IDENTIFYING HARMFUL COMMENTS FOR TAMIL LANGUAGE ON SOCIAL MEDIA

Prepared by

Ms. Disne Sivalingam

198773R

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

July 2022

DECLARATION

We declare that this thesis titled “Identifying Harmful Comments for Tamil Language on Social Media” is my own work. Where any part of this thesis has not been submitted in any form for another degree or diploma at this university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged. Due references have been provided on all supporting works of literature and resources.

Name of the Student: Signature of Student:

Ms. Disne Sivalingam *UOM Verified Signature*

Date: 30/07/2022

UOM Verified Signature

Supervised By: Signature of Supervisor:

Dr. S. C. Premaratne
Senior Lecturer
Faculty of Information Technology Date: 30/07/2022
University of Moratuwa

DEDICATION

I dedicate this research report affectionately to the following:

My Research Supervisor, Dr. S. C Premaratne for dedicating his valuable time for guiding me throughout this.

My Husband and Parents for providing me with continuous encouragement throughout my years of study.

Finally for my friends who helped me on giving ideas and support.

Thank you.

ACKNOWLEDGEMENT

I would like to acknowledge and convey my appreciation for those efforts and support in one way or another contributed to the successful completion of this research.

First, I am deeply grateful for the supervision throughout the one year of my research and the help received from my supervisor Dr. S. C. Premaratne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa. I have learned so much from our project discussions. His willingness to motivate me contributed tremendously to my job.

Besides, I would like to thank all academic staff from the Faculty of Information and technology, who shared their vast knowledge throughout these two years by providing me with a good environment which influenced a lot to achieve this goal.

It is with great pleasure to thank the University of Moratuwa, Sri Lanka, for all the efforts and facilities that it has taken to contributions this postgraduate programme. Especially, my thanks and gratitude to my colleagues for their help and support to complete this research in many ways.

I am as ever, especially indebted to my parents, brother, and husband for their love and support throughout my life to improve my career.

ABSTRACT

The era of social media, such as YouTube, Facebook, and Twitter adding comments to posts are being fun in the daily life of people. But this is also used to spread hate speech and organize hate based activities increasingly nowadays. Harmful and offensive text identification on social media platforms is being a trending research area over the last few years. In a country like Sri Lanka with multiple native languages, people like to comment on social media mostly in their native language. Tamil is one of the Languages commonly used and spoken in the North and East part of Sri Lanka. In recent years people like to comment not only in their native language they also comment in more than one language. In Sri Lanka, people use Singlish (Sinhala + English) or Tanglish (Tamil + English).

Because of the rapid growth of hateful content on social media, there is an immediate need for an efficient and effective method to identify harmful content. A huge number of researches have been done and are being done for automated harmful content detection online. The complication of the Natural Language constructs builds this task very challenging.

A maximum of the research are done in the English Language. This research work aims to classify the code-mixed Tamil comments on social media by categorizing them as harmful and non-harmful by using machine learning models.

TABLE OF CONTENTS

ABBREVIATIONS

TF-IDF	Term Frequency-Inverse Document Frequency
WEKA	Waikato Environment for Knowledge Analysis
NB	Naïve Bayes
SVM	Support Vector Machine
LR	Logistic Regression
TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

LIST OF FIGURES

LIST OF TABLES

CHAPTER 01

Introduction

1.1 Prolegomena

Social media is the most interactive and important platform in today's situation, using people from all over the world in their daily life. During the last ten years, there has been a tremendous increase in user-generated content via social media platforms. People use these platforms (e.g., YouTube, Facebook, and Twitter) to post and share information about themselves and express their views about others' posts and videos. But there are not only positive comments but also negative comments in the form of threats, abuse, and insults. Therefore, it is important and urgent need to detect the harmful text from the comments.

During the last few years, there have been a lot of discussions made on the topic of identifying harmful text on social media platforms. In multiple native language countries, social media comments are mostly written as code mixed. One of the most commonly used languages in Sri Lanka is Tamil Language. In the period of social media, adding comments is being entertained in the day to day routine of individuals. People can comment on any opinions, it can be attacking an individual or attacking a group of individuals or bad opinions, etc.

Many types of research have been done on the topic of finding harmful or offensive words in comments for global languages like English. In a multidisciplinary country like Sri Lanka, many numbers of individuals in online use a minimum of two languages, primarily English+Sinhala (or say Singlish) or English+Tamil (or say Tanglish). If these comments are written in two languages then it is called a bilingual comment. People in multilingual communities express their views and opinions by adding multiple languages in the same sentence. In a bilingual setting, the whole post might be in both languages with similar words called code mixed texts or mixed-code texts. Therefore, harmful texts are generally written in non-native code-mixed scripts.

Therefore in this research, the aim is to identify harmful content in the English-Tamil

comments gathered from YouTube movie reviews. Here, the task aims to classify the Tamil comments on social media by categorizing them as harmful and non-harmful by using machine learning models.

1.2 Background & Motivation

With population growth, the amount of social media users has increased dramatically over the last ten years. People can share and convey their thoughts on online sites like YouTube, Facebook, and Twitter. This has arisen in a flood of text data on online sites. While these social media sites let people interact with one another and help to share pieces of information, they also have negative effects such as the spread of harmful and hateful content, leading to cyber violence, and spreading misinformation leads to dividing people into groups. In recent years, both in-person and online hate speech has increased. Harmful content is being raised and propagated via social media and other internet applications, which ultimately leads to hate crimes.

Harmful speech is a type of verbal or nonverbal communication in which bias and assailment are expressed. It is characterised as deprecating an individual or community because of their gender, religion, nationality, etc[10], [11]. These harmful contents cause a lot of violence, and companies like Facebook, Twitter, YouTube, and others are under fire for increasing offensive content on their sites [12]. Hate content can corrupt not only humans but also chatbots.

It is clear that nowadays, all the social media platforms facilitate the users' native language, so the users express their thoughts without any restraint. There are many advanced controlling structures to find and eliminate harmful comments in the English language. Due to the lack of knowledge in these types of languages and there is a failure to control the harmful comments. There is a need to control and avoid these types of situations.

Considering these factors, this research focuses on identifying the harmful content in the code-mixed (Tamil+English) text on social media. Among many social media platforms,

comments are collected from one of the popular platforms YouTube to create the dataset for this research.

Automatic harassment detection relies heavily on machine learning techniques. According to our studies of online content, the number of offensive and hateful content on social media in comparison to ordinary content is less, yet it can make a serious impact on a targeted person or group. Machine learning uses training data to build a mathematical model that can make predictions or decisions without being explicitly programmed.

