherwise, data may be classified mistakenly by NB. With a Naive assumption of no link between distinct features, the algorithm employs Bayes theorem. As per the Bayes theorem:

Posterior = likelihood * proposition / evidence

P(A|B) = P(B|A) * P(A)/P(B)

When it comes to text classification, a database of probabilities for terms appearing in data set should be calculated. Using calculated probabilities of each term and class probabilities, final predictions are done.The probabilities can be calculated using different density functions. Gaussian Naive Bayes, multinomial, Bernoulli or kernel naive Bayes can be used to achieve this purpose. Naive Bayes is a suitable algorithm for solving multi-class classification problems that works fast and saves a significant amount of time.

Real-time predictions can be made using the Naive Bayes algorithm because it is fast and efficient. For multi-class predictions, this algorithm is widely used. Using this algorithm, the probability of numerous target classes can be quickly determined. The Naive Bayes is used for sentiment analysis like customer feedback analysis. Sentiment Analysis is the process of analyzing if a target group's sentiments are positive or negative. This algorithm is used by Gmail to ascertain whether or not an email is spam. This algorithm is outstanding at detecting spam(Vadapalli, 2020). To develop recommendation systems, Collaborative Filtering and the Naive Bayes algorithm operate together. These systems employ data mining and machine learning to forecast whether or not a user will prefer a certain resource.

## 3.7 Comparison between LR and SVM

| Logistic Regression | Support Vector Machine |
|---|---|
| More sensitive to outliers | Less sensitive than LR |
| Produces probabilistic values | Produces 1 or 0 |
| Vulnerable to overfitting | risk of overfitting is less |
| Based on statistical approaches | Based on geometrical properties of the data |

Table 3.4: Comparison between LR and SVM

## 3.8 Comparison between LR and NB

| Logistic Regression | Naive Bayes |
|---|---|
| No independence assumption | Conditional independence |
| Learning - Discriminative model | Learning - Generative model |
| Even when some of the features are correlated, it performs quite well | if some of the features are dependent, on each other the prediction might be poor |

Table 3.5: Comparison between LR and NB

## 3.9 Design

The basic design of the proposed model has been presented below.

