HATE SPEECH DETECTION FROM SOCIAL MEDIA IN MALAYSIAN CONTEXT

YEE QING WEI

XIAMEN UNIVERSITY MALAYSIA

2024

XIAMEN UNIVERSITY MALAYSIA
厦門大學馬來西亞分校

FINAL YEAR PROJECT REPORT

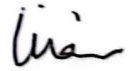
HATE SPEECH DETECTION FROM SOCIAL MEDIA IN MALAYSIAN CONTEXT

NAME OF STUDENT	:	YEE QING WEI
STUDENT ID	:	AIT2104019
SCHOOL/ FACULTY	:	SCHOOL OF ELECTRICAL ENGINEERING AND ARTIFICIAL INTELLIGENCE
PROGRAMME	:	BACHRELOR OF ENGINEERING IN ARTIFICIAL INTELLIGENCE (HONOURS)
INTAKE	:	2021/04
SUPERVISOR	:	DR SHAIDAH JUSOH ASSISTANT PROFESSOR

JULY 2024

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at Xiamen University Malaysia or other institutions.

Signature : 

Name : Yee Qing Wei

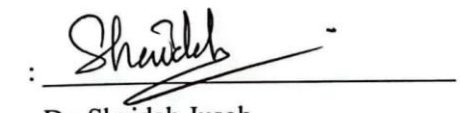
ID No. : 020805-05-0397

Date : 1/7/2024

APPROVAL FOR SUBMISSION

I certify that this project report entitled "**HATE SPEECH DETECTION FROM SOCIAL MEDIA IN MALAYSIAN CONTEXT**" that was prepared by YEE QING WEI has met the required standard for submission in partial fulfillment of the requirements for the award of Bachelor of Engineering in Artificial Intelligence (Honours) at Xiamen University Malaysia.

Approved by,

Signature : 

Supervisor : Dr. Shaidah Jusoh

Date : 10 July 2024

ACKNOWLEDGEMENTS

I would like to thank all who have contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Prof. Dr. Shaidah Jusoh for her invaluable advice, guidance and her enormous patience throughout the development of the research. I would also like to thank all of my correspondents, who had filled up the form I had created so I can gather the ideas on hate speech is defined in Malaysia

In addition, I would also like to express my gratitude to my loving parents and friends who have helped and given me encouragement.

ABSTRACT

Malaysia never has a dataset that is made to train detection system to combat against hate speech against Malaysians in social. This study has created a new hate speech dataset for a context of Malaysians. This dataset was extracted from TheThe novelty of the study is to create a brand new hate speech dataset in Malaysian context by extracting social media posts in Facebook, Twitter and Lowyat.net. Each of these posts is then tokenized into list of word token, before the preprocessing system had removed non-alphabetical letters, negation and stop words. After creating and preprocessing the dataset, the dataset has been converted into word embeddings. Each of the embeddings model used in this study provides different results. LDA , Skip-Gram and CBOW model convert word tokens into LDA , Skip-Gram and CBOW embeddings respectively, while BERTopic model convert word tokens into BERTopic representations, topic possibilities and representative documents. Meanwhile, fastText embedding model could directly train hate speech detection model. Machine and deep learning algorithms are then trained using these word embeddings, and also the original pre-processed word tokens. Naïve Bayes machine learning algorithm can produce hate speech detection system with accuracy up to 60% when trained with CBOW embeddings. Otherwise, Naïve Bayes provides poor accuracy between 45 to 55%, depending on what type of test data is used. LSTM and GRU however, when trained with the original pre-processed data text has high accuracy of 80% and 85% respectively. DCNN and AlexNet can only achieve around 50%, regardless what kind of test data is being used to train. LSTM and GRU only achieve around 55% respectively, when trained with other word embeddings. Results show that pre-processed word tokens either perform better than or as same as word embeddings. Results also state that deep learning algorithm performs better than machine learning algorithms provided that it is trained with pre-processed word tokens.