The objective of machine learning is to develop a model that can predict the results accurately. For that, a classification model is developed by learning from the labeled training corpus. After that using the test corpus, the performance of the developed model is obtained. Machine learning can be characterized into three types based on the data used as a training data set. These are Supervised, Unsupervised, and Semi-supervised learning. This research work used the labeled data as a training dataset so that this research work is followed as a supervised learning methodology.

1.3 Research Problem Statement

The social media platform has rules that are available to the user in the form of guidelines. However, all users do not follow the rules while commenting on a post. So there is a need to identify the harmful comments on social media comments. This research focuses on developing a simple architecture for identifying the harmful text from code-mixed comments on social media.

1.4 Aim and Objective

1.4.1 Aim

Our aim is to achieve a simple solution for harmful comments identification in Tamil Language using machine learning approaches and intend to afford an analysis study with our results of the code-mixed dataset.

1.4.2 Objectives

The main objectives of this research are:

- To collect code-mixed Tamil comments from social media as data set.
- To identify suitable preprocessing methods.
- To identify the features to extract using suitable procedures.
- To identify the best suitable classification algorithm and a deep learning model to predict harmful comments from the dataset.

1.5 Scope of Research

This research aims to predict the harmful words from the comments collected from social media by using machine learning models.

1.6 Proposed Solution

In this research, I proposed to use Labeled Tamil code mixed comments as a dataset to identify the harmful comments in Tamil code mixed on YouTube. It classifies the YouTube comments as harmful and non-harmful based on the classification model using different machine language techniques. Those models can be used to identify harmful words in Tamil code mixed comments on social media. It will assist in limiting the adverse consequence that Social Media has on society.

1.7 Overview of the Report

This chapter provides the introduction to identifying harmful comments in the Tamil Language on Social Media using machine learning techniques. The next chapter describes similar related work done by others. Chapter Three explains the technology adapted and Chapter Four presents the novel approach of the research. Then chapter Five elaborates on the analysis and design of the research, meanwhile, chapter Six is about the

implementation part of the research component. Evaluating the solution is discussed in Chapter Seven. Then Chapter Eight presents the discussion part and the last section provides a list of references.

1.8 Summary

Harmful comment identification is essential in daily life to minimize social media's negative impact. So, this research has been conducted to check the most suitable techniques for the particular dataset.

CHAPTER 02

Literature Review

2.1 Introduction

These days, the use of harmful content is becoming common via social media. In this manner, over the most recent couple of years, many researchers have used various types of feature representation techniques, Text processing, and analysis to classify hate speech content from the texts. Especially this chapter emphasizes on main deciding factors of harmful content identification for the Tamil Language in code-mixed comments on Social Media. Furthermore, this chapter summarizes most of the Text mining approaches with their pros and cons, and a list of different algorithms to apply Text mining techniques is also contained within this chapter.

2.2 Related work in Text mining for Identifying Harmful Content on Social Media Comments

The harmful content identification process in social media texts experiences many difficulties; for example code mixed online content, script mixed online content, etc. This part reveals insight into a couple of state of art methodologies introduced to deal with such problems. The vast majority of the research proposed was validated with the monolingual model for the detection of harmful content. It is nearly simple to develop a monolingual model because (a) unknown token frequency is lower in the test data, (b) it learns a single language dictionary, and (c) it is readily accessible, [1].

Aside from this, some research works were done on multilingual datasets where the comments are blended with two or more languages and a variety of methodologies were used for the identification of hateful content. (Saumya et al.) proposed broad tests utilizing various types of techniques to identify harmful content on Twitter in different code-mixed and script-mixed comments. They obtained better results using one to the six-gram characters along with TF-IDF features vectorization. And Logistic Regression,

Naive Bayes, and Vanilla Neural Network models are performing well for Malayalam code-mixed, Tamil code-mixed, and Malayalam script-mixed, respectively. Finally, concluded that machine learning models perform better than transfer learning models and hybrid deep models for their dataset[1].

(Kedia and Nandy) used an ensemble of an AWD-LSTM-based model with two different transformer model architectures based on BERT and RoBERTa. They have accomplished the values 0.77, 0.97, and 0.72 as weighted-average F1 scores in the English-Tamil, English-Malayalam, and English-Kannada datasets respectively[2]. (Dias et al.) Proposed an approach to identify the racist online comments in Sinhala by using machine learning techniques. Support Vector Machine with two classes was trained by comments which are collected from Facebook and labeled as racist or non-racist based on the intention. A value of 70.8% is obtained as the accuracy [3].

(Veena et al.) developed and evaluated four systems consisting of Long Short Term Memory networks, Logistic regression, XGBoost, and Attention networks. Character level n-grams with TF-IDF vectorized techniques were used as features and then Malayalam code-mixed comments, and Manglish (Malayalam written in Roman script) and Tanglish (Tamil written in Roman script) comments were used as datasets. The best results were obtained with an F1 score value of 0.93 [4]. (Bohra et al.) used Hindi-English (Hinglish) dataset collected from Twitter and then extracted with n-grams and lexicon-based features classified using SVM and Random Forest algorithms with a precision value of 0.71[8]. (Samghabadi et al.) proposed ensemble learning based on different machine learning algorithms, for example, Logistic Regression, SVM with word n-gram, character n-gram, word embedding, and sentiment as a feature set. In their research word embedding and character n-gram features are found as best compared to an individual feature [9].

(Aditya et al.) gathered nearly four thousand five hundred seventy-five Hindi-English tweets. Then it is preprocessed using emotion count, and punctuation count, and then the feature word n-grams are extracted. A large matrix is created by these features to reduce

the size of this large matrix chi-square is used as a dimensionality reduction strategy. To classify the data algorithms such as SVM and Random Forest was utilized, The value of 71.7% accuracy and Random Forest 66.7% were achieved with SVM[13].

(Edel Greevy et al) presented supervised learning methods for classifying the racist comments. Bi-gram feature extraction method is used by the authors to convert the text into vectors. And then Bag of Words feature representation is used. An accuracy value of 87% is achieved using the Support Vector Machine (SVM) classifier[14]. Irene Kwok et al. [15] dealt with the automatic identification of racism against black in Twitter groups using a machine learning-based method. In their work, the unigram feature extraction technique along with the Bag Of Word feature vectorization technique is used to produce the large numeric vectors. The created numeric vector is then fed into the Naïve Bayes classifier to create a mode. They got a value of 76% as maximum accuracy as outcomes. (Sanjana Sharma et al.) used Twitter data to identify the hateful content. In their work, Bag Of Word features are used. Then the created vector is fed into the Naïve Bayes classifier and a value of 73% is obtained as a maximum accuracy[16].