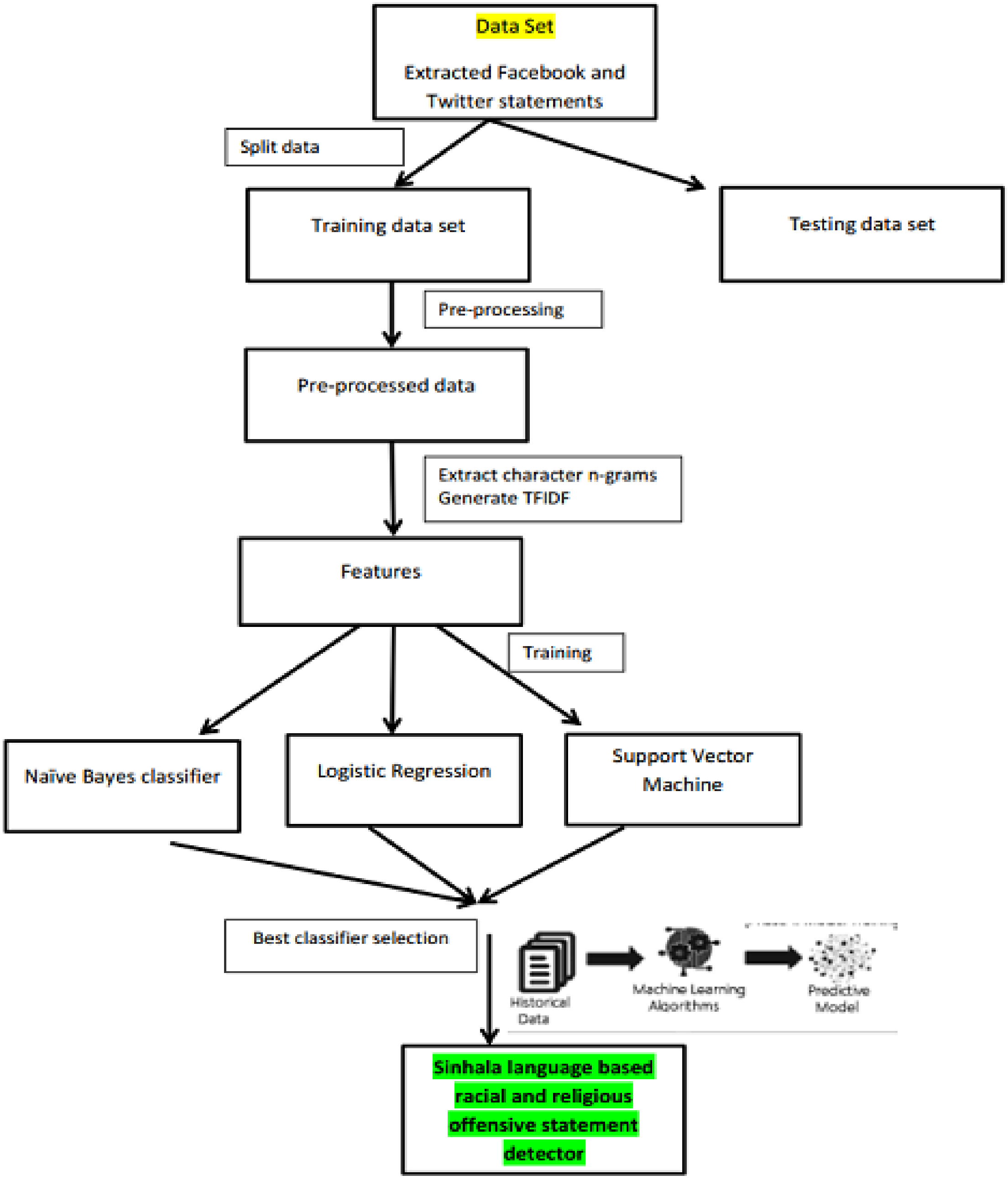

Figure 3.4: Architecture of the proposed model

## 3.10 Summary

The model is designed based on supervised machine learning technique. Once the data set is constructed, it should be pre-processed in order to remove the noise. After that, character n-grams features are extracted and those are weighted using TF-IDF.Finally, the Naive Bayes, Logistic Regression and SVM classifiers are trained separately with extracted features. As per the evaluation results, best classifier can be selected to continue with for this task.

# Chapter 4

# Evaluation

## 4.1 Evaluation approach

As discussed in previous chapters, three prominent machine learning models have been used: Logistic Regression, Support Vector Machines and Naive Bayes. A set of experiments was conducted to evaluate the proposed model for Sinhala language based offensive statements detection in social media. A binary classification task has been performed in each experiment in which statements were classified to ”Offensive” or ”Not Offensive on the grounds of race or religion” classes. In this chapter,the data sets used in experiments and the experimental setup, results including evaluation metrics were presented. Since it’s expected to measure how accurate the prediction are, an experiment based evaluation approach is the most suitable approach for this kind of study. Basically, experiments are conducted across two paths as with pre-processed training data and raw data without pre-processing. Finally, a comparative study has been done in order to select the best classifier.

## 4.2 Experimental setup

The database consists of 1250 statements and the data set was divided between training and validation data sets as 80:20. The proposed model was trained with 1000 data instances for three classifiers separately. The rest of the data set was kept as the validation data set. Ultimately, using the validation set that had been separated from the data corpus, the trained model was scored and evaluated. The composition of the two data sets were shown below.

| Label | Number of Instances | Percentage(%) |
|---|---|---|
| Offensive | 352 | 35.2 |
| Not Offensive on the grounds of race or religion | 648 | 64.8 |

Table 4.1: Composition of training data set

| Label | Number of Instances | Percentage(%) |
|---|---|---|
| Offensive | 159 | 63.6 |
| Not Offensive on the grounds of race or religion | 91 | 36.4 |

Table 4.2: Composition of validation data set

### 4.2.1 Experiments with pre-processed data

Basically, experiments have been conducted across two paths. First the proposed model was trained and validated with pre-processed data. After that, To determine how accurate the model's predictions are for the validation data set, a confusion matrix was used. Maximum features was set to 6000 and n-gram range of (1-4) was used.

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 126 | 11 |
| | Not Offensive | 33 | 80 |

Table 4.3: Confusion matrix for results obtained by LR

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 60 | 1 |
| | Not Offensive | 99 | 90 |

Table 4.4: Confusion matrix for results obtained by NB

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 115 | 11 |
| | Not Offensive | 44 | 80 |

Table 4.5: Confusion matrix for results obtained by SVM

### 4.2.2 Experiments with raw data

Experiments were conducted with raw data. Any sort of pre-processing was not applied for training or validation data in this case.