KEYWORD: Hate Speech Detection System, Social Media Post, Malaysian, Deep Learning, Embedding

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

x	input in the current GRU layer
h	information from the previous GRU layers
W	weights of the input embeddings
U	weights of the previous information in previous embeddings
σ	sigmoid function
z	update gate value
r	reset gate value
$\odot$	element-wise multiplication

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, a significant high amount of Malaysia netizens have used hate speech while posting in social media platforms. According to Teoh (2023), hate speech has been surging in Malaysia social media ever since the November election campaign cycle. Malaysia authorities have also taken down 3752 posts related to fake news or hate speech from January 1 to October 31 2023 (Choy, 2023).

The exposure of hate speech against a group not only normalize discrimination towards them, but also encourage others to segregate, harass and even commit violence against the discriminated group (Bilewicz & Soral, 2020). Therefore, it is important for social media companies to remove any hate speech in their platform. An automatic hate speech detection system should be implemented to effectively remove hate speech that discriminates races, religion and genders among Malaysians.

Even though social media company has started to use AI moderation systems to detect and remove hate speech from their platforms, they rely on hate speech datasets in Western countries' context (Vincent, 2020; Paul & Dang, 2022) and are unable to detect most of the hate speech posted by Malaysians. Even if their hate speech datasets include slurs against every community in Malaysia, they are still unable to detect social media posts that seems innocent by Westerners but are actually hate speech in Malaysian context. Therefore, it is necessary for social media company to apply hate speech datasets in Malaysian context to detect hateful Malaysian posts.

When searching for literatures in internet for literature review, zero literature paper has discussed about building a hate speech detection system to specifically moderate Malaysians, let alone finding a public hate speech dataset in Malaysian context. Therefore, a dataset of hate speech in Malaysian context needs to be built first, to train a hate speech detection system. If hate speech detection system can effectively classify the instances in this dataset as hate speech or non-hate speech, the algorithm can effectively detect and remove any hateful social media post made by Malaysians.

1.2 Problem Statement

It is true that there is a similarity between hate speech in Western world and Malaysia. This also means that it is possible to train the detection system only with datasets found from online. However, the overall accuracy of the hate speech detection system might be decreased due to the system's bias of detecting hate speech in a Western context. Basically, this detection system is unable to recognise specific Malaysian contexts in sentences because it has been trained with data made by Western data engineers. Detection system would never understand the hidden meaning of terms such as cow, "Type-C" and "kering" if it has never seen them in the training data model. Therefore, it is necessary for this thesis study to use social media posts in Malaysian context to train the hate speech detection system. However as previously mentioned, there is no available hate speech dataset in Malaysian context in the internet. Therefore, this thesis study will propose a hate speech data model. The data in the model is obtained by extracting suitable Malaysian social media posts and classifying them. Multiple machine and learning algorithms are then trained using the hate speech dataset to create a hate speech detection system.

1.3 Aims and Objectives

The aim for this study is to detect hate speech from social media among Malaysians. Therefore, two objectives have been proposed to complete this aim. They are:

1. build a hate speech corpus in Malaysian context.
2. propose a suitable model to detect hateful social media posts among Malaysian community.

In order to complete the first objective, raw texts have to be collected from Malaysian social media posts and ground rules have to be created. Each social media post shall be classified to be hate speech or non-hate speech. The ground rules are ensured to be aligned with the consensus of Malaysians, while matching the definition of hate speech. For the second objective, machine and deep learning models are implemented and trained with different types of word embeddings. The accuracies of these models are then compared to find out the best model. Finally, the best model is used to build a prototype with user interface for user to enter any input. The prototype will then determine and inform the user hate speech in real time. Experimentations are then performed on the prototype to ensure that it has a decent accuracy in detect hate speech.

1.4 Thesis Contribution

This dataset is a unique dataset that contains raw posts posted by Malaysians in social media. It contains many hate speech with new localized contexts and meanings. If an algorithm can successfully classify Malaysian social posts, the algorithm can contribute to the reduction of Malaysian hate speech spreading around the social media. This is because social media with automatic detection system would be easily removing any hateful social media posts before it gets spread to too many users.