(In the existing research (Zeerak Waseem et al.) proposed hateful content identification on Twitter data. In their study, to generate the numeric vectors character N-gram features are used. The generated numeric vectors are fed into the Logistic Regression model and the overall 73% F-score is obtained[17]. The study (Chikashi et al.) employed a Machine Learning model to detect abusive content in web-based texts. In their work, they used the character Ngrams as features and then fed them into an SVM classifier. An overall value of 77% F-score is obtained using the above classifier[18].

Similarly in (Shervin Malmasi et al) proposed a method for hate content detection online using machine learning techniques. They used the character four grams as a feature extraction method. Finally, the generated numeric feature vector is fed to the SVM classifier to obtain the results. The value of 78% is obtained as the maximum accuracy[19].

Over the last few years, some researchers used Machine learning models to detect automatic hateful content identification. A similar study conducted by, (Karthik Dinakar et al.) presents a classification approach to sensitive content on social media posts and comments. To generate the feature vectors, word unigram features along TF-IDF feature vectorization are used. Then the generated features are fed to different four Machine Learning classifiers such as SVM, Naïve Bayes, rule-based, and J48. An accuracy value of 73% is obtained by using a rule-based classifier and they have concluded that a rule based classifier gives the highest performance compared to other classifiers such as Naive Bayes, J48, and SVM[20].

Similarly,(Shuhua Liu et al.) presented an approach for the classification of web pages classified as hate or violence. They used tri gram features with TF-IDF representation techniques in their studies. A Naive Bayes classifier is used for classification and obtained the highest accuracy value of 68%[21].

Based on the analysis done, features such as N-gram, TF-IDF, word features, character features, and word embedding were used. Classification is done using conventional learning-based models (such as RF, SVM, NB, LR), neural network-based models (such as Vanilla Neural Network, CNN, Bi-LSTM, AWD-LSTM), and transfer learning Models(BERT, RoBERTa ULMFiT, and hybrid deep models) were used. Other than the languages used by wide users of social media like English, a lot of research has also been going on for the development of social media data of the under-resourced languages[4]. This work addresses this gap of less research in harmful comments identification methods for the Tamil language with code mixed dataset.

The below Table2.1 summarise the Feature Extraction Techniques used in the existing studies.

Table2.1: Summary of Feature Extraction Techniques

Features	**References**
Word N-gram	[4], [8], [9], [13], [14], [15]
Character N-gram	[1], [4], [9], [17], [18], [19], [21]
Bag-of-words (BoW)	[14], [15], [16],
Lexicon based features and embedding	[8]
Sentiment	[9]

The below Table 2.2 summarises Feature Vectorization Techniques used in the existing studies.

Table 2.2: Summary of Feature Vectorization Techniques

Features	**References**
TF-IDF	[1], [4], [20], [21]

Based on the identifying harmful comments on the social media the traditional methods are partitioned into four parts. The techniques that are used are listed below table.

Table 2.3: Summary of Machine Learning Algorithms used for harmful content identification

Algorithms	**References**
Support Vector Machine (SVM)	[3], [8], [9], [13], [14], [18], [20]
Linera Regression (LR)	[1], [4], [9], [14]
Naïve Bayes (NB)	[1], [15], [16], [21]
Random Forest (RF)	[8], [13]
J48	[20]
Rule-bases	[20]

Most of the papers selected for review using any of the following methods for evaluating the performance as Accuracy, Recall, F1-score, and Precision. The highest performance value obtained using different classification methods from the selected papers is shown in table 2.4. The same classification method gives different performance values based on the dataset used for the research work.

Therefore, For evaluating the developed models "Accuracy", "Recall", "Precision" and "F1-score" values will be used.

Table 2.4: Summary of Performance values

Reference	**Classification method**	**Accuracy(%)**	**F1-score(%)**
[2]	AWD-LSTM	N/A	77%
[3]	SVM	70.8%	N/A
[8]	SVM	71%	N/A
[13]	SVM	71.7%	N/A
[14]	SVM	87%	N/A
[15]	NB	76%	N/A
[16]	NM	73%	N/A
[17]	LR	N/A	73%
[18]	SVM	N/A	77%
[19]	SVM	78%	N/A
[20]	Rule-based	68%	N/A

2.3 Harmful Content Identification

We are using Youtube comments, so narrow down and have a look on Youtube. Youtube has developed hateful content approaches in conferences with makers who shared their points of view, as well as master companies that concentrate on web based harassment and the spread of harmful thoughts online. Similarly, with their way to deal with other

policy violative substances, they work rapidly and delete the violating contents of their hate policies[25].

They are working mainly in the English language. But In our case, we are using English-Tamil code-mixed comments. It is more difficult to identify the harmful contents of these code-mixed comments than the English ones.

2.4 Available Tools

There are so many tools available freely to do text mining. Some of the tools are RapidMiner, Weka, Python anaconda, Phycharm, R, and Orange. Each tool has its own pros and cons.RapidMiner can be used with the Text mining Extensions and other tools also can be used by installing different plugins, libraries, or extensions.

Among the tools, we have to find out the best and most flexible tool to use for our purpose. RapidMiner with the text mining extension is selected to do this research work.

2.5 Summary

This chapter elaborates on the findings of our literature study and a summary of the previous related work. It shows different Text mining techniques and effective Text mining algorithms which are used previously to classify a large set of data to achieve dissimilar multipurpose goals.

Findings of the literature and a summary of the existing works were presented in this chapter. To identify the hate content from the social media texts, there are different types of methods used such as Machine Learning, Deep Learning, and transfer learning methods. If we consider the machining learning method, it has different sub processes such as preprocessing, Feature extraction, Feature vectorization, and building a model using the best classifiers. Finally, evaluate the developed models using different types of performance measures.

Learning from the findings we choose to do the research with machine learning methods. The experiment will be started by collecting comments from youtube and then comments will be preprocessed, feature extracted, feature vectorization and then it will be followed by developing models and different classification algorithms. Finally, the best fit classification model is found by analysing the performance measures.

In this study word n-gram technique will be used for feature extraction and TF-IDF will be used as the feature vectorization process. Then LR, SVM, and Naive Bayes will be used as machine learning techniques to build the model. After that, the models will be evaluated using performance values like Accuracy, Recall, Precision, and F1-score. Finally, the best fit model is validated using an external dataset to avoid the overfitting of the data.