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 134 | 30 |
| | Not Offensive | 25 | 61 |

Table 4.6: Confusion matrix for results obtained by LR

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 130 | 22 |
| | Not Offensive | 29 | 69 |

Table 4.7: Confusion matrix for results obtained by NB

| | | True Class | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Predicted Class | Offensive | 130 | 21 |
| | Not Offensive | 29 | 70 |

Table 4.8: Confusion matrix for results obtained by SVM

| Classifier | Accuracy without pre-processing | Accuracy after pre-processing |
|---|---|---|
| Logistic Regression(LR) | 0.780 | 0.824 |
| Naive Bayes(NB) | 0.796 | 0.6 |
| Support Vector Machines(SVM) | 0.8 | 0.78 |

Table 4.9: Accuracy of predictions

### 4.2.3 Performance analysis

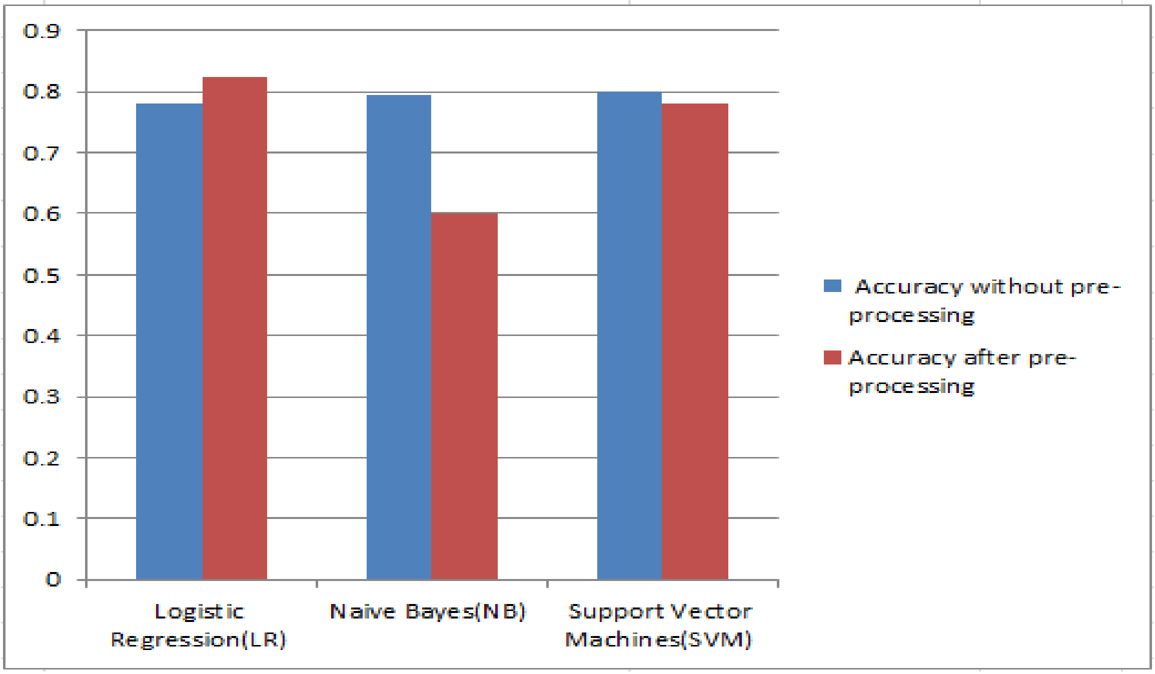

Figure 4.1: With pre-processing and without pre-processing

The accuracy of the predictions was calculated using confusion matrix. The SVM classifier recorded the highest accuracy with raw data set while LR produced the lowest.HaCohen-Kerner et al.(2020) has mentioned that,for all the tested data sets, there was always at least one combination of basic pre-processing methods that could be used to significantly improve the text classification. in addition, It was discovered that removing stop words allows for a significant improvement over the baseline result. Having this knowledge, experiments were conducted with pre-processed data. The prediction accuracy of the Naive Bayes classifier dropped significantly and the accuracy of SVM also dropped slightly.However, the LR recorded the accuracy as 0.824 and it was the the highest prediction accuracy achieved by a classifier for this proposed model.