1.5 Thesis Overview

Before actually deciding on what kind of research methodologies this thesis study shall have performed, a literature review of the topic of hate speech and its detection system

shall be performed in this study. The literature review first focuses on finding definitions of hate speech and categorize different classes of hate speech. Then, the literature review will find similarities and difference of hate speech in different countries, including United States of America and Malaysia. Finally, literature review will look at different hate speech models that have built in the past. Even though these hate speech detection models are trained with data in Western context, it is still necessary to find out their performance and the advantages and disadvantages of using each of them.

After that, the thesis then shows the methodologies used in the research. In the beginning, numerous posts made by Malaysia netizens have been collected from social media websites. A form have also been created to determine the Malaysians' boundaries for a speech to be hateful or not. After that, the posts can be categorized as hate speech or non-hate speech based on the aforementioned boundaries. After a data model has been built, its text data will be converted into word embeddings such as BERTopic and Continuous Bag of Words (CBOW). The text data, whether it has been converted into word embeddings or not, has been trained in multiple machine learning and deep learning models. After that, a User Interface (UI) is created for users to input any sentences he wants to know whether it is hate speech or not.

Multiple interesting results have been observed inside the hate speech models. For example, every models with BERTopic embeddings did not perform well in classifying Malaysian social media post as hate speech or non-hate speech. RNN and GRU algorithm also performed significantly better than DCNN algorithms. The accuracy of fastText model is also very sensitive to its learning rate.

CHAPTER 2

LITERATURE REVIEW

2.1 Definition and Categories of Hate Speech

Before reviewing various hate speech detection system, the definition of hate speech needs to be clarified first. In Fortuna & Nunes (2019), multiple definitions of hate speech made by different authorities has been discussed. For example, European Union has released a code of conduct which specifies that any action publicly inciting to violence and/or hatred against one or more members of a group with same race, religion, nationality or ethnicity (Jourová et al., 2017). Facebook has stated that any content that attacks people based on their race, national identity, gender, sexual orientation, disease or disability can be considered as hate speech. Facebook has also specified that, humour and satire content that might be deemed as offensive by some people are not hate speech (Fortuna & Nunes, 2019).

Meanwhile, Twitter combined the previous two definition and stated that someone who either inciting violence/hate or attacking others, is performing hate speech (Fortuna & Nunes, 2019). Fortuna & Nunes (2019) then concludes their definition that is similar to Twitter. To them, hate speech is language that attacks and incite violences against specific groups due to their religion, descent, nationality, physical appearance, sexual preference or gender. They also believe that hate speech can still occur in subtle forms or humours. This is due to this kind of “harmless” content can negatively affect someone’s mental health and promote incorrect stereotypes.

People normally determine whether a sentence can be determined as hate speech simply based on the words used inside. One example is when someone is using slurs or referencing hateful organizations and ideologies (Fortuna & Nunes, 2019).. They did this to promote their hateful ideas and/or harass the targeted group. This kind of action usually will be deemed as hate speech. However, the usage of slurs, hateful ideologies and organizations in a neutral or historical context are not considered as hate speech. For example, the term “nigger” is a racial slur. Yet, a sentence containing the term is not considered as hate speech if it discusses the languages used to discriminate Africans back in 19th century. Furthermore, there exist certain words that can only become hate speech when they are used in certain context (Fortuna & Nunes, 2019). For example, the phrase “black monkeys” doesn’t attack on any group of people in normal context. However, this phrase will have a hateful meaning to Africans if the phrase is referring to them instead of actual monkeys.

The same can be said on the context of multiple sentences. Take this example “You are a Malay. I don’t want to see you.”, if we separate these sentences in the text passage into different independent social media post, none of them would have deemed as hate speech by Malaysians. However, if these sentences are combined into one passage, most Malaysians would agree that it will be standard definition of hate speech. These exceptions are huge challenges for artificial intelligence to determine whether a sentence is hate speech.