CHAPTER 03

Technology adapted in Harmful Content identification

3.1 Introduction

This chapter presents text mining technology which we selected to identify hate speech effectively and efficiently in detail. Moreover, chapter three gives an impression of the selected technology which was perfectly adopted for our research. And also it explains the way of using the selected techniques that differ from the technologies chosen in the past literature.

3.2 Text Mining Techniques

By analysing, and discovering the relationship or some patterns from the large datasets we can get the information and then be able to predict something related is called text mining.

There are many techniques used, a few of them are Classification analysis, Association rule learning, cluster, and regression outlier detection. The classification technique is chosen in this study to categorise the comments into two classes (Harmful or non-harmful) using a trained model. In this study LR, SVM, and NB were used as the classification algorithm.

3.3 Machining Learning Classifiers

Building an effective way to avoid hate speech on social media requires identifying and tracking harmful content. Social Media companies like Facebook, YouTube, and Twitter invest millions of money each year in this task. This is generally because such attempts are basically based on manual control to distinguish and remove hateful or harmful

contents. The process is time taking and not possible and suit for reality [22, 23].

Machine learning deals with the computer's capacity to learn without being explicitly programmed. Rather than a person with programming knowledge defining rules for a specific work, a machine is used to carry out the work using different algorithms for the given data. As a result, machine learning is called a data-driven approach. Currently, in many areas of computer science, research machine learning approaches are broadly used by researchers. Machine learning is also important in text classification tasks.

The two main machine learning strategies are supervised and unsupervised learning. Labeled input data is used in supervised learning strategies. When the training data is not labeled, it is referred to as unsupervised learning. The data corpus is having two parts training data and testing data. Until a significant performance value is achieved the algorithm is trained using training data. This is known as the learning phase, followed by the testing phase. During the testing phase, predictions are made on the test data, and then performance is evaluated to compare the predicted label and actual labels.

Supervised learning classifiers include k-NN, LR, SVM, NB classifiers, and decision tree models. Meanwhile, because the number of researchers who attempted unsupervised learning for hate speech detection is moderately less, harmful content identification has been framed as a supervised learning task. Classifiers LR, SVM, and Naïve Bayes were selected For this research work.

3.3.1 Logistic Regression (LR)

According to (Abro et al.) Logistic Regression is defined as a prescient examination. It utilizes a sigmoid function to make an explanation of the connection between one independent factor and one or more independent factors [27].

(Ketsbaia et al.) states that Logistic Regression permits a method for joining pieces of relevant proof to estimate the probability of a specific class happening inside a specific context. Its task is to check the probability of class 'y' occurring inside context 'X'[26].

3.3.2. Support Vector Machine (SVM)

(Abro et al.) describes that "Support Vector Machine is one of the supervised learning algorithms. It builds an ideal hyperplane by gaining from training data that isolates the classifications while characterizing new data" [27].

SVM has to a great extent been utilized to construct cyberbullying prediction models and has so far been identified as successful and efficient [26]. (Ketsbaia et al.) states that SVM tries to find the most possible surface to separate positive and negative samples used for training.

3.3.3 Naïve Bayes (NB)

The research of (Abro et al.) describes the thatNaïve Bayes as a probabilistic-based supervised learning algorithm used for predicting unknown data it using the "Bayes Theorem". It deals with restrictive independence among features [27].

(Ketsbaia et al.) states that NB has progressively been carried out to build cyberbullying predictions and can be found in models delivered by various specialists. "This model expects that the text is created by a parametric model. And It calculates Bayes-optimal evaluations of the model parameters by using training data "[26].

3.4 Rapid Miner Studio

There are several Data mining tools such as SPSS Modeler, RapidMiner, Weka, and Orange available to do the data classification. Some of the tools provide a user-friendly environment by providing different analysis methods. (Jovanovic et al.) states that RapidMiner is a software tool used for text mining work processes for different tasks, going from various areas of data mining applications to various parameter optimization plans. One of the fundamental qualities of RapidMiner is its high-level capacity to program execution of complex work processes, all finished inside a visual UI, without the requirement for traditional programming skills[28]. RapidMiner gives several extensions

according to the need we can install and use. For this Research work, RapidMiner with the Text processing extension and WEKa extension is used.

3.4.1 WEKA extension

This extension has 132 operators such as 2 performance operators, 1 import operator, and 128 Modeling operators. For this research, we used the performance operator for the evaluation purpose.

The WEKA extension is shown in below Figure 3.1

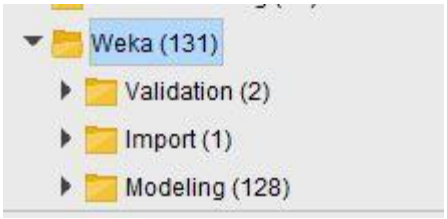


Figure 3.1: WEKA Extension

3.4.2 Text Processing

This extension has 53 operators and each is used for different purposes some of them are n-gram generation, Tokenization, stemming, etc. For the experiment tokenization, process documents from data, and Generate n-gram(Term) were used. The Text Processing extension is shown in below Figure 3.2

Figure 3.2: Text Processing Extension

3.5 Summary

This chapter describes Text mining as the technology proposed to identify the harmful contents of the Tamil Language on Social Media comments and the tools and its extensions available to do so. The next chapter shows the approach to identifying the harmful content for the Tamil Language on Social media through technology used.

CHAPTER 04

A novel approach for Identify Harmful Contents

4.1 Introduction

Chapter three discussed the technology adopted for our studies to identify harmful content from Social Media content. This chapter presents our approach to identifying harmful content from social media in detail by highlighting the Hypothesis, input, output, and process. This chapter emphasizes the key features that distinguish our novel approach from the existing approaches in various scenarios.

4.2 Hypothesis

We hypothesise that to identify the harmful comments on social media sites in code mixed Tamil-English language. This could be an issue without an appropriate method and may be accomplished via the utilisation of classification. We are trying to use methods based on machine learning algorithms for that. Under machine learning algorithms k-Nearest Neighbour, Naïve Bayes, SVM, Random forest, and Decision tree algorithms are utilised. At last, we are planning to choose the foremost reasonable and exact classifying model.

4.3 Input

In this approach, we are planning to implement a model for social media comment analysis to identify hurtful comments which are made in code mixed with Tamil-English language. For this, I have collected code mixed comments from YouTube made in the Tamil language which are manually labeled as positive or negative. At that point, the comments are saved within the excel sheet for doing the task.