The accuracy of 0.824 in LR classifier means it has classified 206 statements correctly out of 250 statements.It has failed to detect 33 offensive statements out of total 159 offensive statements. Still 11 false positives has been recorded by LR classifier using pre-processed data sets.Since the TFIDF weighted character n-grams have used as features, pre-processing techniques such as simplification characters and removal of special characters have been positively effected to achieve a higher accuracy of 0.824. However, the accuracy of Naive Bayes algorithm dropped significantly from 0.796 to 0.6. As per the results LR has recorded the best performance.

### 4.2.4 Comparative analysis of classifiers

To get the better understanding of the performance of classifiers, metrics such as F1score, recall and precision in addition to accuracy were calculated. Basically, When it comes to choosing the best model, accuracy is indeed not the be-all and end-all criterion. That is the main reason behind the selection of other metrics in addition to accuracy. Here, the experiments were done using pre-processed data sets. The comparative analysis of LR,NB and SVM classifiers is shown below.

| Classifier | Accuracy | Precision | Recall | F1score |
|---|---|---|---|---|
| Logistic Regression(LR) | 0.824 | 0.920 | 0.792 | 0.851 |
| Naive Bayes(NB) | 0.6 | 0.983 | 0.377 | 0.741 |
| Support Vector Machines(SVM) | 0.78 | 0.913 | 0.723 | 0.801 |

Table 4.10: Comparative analysis

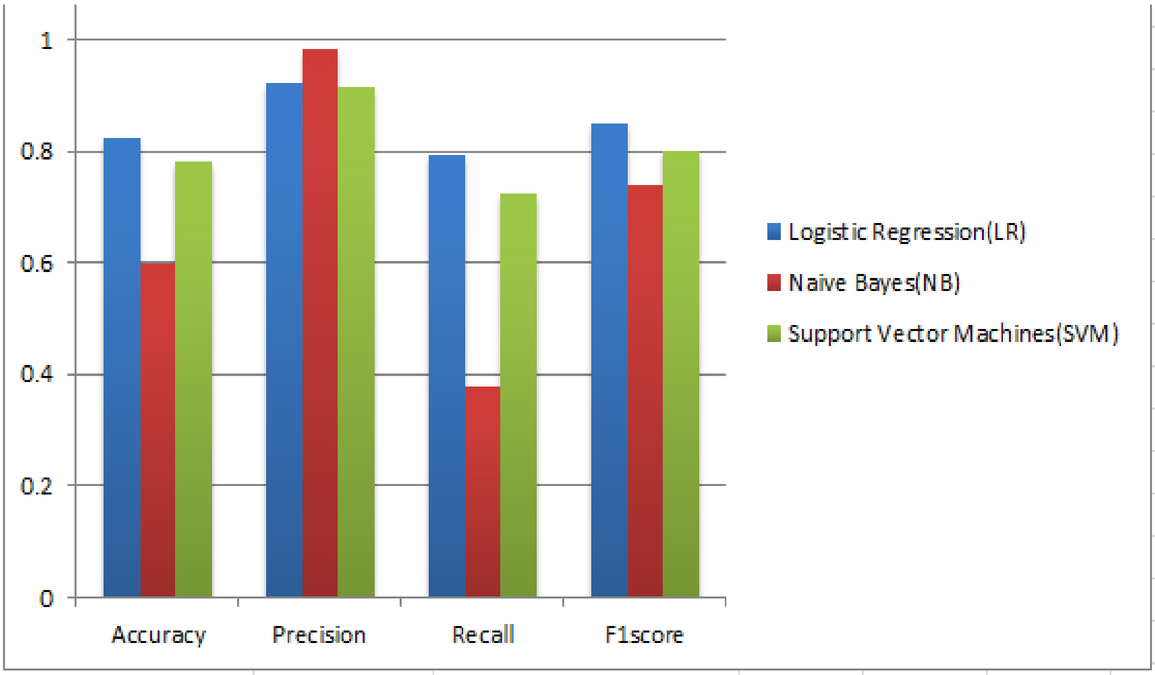

Figure 4.2: Graphical representation of experimental results

For an imbalanced classification problem like the one focused in this study, accuracy alone as a performance metric is inappropriate and inadequate. The key reason for this is because the number of instances from the majority class overwhelm those from the minority class, meaning that even a basic model can achieve higher accuracy scores of more than 90 percent depending on how severe the class imbalance happens to be.Therefore, it's much needed to use an alternative metrics such as precision and recall. It's obvious that precision should ideally be 1 high for a good classifier. The LR model recorded a higher precision value as 0.920 while Naive Bayes produced the highest precision. However the recall of the Naive Bayes was very low as it was 0.377. Therefore, ideally in a good classifier, both precision and recall should be equal to 1 which also means False Positives and False Negatives are zero. It should be highlighted that Naive Bayes has produced 99 false negatives which means it recognised only 60 offensive statements out of 99 offensive statements. Even though the precision is high Naive Bayes can't be names as good classifier here because of low recall value.