Hate speech can be divided into multiple categories. Each category discriminates against people in a certain group within the category. Table 2.1 summarises the hate speech category and the examples inside this category (Aditya et al., 2022; Fortuna & Nunes, 2019)

Table 2.1: Categories of Hate Speech

Category	Definition	Example
Race	A group of people with common cultural, geographic, linguistic, ancestral and/or religious background (Merriam-Webster, 2023a)	Malay Chinese Indian
Religion	A system of belief and worship of a god or gods (Cambridge University Press, 2023)	Islam Buddhism Christianity Hinduism
Gender	The behavioural, cultural and psychological traits associated with men or women (Merriam-Webster, 2023b)	Male Female
Body Physics	Bodily proportions, muscular developments or appearances of people (Houghton Mifflin Harcourt, 2011)	Fat Skinny Short
Sexual Orientation	Emotional, sexual or romantical attraction of an individual (Eldrige, 2023)	Straight Gay Lesbian
Disability	A condition of the body or mind impairment (Centers for Disease Control and Prevention, 2020)	Autism Down's Syndrome Amputation

In Bahador (2020), hate speech has been categorised into three categories. which are "early warning", "dehumanization & demonization" and "violence & incitement". The first category is subtle as it first only creates and distinguish between an "us" group and a "them" group. This leads to the emerge of criticism of "them" group's negative actions. Eventually, this group will start to have negative characterizations.

It is necessary to understand the meaning of both terms in the second category first to understand the category itself. Dehumanization is the portrayal of "them" group as non-human entities that not well liked by the public, such as rats and germs. It attempts to normalize abusive and hateful behaviours towards the dehumanized group as they are now considered not as human beings (Bahador, 2020). People are told to treat them as animals instead of humans. Demonization is the portrayal of "them" group as evil entities, such as demons and monsters, that will destroy the "us" group and bring end to the world. It justifies starting a conflict and provoke violence towards the demonized group (Bahador, 2020). This is because people think that it is justified to commit any immoral action towards an evil figure to survive.

The last hate speech category is straight up calling violence against "them" group. This is the only hate speech category which violates the law (Bahador, 2020). While various institutions have clearly defined hate speech and its category, the context of hate speech in each category are still different across countries due to different cultures, religions and moral views.

2.2 Hate Speech in Different Countries

2.2.1 United States of America

United States of America is an ethnically diverse country. According to U.S. 2022 census, 58.9% of Americans are from the White group. Latinos and Blacks made up of 19.1% and 12.6% of population respectively (USAFacts, n.d.). Due to their large populations, racial conflicts in United States mainly involves with at least one of these three groups.

The conflict between White and African Americans started to emerge when Africans were transported to United States as slaves. The African slaves were sold in auction house to slave owners to work and treated poorly in plantation farms. The authorities even helped slave owners who wish to torture slaves but couldn't do it by themselves (Hinton & Cook, 2021). After these Africans were liberated in Civil War, the tensions between former slave owners and slaves had been increased. This led to Black Americans being discriminated and segregated from the rest of the community. Many convicted African Americans were forced to do labours inside the prison day and night. These exhausting labours caused many Blacks to die in the prison (Hinton & Cook, 2021). This situation was only improved after the enactment of 1965 Civil Rights Act. Radical Black Nationalists organisations such as Black Liberation Army, had tried to combat against this situation by performing terrorism acts such as hijacking an airplane (Susman, 2011) and robbing an armoured truck (Gado, n.d.).

The history of conflict between White and Latino people are rather short. It emerged due to large number of refugees coming from Central America. These refugees came to United States to seek for new economic opportunities and escape from violence in their home countries (Morales et al., 2022). However, many blue-collar workers were fired from their jobs because they couldn't compete with the refugees who only want insanely low wages (Galemba, 2021). Furthermore, these refugees can't speak English well and are unable to join into the local area communities. These factors caused conflict between these three races. About 30-40% of people from each of these three races have used racial slurs and other negative sentiments against each other in social media (Ngwainmbi, 2022). For example, many Blacks often refer White people as "white trash" and "cracker".