4.4 Output

The fundamental results of this research work are to discover the leading classification model based on machine learning techniques to identify whether the given code mixed Tamil comments are harmful or non-harmful.

4.5 Process

The chosen corpus is divided into training and testing. A RapidMiner program installed with a text processing extension is utilised for the research. At that point, the text preprocessing, feature extraction, and feature vectorization are done. After That, several models were made by utilising a few machine learning algorithms. After that, the accuracy of all the over models was evaluated and at last, the best model was found. Then the external dataset is collected and labeled by the distinctive users. At that point, the dataset is validated using the model developed.

4.6 Summary

This chapter presented an overview of our approach to identifying the harmful content for Tamil Languages on social media. In this scenario, it is highlighted out our approach offers an efficient and precise solution for harmful content identification using machine learning techniques. The next chapter provides the design of our approach presented here.

CHAPTER 05

Analysis and Design

5.1 Introduction

This section provides the approach to developing the harmful comment identification model utilizing machine learning algorithms. Moreover, this presents the analysis and design of the proposed methodology.

5.2. The High level design of the Framework

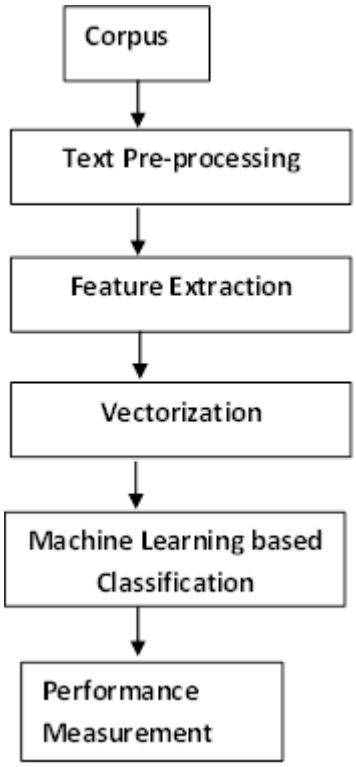


Figure 5.1: The high level design of the Framework

Data corpus was collected from YouTube comments in the initial stage. Then the collected corpus was divided into two sections testing and training. After that, the

training corpus was tokenized and removed duplicates in the preprocessed phase. Then the preprocessed corpus goes under the feature extraction and vectorization phase. Then the modeling was performed using different machine learning algorithms. Following that the test corpus was classified using the developed model. For each algorithm, the performance was measured and finally, the accuracies were evaluated to find the best classification algorithm. At that point, the best fit model can be utilised to identify the harmful words or contents in Tamil-English code-mixed comments from social media.

5.3. Summary

This chapter presents the high-level design of identifying harmful content from Youtube comments. Conjoint steps were followed to achieve the identification of harmful content in the “English-Tamil” Language on Youtube was described briefly. The implementation section is explained in the following chapter.

CHAPTER 06

Implementation of the solution

6.1 Introduction

The high-level framework of the research was discussed in the previous section. In this chapter overall implementation part of the research explains in detail.

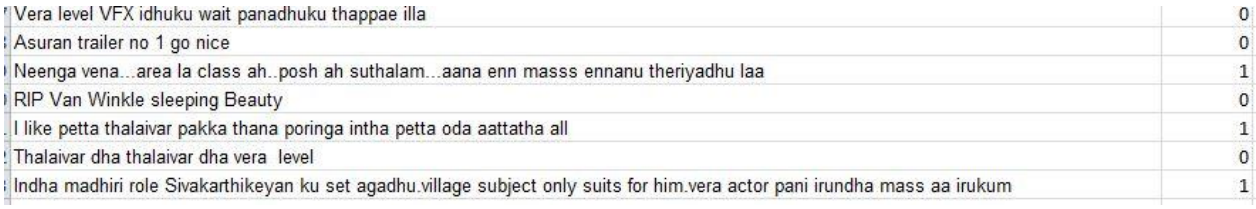
6.2 Data Corpus Construction

Corpus was gathered from YouTube comments and it includes 3000 manually labeled code mixed English-Tamil data corpus. Within that 1500 harmful and 1500 non-harmful comments are used to balance the data corpus. Excel files are used for saving the training and test data corpus.

Table 6.1: Details of Data Corpus

Type	Data Corpus	
	Training	Testing
Harmful comments	1000	500
Non-Harmful comments	1000	500
Total	2000	1000

Collected data corpus after manually labelled is shown in Figure 6.1



Vera level VFX idhuku wait panadhuku thappae illa	0
Asuran trailer no 1 go nice	0
Neenga vena...area la class ah..posh ah suthalam...aana enn masss ennanu theriyadhu laa	1
RIP Van Winkle sleeping Beauty	0
I like petta thalaivar pakka thana poringa intha petta oda aattatha all	1
Thalaivar dha thalaivar dha vera level	0
Indha madhiri role Sivakarthikeyan ku set agadhu.village subject only suits for him.vera actor pani irundha mass aa irukum	1

Figure 6.1: Collected data corpus

500 youtube comments were collected and labelled by five different users. Then the comments are labelled by using the maximum possible value as “harmful” or “non-harmful”.This data is used as an External Data to validate the built model.

Google forms including 500 comments are created and sent to different five users. Then users are instructed to label the comments as harmful or non-harmful by reading. After that the comments are labeled with the help of the responses. Google Form used to label the external data is shown in Figure 6.2.