Since even though the precision and recall are useful, they do not tell the whole story, it's much needed to having a metric which can combine precision and recall together to give an idea about the classifier as the whole. The F1-score is a metric that takes both precision and recall into account and is defined as follows:

F1 Score = (2 * Precision * Recall) / (Precision + Recall)

As per the table 4.10, Logistic Regression shows the highest F1 Score while Naive Bayes shows the lowest as 0.741. Having considered the F1 Score, the LR classifier can be identified as the best algorithm for the proposed model which is designed to detect the Sinhala language based racial and religious offensive statements. LR outperformed the other classifiers while maintaining decent accuracy and precision of predictions.LR performed better with the character n-grams range up to 4. Here, the LR has been configured with regularisation parameter C=100 and the ”saga” optimization algorithm has been used.

## 4.3 Summary

Evaluation is an essential part of a construction of machine learning model for the acceptance.Here, the experiment based evaluation technique has been used and metrics such as accuracy, precision, recall and F1 score was used to analyze the performance. The performance of three classifiers studied in this study was compared in a comparative study. Logistic Regression has achieved the highest F1 score comparing with other two algorithms. Regularization has been done in order to improve the performance of LR classifier.

# Chapter 5

# Conclusion

This dissertation presents a machine learning based model to detect Sinhala language based racial and religious offensive statements.The social media engagement is rapidly growing, and some individuals have used it into a platform for spreading racial thoughts within communities. As a multi ethnic country, Sri Lanka needs the ethnic harmony to become a well developed country. Recently, Few tragedies happened in Akurana and Atulugama areas ultimately led to clash between two ethnic groups. There had been incidents where the hate spread out as a result of racial thoughts in Facebbok. In such situations Sri Lankan government had to manage the devastating situation by prohibiting the use of social media for few days. However, once the ban lifted, there were no mechanism to moderate the comments and posts in Facebook. Relevant authorities has failed to stop the spread of hate via social media platforms since they don't have capable Sinhala language interpreters to detect racial and religious offensive statements.

To solve the issue of lack of Sinhala language supported offensive language detectors, a machine learning based model was proposed with TFIDF weighted n-gram features. The data sets were prepared by extracting comments and posts from Facebook and Twitter. Data set were annotated manually and divided in to two sets as training data set and validation data set. After pre-processing the data TFIDF character n-grams were generated as features. Proposed model were trained separately using three prominent machine learning classifiers as Logistic Regression, Naive Bayes and Support Vector machines.

The validation results showed an accuracy of 0.824 and precision of 0.920 for Logistic Regression. Logistic Regression also recorded the F1 Score of 0.851 which is the highest with comparing to other two classifiers. As per the F1 Scores of three classifiers highest was 0.851 of LR classifier and lowest was 0.741 of Naive Bayes classifier. Here, the F1 Score is used to decide the best classifier since accuracy alone is not sufficient for imbalanced data set which we focused here. LR achieved this accuracy and precision, with pre-processed data sets and the classifier has been configured with regularisation parameter C=100. in addition to that, LR classifier used saga as the optimazation algorithm.

## 5.1 Future works

This dissertation presents a simple yet realistic model for detecting Sinhala language based racial and religious offensive statements. The precision of the model should be increased using optimzation techniques. Furthermore, more statements with code-mixed data can be seen in social media where people use Sinhala words but they have published those comments using English alphabet. Therefore, as future work, this model should be extended to identify offensive statements with code-mixed data. Another main thing is that there is a latest trend that uses sarcasm to spread racial thoughts or hurt another community. That is very difficult classify as offensive since the context should be carefully analysed in order to decide whether a statement based on sarcasm or not. Sarcasm related offensive statements classification is another main area to investigate in future.

Example statement: නානා ඔයා බබා. එන්න හරහා දාලා නළවන්න

Statements like example statement had been published in a page to indirectly insult a particular group. However, these statements are very difficult to classify since they are based on sarcasm. This study can be extended to address this issue as well.