There is also large presence of islamophobia in United States. This distrust of Muslim has been largely increased after the Islamic terrorist attacks that happened on 11th September 2001 (Haque et al., 2019). Many Americans are scared of their Muslim friends and co-workers and worried if they are terrorists in disguise. American Muslims often feel depressed when they witness the increase of anti-Islam hate speech in social media. For example, Muslims were told to go back to Middle East even if their ancestry are not from there. Some of them are even forced to change their appearances and stop defending Islam in order to not look like an extremist in the eyes of others (Haque et al., 2019). Christians in America also often get verbally attacked, harassed and sometimes physically assaulted by secularists and atheists. A report from Observatory on Intolerance and Discrimination against Christians (OIPAC) stated that there are about 1500 anti-Christian attacks in United States of America and Europe in the years 2020 and 2021 (Duvall, 2022).

Issue involving genders and sexual orientations are among one of the hotly debated topics in United States. Some Americans believed that their birth assigned gender isn't their "actual gender" (transgender) or they have no gender (non-binary) (Price et al., 2023; HRC Foundation, 2023). Meanwhile, some Americans are attracted to people same gender as them (gay and lesbian), to male and female at the same time (bisexual), or to no one at all (asexual) (Zambon & Kuehnle, 2023). This caused heated debates between people who support and oppose normalizing these genders and sexual orientations. This naturally produced a lot of hate speech against people with different genders and sexual orientations. One thing to note is that United States' mainstream narrative is to accept whatever gender and sexual orientation people believe. Thus, it is considered as a hate speech in United States of America to relate sexualities and genders as mental illnesses and advise someone to seek for mental help. An individual will be deemed hateful by majority of Americans for referring a transgender with his birth assigned gender.

2.2.2 Malaysia

The racial relationship in Malaysia is only slightly better than in United States. The differences between the two countries are that the races involved in the argument are different, and Malaysian needs to face consequences in court for any statement he said online. The three main races in Malaysia are Bumiputras, Chinese and Indians. They respectively represent 69.6%, 22.6% and 6.8% of the population (Mahidin, 2020). Furthermore, Bumiputra is a large category consists of several races, such as Malays, Ibans and Dusuns.

One of the most common hate speech in Malaysia is about Malay supremacy ("Ketuanan Melayu"). Politicians often use the fear of Malays losing their identity and economic status to promote hatred and discrimination against other races (Sani, 2011). Sometimes, different races in Malaysia will temporarily have a truce and instead focus picking on refugees, immigrants or just anyone from a different country (Fernandez, 2020).

The ratio of a hate speech containing racial undertone to a hate speech containing religious undertone is 4:1 in Malaysia (Zamri et al., 2023). However, it is still easy to find Malaysians spreading anti-Islam sentiment in social media. Furthermore, there are large presences of anti-Christian, anti-Buddhism and anti-Hinduism rhetoric on the internet. Meanwhile, there's not much hate speech regarding someone's sexual orientation in Malaysia. Malaysians also only spread hate speech degrading either men or women in social media. There's simply no presence of hate speech against someone for being transgender or nonbinary in Malaysia. Unlike United States, a Malaysian can be deemed as a rational and a morale person if he advises someone to seek for mental help to correct its sexual orientation

Hate speech in Malaysia appears the most when there's instability in the country. In Fernandez (2020), three waves of conflict occurred in the social media during COVID-19 pandemic. The first wave caused a vendetta against tourists and immigrants from Mainland China. This is because COVID-19 virus originated from Wuhan, China. Surprisingly, Malaysia Chinese didn't get attacked in this wave. The second stage of conflict was started from a Taligbh meeting that started a virus outbreak and caused a spike of national COVID-19 cases. Muslims and non-Muslims debated over the necessity of Taligbh meetings and eventually attacked each other. Many Malaysians targeted Rohingyans in the third wave. It was started by a rumour about Malaysian government planning to give citizenship to Rohingyans, and misconceptions that Rohingyans brought COVID-19 virus from Myanmar.