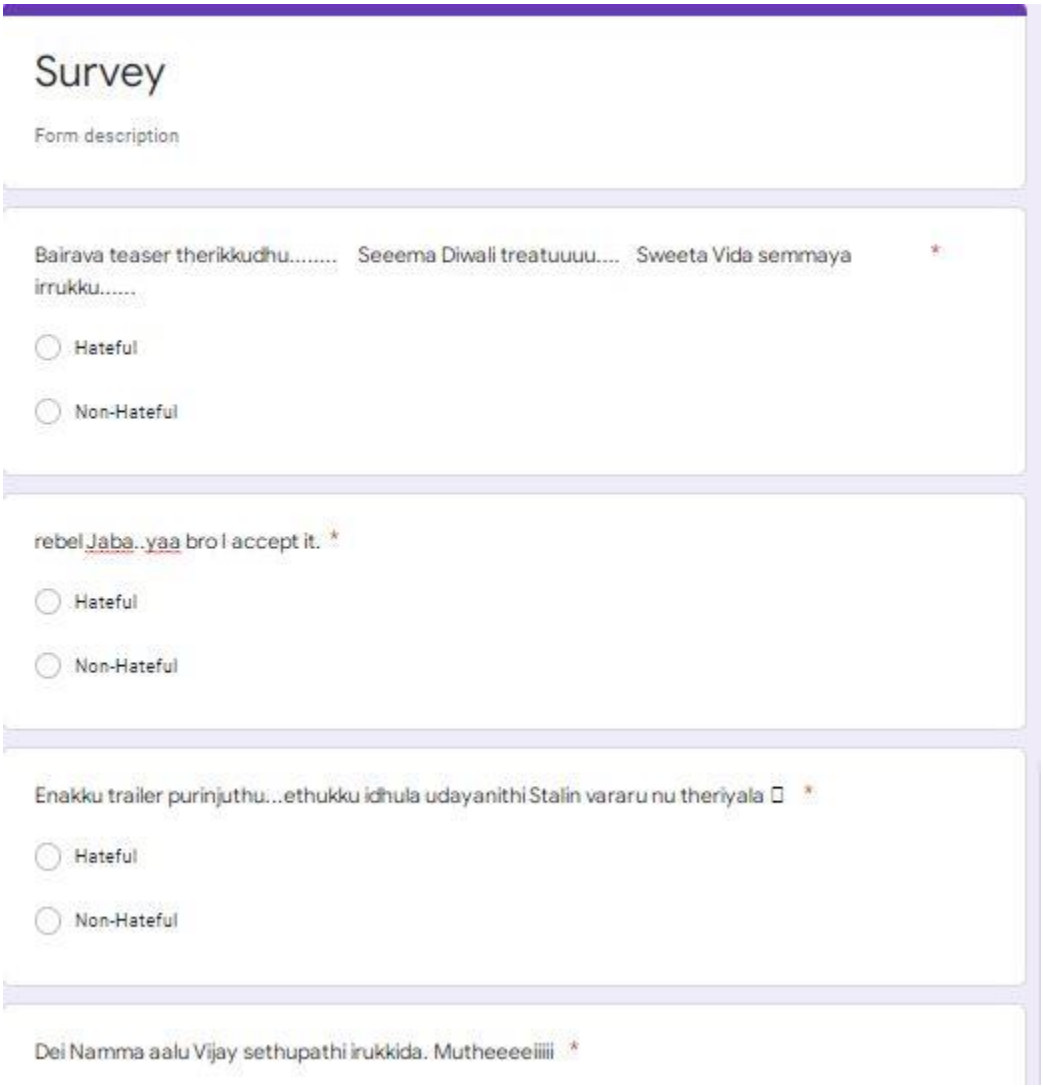


Figure 6.2: Google form for labelling External data

6.3 Text Pre-processing

The collected corpus was first preprocessed before analysis with the following steps:

- **Removal of Duplicates**

  As the first step in the preprocessing, the duplicates are removed from training and corpus. This step removes similar phrases from the corpus. This step is utilized to get an efficient model and to increase efficiency and accuracy.

- **Transform cases (to lower case)**

  An English-Tamil code mixed data set is utilized for the experiment. By using this step, all the sentences are transformed into lower cases.

- **Tokenization**

  By using this operator, paragraphs are tokenized into sentences and then sentences are tokenized into words. In the tokenizing process, some characters like punctuations will be discarded.

6.4 Feature Extraction

After the preprocessing, the feature extraction was done. Feature extraction is a vital step in machine learning-based classification. The determination of features to create vector space is the complex part. According to the literature relevant to English-Tamil code mixed language based text classification, For this study feature extraction techniques such as character n-gram and word n-gram was utilized.

Word n-gram features:

N-grams are continuous sequences of symbols or words or tokens in a record. In specialized terms, it can characterise as the neighbouring sequences of things in a record. It becomes possibly the most important factor when we manage text data in NLP(Natural Language Processing) tasks. N-gram features can be either a unigram, bigram, trigram, and so on. According to the recent literature, extracting word n-gram features gives a better way to distinguish harmful comments.

6.5 Feature Vectorization

Turning a collection of text documents into numerical feature vectors is a common process of Feature vectorization. In machine learning based content classification Feature Vectorization is one of the main steps. According to the past literature, the most used feature vectorization method is TF-IDF.

Term Frequency- Inverse Document Frequency (TF-IDF)

It is an important term inversely connected with its recurrence over documents. Information on how much of the time a term appears in a document is given by TF. And the information about the relative uncommonness of a term within the collection of documents provided by IDF. Can get our final TF-IDF value by expanding these two values together.

6.6 Machine Learning based Classification

Text classification is the automatic process of categorizing text into organized groups according to its contents. There are three kinds of classification methods such as supervised, unsupervised, and semi-supervised. In this research, a labeled corpus is used as a dataset. So supervised classification algorithms are used to do the classification part. Machine learning-based models such as SVM, LR, and Naïve Bayes Classifiers are used.

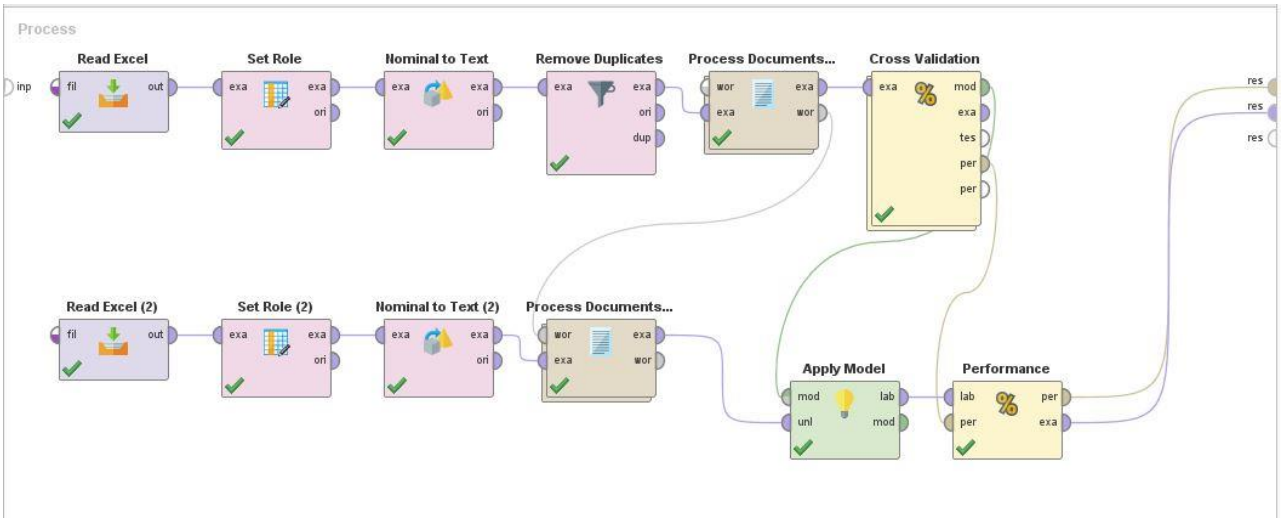


Figure 6.3: Machine Learning-based Classification Model

6.7 Performance Measurement

In the text classification approach performance measurement is the final stage. Many performance measures are available, from which we have chosen the mostly used measures. In common, numerous measures are utilised such as accuracy, recall, precision, F1-score, etc.