# References

[1] ALJAZEERA.COM. 2020. Sri Lanka: Facebook apologises for role in 2018 anti-Muslim riots. [online] Available at: <https://www.aljazeera.com/news/2020/5/13/sri-lanka-facebook-apologises-for-role-in-2018-anti-muslim-riots> [Accessed 22- Nov- 2020].

[2] Alshalan, R. and Al-Khalifa, H., 2020. A Deep Learning Approach for Automatic HateSpeech Detection in the Saudi Twittersphere. applied sciences,.

[3] Bretschneider, U. and Peters, R., 2017. Detecting Offensive Statements towards Foreigners in Social Media. Hawai: Proceedings of the 50th Hawaii International Conference on System Sciences, pp.2213-2222.

[4] Hate Speech on Social Media: Global Comparisons", Council on Foreign Relations, 2020. [Online]. Available: https://www.cfr.org/backgrounder/hate-speech-social-media-globalcomparisons. [Accessed: 22- Nov- 2020]..

[5] Dertat, A., 2017. Applied Deep Learning - Part 4: Convolutional Neural Networks. [online] Medium. Available at: <https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2> [Accessed 10 May 2021].

[6] Dias, D., Welikala, M. and Dias, N., 2018. Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning. University of Sri Jayawardanapura.

[7] Jankiev, N., 2018. Practical Text Classification With Python and Keras – Real Python. [online] Realpython.com. Available at: <https://realpython.com/python-keras-text-classification/convolutional-neural-networks-cnn> [Accessed 16 May 2021].

[8] Madisetty, S. and Desarkar, M., 2018. Aggression Detection in Social Media using Deep Neural Networks. Santa Fe: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp.120-127.

[9] Nanayakkara, S., 2018. Sinhala Text Classification based on n-grams. [online] Medium. Available at: <https://github.com/samitha9125/SinhalaTextClassification> [Accessed 10 May 2021].

[10] Pitenis, Z., 2019. Detecting Offensive Posts in Greek Social Media. M.A. University of Wolverhampton.

[11] Sapkota, U., Bethard, S., y-Gomez, M. and Solorio, T., 2015. Not All CharacterN-grams Are Created Equal: A Study inAuthorship Attribution. Colarado.

[12] Sarjoon, A., Yusoff, M. and Hussin, N., 2016. Anti-Muslim Sentiments and Violence: A Major Threat to Ethnic Reconciliation and Ethnic Harmony in Post-War Sri Lanka.

[13] Scott, W., 2019. TF-IDF from scratch in python on real world dataset. [online] Medium. Available at: <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089> [Accessed 10 May 2021].

[14] Smith, I. and Thayasivam, U., 2019. Language Detection in Sinhala-English Code-mixed Data. International Conference on As ian Language Pro cessing (IALP), pp.228-233.

[15] Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B. and Daeleman, W., 2016. The Automated Detection of Racist Discourse in Dutch Social Media. Computational Linguistics in the Netherlands Journal 6,.

[16] Zaghi, C., 2018. Automatic detection of hate speech in social media. MSc. University of Malta.

[17] Gour, R., 2019. 8 Unique Real-Life Applications of SVM. [online] Rinu Gour. Available at: <https://medium.com/@rinu.gour123/8-unique-real-life-applications-of-svm-8a96ca43313> [Accessed 6 August 2021].

[18] Pant, A. (2019). Introduction to Logistic Regression. towards data science. Retrieved 17 July 2021, from https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.

[19] HaCohen-Kerner, Y., Miller, D. and Yiga, y., 2020. The influence of preprocessing on text classification using a bag-of-words representation. plos.org, [online] Available at: <https://doi.org/10.1371/journal.pone.0232525> [Accessed 17 July 2021].

[20] Haddad, B., Orabe, Z., Al-Abood, A. and Ghneim, N., 2020. Arabic Offensive Language Detection with Attention-based Deep Neural Networks. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools. pp.76-81.

[21] Malaviarachchi, I. and Jayalal, S., 2020. Classification of Cyberbullying Sinhala Language Comments on Social Media.

[22] Vadapalli, P., 2020. Naive Bayes Classifier: Pros Cons, Applications Types Explained | upGrad blog. [online] upGrad blog. Available at: <https://www.upgrad.com/blog/naive-bayes-classifier/> [Accessed 6 August 2021].