In conclusion, hate speech in Malaysia is more focused on birth assigned gender while hate speech in United States of America is more focused on sexual orientation and transgender. Toxic statements about someone's race, religion or nationality still can be commonly found in the social media communities of both countries

2.3 Hate Speech Detection Systems

Several research articles on hate speech detection systems were reviewed and compared to figure out which algorithms are more suitable to train a hate speech detection system. The author generally checks the dataset used and the results in each of the literature papers.

For example, in Roy et al. (2020), hate speech dataset was obtained from Agarwal (2018) and then transformed into TF-IDF embeddings. Multiple machine learning models were trained in the literature, which were RF, NB, SVM, DT, GB, and KNN. Furthermore, deep learning models such as CNN, LSTM, C-LSTM and DCNN were trained as well. It showed that SVM had the best results (0.53) among the machine learning algorithms. However, every single deep learning model performed better than SVM. While CNN, LSTM and C-LSTM had similar accuracy score, detection system trained with 10-fold-cross validation DCNN performs significantly better than the rest of deep learning algorithms. This DCNN model predicted 88% of hate speech and 99% of non-hate speech correctly.

In another example, Al-Hassan & Al-Dossari (2021) studies Arabic posts that were posted on Twitter. The datasets were then transferred into TD-IDF and Keras embeddings respectively, for base SVM and deep learning models. Four deep learning models were developed and trained in this study, which are LSTM, LSTM + CNN, GRU and GRU + CNN. They believes that combining two deep learning algorithms in one will show a better performance as they can capture more contexts. They found out that the base SVM model had the worst accuracy in predicting hate speech. While

LSTM trained slower than GRU model, it can result in higher accuracy. They believe it is due to LTSM classifying instances with multi classes better than GRU. Furthermore, adding CNN to both models increased the percentage of Twitter posts correctly labelled as hate speech or non-hate speech. Thus, LSTM + CNN model gave the best result in this literature.

Muslim et al. (2021) is also one of the literature that have been reviewed in this thesis study. It transforms OLID dataset into a BERT embedding. This BERT embedding is then fine-tuned into two additional word embeddings with sensitive learning and ensemble models respectively. These three embeddings were then trained to create a hate speech detection system. The result of this literature paper is that both fine-tuned BERT model had higher accuracy than the original BERT model.

The review on literatures does not end by simply looking at their results. Further observations need to be done on the literature papers to actually figure out which algorithm previously performs the best in the literature papers. One observation found is that most papers agreed that deep learning NLP models generally outperform machine learning models. Jahan & Oussalah (2023) had reviewed multiple review papers of hate speech detection on different languages through the years. It found out that the outperformance of machine learning models has caused researchers to change into them in their detection system. When researchers began to implement hate speech detection system, they would use machine learning algorithms such as SVM (Support Vector Machine) and TF-IDF to train their model. Eventually, researchers decided to use deep learning algorithms such as RNN and CNN to implement model. Jahan & Oussalah (2023) had also found out that there was no significant performance difference when using RNN or CNN model. Al-Hassan & AlDossari (2021) had found out that its deep learning models (LSTM & GRU) has higher performance than its base models (SVM & TF-IDF). Aditya et al. (2022) had also found out that deep learning models were more suitable to be used to solve multi-label classification problems. Roy

et al. (2020) found out that the performance of SVM has the highest accuracy comparing to other machine learning models (LR, RF, KNN & etc.). However, SVM algorithm was still unable to effectively predict whether a tweet could be categorized as hate speech or not. 53% of the tweets in the dataset is correctly classified. Meanwhile, deep learning models in this study can obtain high accuracy score. DCNN with 10-fold cross validation was the best deep learning model observed in the study. It had prediction recall value of 0.88 for hate speech and 0.99 for non-hate speech. These studies had proved that it is better to use deep learning algorithm to train hate speech detection model.