"Accuracy", "Recall", "Precision" and "F1-score" were measured to evaluate the performance. The formulas utilised to calculate the measurements are as per the following:

- Accuracy: The accuracy metric defines how many predictions (positive and negative) are really predicted right. To calculate this metric, utilize the following formula:

Accuracy= (TP+TN)/Total number of cases.

- Precision: Precision metric defines how many positive cases are actually predicted right. To calculate this metric, use the following formula:

Precision= TP/ (TP+FP)

- Recall: The recall or True Positive Rate (TPR) metric defines how many positive cases that the model predicted are really predicted right. To calculate this metric, utilize the following formula:

Recall= TP/ (TP+FN)

- F1-score: A metric combines Precision and Recall. To calculate this metric, use the following formula:

F1 score= 2TP/(2TP+FP+FN)

Where,

True positive (TP) =the number of harmful comments that the model predicted as harmful comments.

False positive (FP) = the number of harmful comments that the model predicted but actually are non-harmful comments.

True negative (TN) = the amount of non-harmful comments that the model predicted correctly.

False negative (FN) = the amount of non-harmful comments that the model predicted but actually are harmful comments.

accuracy: 66.24%

	true 0	true 1	class precision
pred. 0	364	195	65.12%
pred. 1	147	307	67.62%
class recall	71.23%	61.16%	

Figure 6.4 Performance Measurements

Then the best fit model was identified. After that, the identified model is validated using an external dataset.

6.8 Summary

This chapter presents an overall implementation of the model proposed in the research work. Furthermore, it also gives a description of the RapidMiner Studio software used to develop the model and how to use classification techniques to build the model. All the models developed in the work will be evaluated and identify the best fit model in the following chapter.

CHAPTER 07

Evaluation

7.1 Introduction

This Chapter focuses on how testing methodologies are carried out according to the objective in terms of evaluation measurements for the chosen text mining procedure such as accuracy rate, TP & FP Rate, Precision, and F1-Measure.

7.2 Evaluation for classification

As described earlier, precision, F1-score, and recall were determined for each model's accuracy. Cross-validation techniques with ten folds were used. The collected data corpus was trained and the following experimental results were obtained with the help of Rapid Miner tools.

Table 7.1: Results with LR Classification Technique

LR	Class	Vectorization Method – TF-IDF				
		Feature Extraction				
		Word n-gram features			Character n-gram features	
		Unigram	Bigram	Trigram	Unigram	Bigram
Precision	Harmful	73.01	82.64	85.64	34.62	60.63
	Non-Harmful	63.21	61.03	59.20	49.63	53.60
Recall	Harmful	51.92	40.40	33.74	3.64	27.07
	Non-Harmful	81.15	91.67	94.44	93.25	82.74
F1-Score	Harmful	51.92	40.4	33.74	3.64	27.07
	Non-Harmful	81.15	91.67	94.44	93.25	82.74
Accuracy		**66.67**	66.27	64.36	48.85	55.16

Table 7.2: Results with SVM Classification Technique

SVM	Class	Vectorization Method – TF-IDF					
		Feature Extraction					
		Word n-gram features			Character n-gram features		
		Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
Precision	Harmful	65.20	78.55	83.47	36.52	68.18	73.08
	Non-Harmful	73.18	63.07	61.65	48.76	51.77	53.18
Recall	Harmful	56.17	48.08	41.82	8.48	9.09	15.35
	Non-Harmful	79.56	87.10	91.87	85.52	95.83	94.44
F1-Score	Harmful	60.35	59.65	55.72	13.76	16.04	25.37
	Non-Harmful	76.24	73.16	73.79	62.11	67.22	68.04
Accuracy		**68.27**	67.77	67.07	47.35	52.85	55.26

Table 7.3: Results with Naïve Bayes Classification Technique

Naïve Bayes	Class	Vectorization Method – TF-IDF					
		Feature Extraction					
		Word n-gram features			Character n-gram features		
		Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
Precision	Harmful	67.18	67.18	67.56	49.59	59.56	68.64
	Non-Harmful	64.96	64.96	65.04	51.85	51.56	53.01
Recall	Harmful	61.21	61.21	61.06	97.37	12.12	16.36
	Non-Harmful	70.63	70.63	71.23	2.78	91.87	92.66
F1-Score	Harmful	64.06	64.06	64.15	65.71	20.14	26.42
	Non-Harmful	67.68	67.68	67.99	5.28	66.05	67.44
Accuracy		65.97	65.97	**66.17**	49.65	52.35	54.85

According to the results obtained above SVM gives higher accuracy than the other models with the uni-gram feature representations. So we can say SVM is the best fit model for identifying harmful content for this particular dataset. To keep away from the overfitting of the data we have to validate the best model using an external dataset. Results are shown in Table 7.4

Table 7.4: Comparision between the test data results and external data desults

Perfomance Measures	**Class**	**Results abtained using test data**	**Results obtained using external data**
Precision	Harmful	65.20	48.30
	Non-Harmful	73.18	82.67
Recall	Harmful	56.17	53.79
	Non-Harmful	79.56	79.29
F1-Score	Harmful	60.35	50.90
	Non-Harmful	76.24	80.84
Accuracy	**Accuracy**	**68.27**	**72.55**

The model works for the external data. So This is the best fit model for identifying harmful comments from youtube data. The best fit model information is listed in Table7.5.

Table 7.5: Details of the best fit model

Preprocessed Techniques	**Removal of duplicates** **Transform cases(to lower case)** **Tokenization**
Feature Extraction	Word n-gram
Feature Vectorization	TF-IDF
Used machine learning algorithm	SVM

7.3 Summary

This chapter concludes with the performance values obtained by the classification models described in the implementation chapter. The last chapter will summarize the overall research by highlighting the findings and discussing the limitation and future work..

CHAPTER 08

Conclusion and Future Work

8.1 Introduction

This section presents an overview of our research and how we provide the solution for identifying harmful content from social media content which belongs to the Natural Language Processing category. And also, this chapter focuses on the limitations and future further work of this research.

8.2 Conclusion

This research work presents a solution for harmful comment Identification in code mixed Tamil-English comments which are collected from YouTube comments.

Our primary objective of this work is to find the best fit model that gives the higher performance value. So the Machine learning models are trained on data of code-mixed language to obtain better performance in all of the models.

We compared the results of three classifiers and observed that SVM outperformed all other classifiers tested here. The highest accuracy of 68.27% is obtained while using TF-IDF feature representation with SVM for uni-gram feature groups.

We found that SVM with uni-gram feature representation gives the highest performance, so this is the best fit model among the model developed. To avoid the overfitting of the data, This model is validated using an external dataset. 500 comments are collected from youtube and labeled by different users. Then the collected labeled data is validated using the developed model. The accuracy is obtained as

Another finding is word n-gram has a higher performance than the character n-gram.

8.3 Limitations

The accuracy rate we got is considerably little low, which means the incorrectly predicated rate is high. The accuracy of the tool highly depends on the dataset we selected.

Comments are code-mixed and it is more complex to identify the harmful contents than the other datasets. In code-mixed English- Tamil language same sentence can be written with different spelling by different persons.

8.4 Future Developments

In future research, we will further consider the different preprocessing methods and different classification methods to further improve the performance of our model.

Research can be done using deep learning and transfer learning methods and then can compare the results. In this research we used 3000 comments as a dataset, we can test the performance by increasing the dataset size.

8.5 Summary

This chapter concluded with the conclusion of the research work, limitations, and extended future work.

CHAPTER 09

REFERENCES

[1] S. Saumya, A. Kumar, and J. P. Singh, "Offensive language identification in Dravidian code mixed social media text," in *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, 2021, pp. 36–45.

[2] K. Kedia and A. Nandy, "indicnlp@kgp at DravidianLangTech-EACL2021: Offensive Language Identification in Dravidian Languages," arXiv [cs.CL], 2021.

[3] Dias, Dulan & Welikala, Madhushi & Dias, N. G. J... (2018). Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning. 1-6. 10.1109/ICTER.2018.8615492.

[4] P. V. Veena, P. Ramanan, and D. G. Remmiya, "CENMates@HASOC-Dravidian-CodeMix-FIRE2020: Offensive language identification on code-mixed social media comments," Ceur-ws.org. [Online]. Available: http://ceur-ws.org/Vol-2826/T2-37.pdf. [Accessed: 13-Jul-2021].

[5] S. Sai and Y. Sharma, "Towards offensive language identification for Dravidian Languages," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 18–27.

[6] K. Yasaswini, K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian Languages," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 187–194.

[7] B. R. Chakravarthi et al., "Dravidian code mix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text," Arxiv.org. [Online]. Available: http://arxiv.org/abs/2106.09460v1. [Accessed: 13-Jul-2021].

[8] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of Hindi-English code-mixed social media text for hate speech detection," in Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, 2018, pp. 36–41.

[9] N. S. Samghabadi, D. Mave, S. Kar, and T. Solorio, "RiTUAL-UH at TRAC 2018 Shared Task: Aggression Identification," arXiv [cs.CL], pp. 12–18, 2018.

[10] Al-Hassan, Areej, and Hmood Al-Dossari.(2019)“Detection of hate speech in social networks: a survey on multilingual corpus. “Computer Science Information Technology (CSIT)9(2):8.

[11] Blaya, Catherine. (2019) “Cyber hate: A review and content analysis of intervention strategies. “Aggression and violent behavior45:163-172.

[12] Biere, Shanita, Sandjai Bhulai, and Master Business Analytics. (2018) “Hate Speech Detection Using Natural Language Processing Techniques”. Diss. Master’s dissertation, Vrije Universiteit Amsterdam.

[13] A. Bohra, D. Vijay, V. Singh, S. Akhtar and M. Shrivastava, "A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection", 2022.

[14] I. Kwok and Y. Wang, “Locate the Hate: Detecting Tweets against Blacks,” Twenty-Seventh AAAI Conf. Artif. Intell., pp. 1621–1622, 2013.

[15] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” Policy and Internet, vol. 7, no. 2, pp. 223–242, 2015.

[16] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In WWW, pages 29{30, 2015.

[17] Zeerak Waseem and Dirk Hovy. Harmful symbols or harmful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop, pages 88–93. Association for Computational Linguistics, 2016. doi:10.18653/v1/N16-2013

[18] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web—WWW ’16, Montreal, QC, Canada, 11–15 May 2016; ACM Press: Montreal, QC, Canada, 2016; pp. 145–153.

[19] Malmasi, S. and M. Zampieri, Detecting hate speech in social media. arXiv preprint arXiv:1712.06427, 2017.

[20] Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. , "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," ACM Transactions on Interactive Intelligent Systems (TiiS) 2 (3): Article 18., 2012.

[21] Liu, S. and T. Forss. Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification. in KDIR. 2014.

[22] Gitari, N.D., et al., A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 2015. 10(4): p. 215-230.

[23] S. Hinduja and J. Patchin, "Bullying, Cyberbullying, and Suicide", Archives of Suicide Research, vol. 14, no. 3, pp. 206-221, 2010.

[24] Malmasi, S. and M. Zampieri, Detecting hate speech in social media. arXiv preprint arXiv:1712.06427, 2017.

[25] "YouTube hate speech and harassment policy - How YouTube Works", *YouTube hate speech and harassment policy - How YouTube Works*, 2022. [Online]. Available: https://www.youtube.com/howyoutubeworks/our-commitments/standing-up-to-hate/. [Accessed: 20- Jun- 2022].

[26] L. Ketsbaia and X. Chen, "Detection of Hate Tweets using Machine Learning and Deep Learning." Accessed: Jun. 21, 2022. [Online]. Available: https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/44126850/Accepted_Manuscript_Ketsbaia_et_al.pdf

[27] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic Hate Speech Detection using Machine Learning: A Comparative Study," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, 2020, doi: 10.14569/ijacsa.2020.0110861.

[28] Jovanovic, Milos & Vukicevic, Milan & Delibašić, Boris & Suknovic, Milija. (2014). Using RapidMiner for Research: Experimental Evaluation of Learners.