Furthermore, studies showed that combination of several deep learning models to train hate speech detection model generally performs better than only using one. In Al-Hassan & Al-Dossari (2021), researchers had trained LSTM & GRU (Gated Recurrent Units) models and compared their performances on predicting Arabic tweets. LSTM categorised the tweets better than GRU but took longer time to train. This was due to the complexity of LSTM model. These two algorithms were then combined with CNN models. The combination of CNN with these two models performed better than the initial deep learning models. The hate speech detection model of CNN combined with LSTM performed the best in the study. Jahan & Oussalah (2023) also agreed that it is better to combine different deep learning models into one training algorithm. This was because there was no single deep learning model that could perform the best on detecting hate speech. Albadi et al. (2018) also supported this claim as their model with best performance was a GRU-based RNN model. These literatures recommended researchers to implement a model consists of at least two deep learning algorithms to detect hate speech.

Moreover, BERTopic created word embeddings that give the highest accuracy. Jahan & Oussalah (2023) stated that BERTopic topic modelling is the most popular word embedding model as of right now. Its accuracy is higher than ELMO topic modelling

algorithnm, which was widely used by researchers during the early stages of the development of hate speech detection systems. Muslim et al. (2021) had created different word embeddings to train in detection system. The result showed that BERTopic embeddings that were fine-tuned had the result with the best accuracy compared to other topic modelling algorithm, including the base BERTopic embeddings. The fine-tuning process of BERTopic included implementing costsensitive learning (modifying loss function by updating weight during backpropagation) and ensembled models (hard and soft majority voting). These studies have shown that it is highly advised to use BERTopic word embedding models to transform corpus into word embeddings to train hate speech detection model.

However, some research had applied machine learning algorithms and other word embedding models to train detection system. In Poletto et al. (2020), WI (Weirdness Index) and PWI (Polarized Weirdness Index) used to train the detection system. WI measured the ratio of the frequency of a term appears in a specific set of documents over the ration of the frequency of the same term appear over the whole corpus. While it managed to obtain quite high accuracy in terms of projecting hate speech, it was unable to give human researchers an insight on how it labelled and annotated the corpus. Ariwibowo et al, (2022) applied FastText + CBOW, Ngram LSTM and Word2Vec + LSTM algorithms to train detection system. FastText + CBOW model performed better than the other two machine learning models. It could correctly classify 83.52% of Indonesian texts to be hate speech or non-hate speech. It could also determine whether an individual or a group was targeted and what category of hate speech (gender, religion, race, etc.) was used in a hate speech. The research also pointed out that word embeddings like BERTopic might have given a better result and there were planning to build models with this embedding. Aditya et al. (2022) had used TF-IDF word embeddings to train multiple hate speech detection models. One vs Rest model was the best algorithm as it obtained a score of 0.91 on detecting the category of hate speech. It essentially trained 17 classifiers for each target feature’s score and

an additional one to predict which feature did a hate speech belong to. In conclusion, deep learning and BERTopic modelling algorithms are not the only selections to build a hate speech detection system. Some alternative choices can still build a detection system with high accuracy.

2.4 **Summary**

The literatures have clearly shown that there is a difference between hate speech in Malaysia and in Western countries (most notably United States of America). While there are many research on implementing different techniques of hate speech detection system, none of them focuses on hate speech in Malaysian context. The English public datasets available have their data obtained from users living in Western countries. Thus, these datasets has not touched on the races in Malaysia. The amount of religious hate speech in Western hate speech dataset has been underperformed than expectations. This is likely due to the large existence of speech on sexual orientation and transgender in these datasets. These data have taken up spaced inside the data models, that were meant to store religious speech. It is likely that the fact that religion does not cause a big conflict in the West like it does in Malaysia contributes to the underperformance as well. During the process of searching and reviewing literatures, the author also could not find a hate speech dataset written in Malay. Therefore, this thesis study cannot obtain Malaysian social media post by directly translating Malay dataset into English. Therefore, this thesis study needs to obtain Malaysian social media posts and use them to build a data model.

This thesis study had also reviewed multiple machine learning and deep learning models and determined which ones could potentially have the highest accuracy to classify any sentence to be hate speech or not. The algorithms that prove to provide high accuracy in various literatures of hate speech detection system, will be implemented in the upcoming methodologies.

CHAPTER 3

RESEARCH METHODLOGY

3.1 Methodology Overview

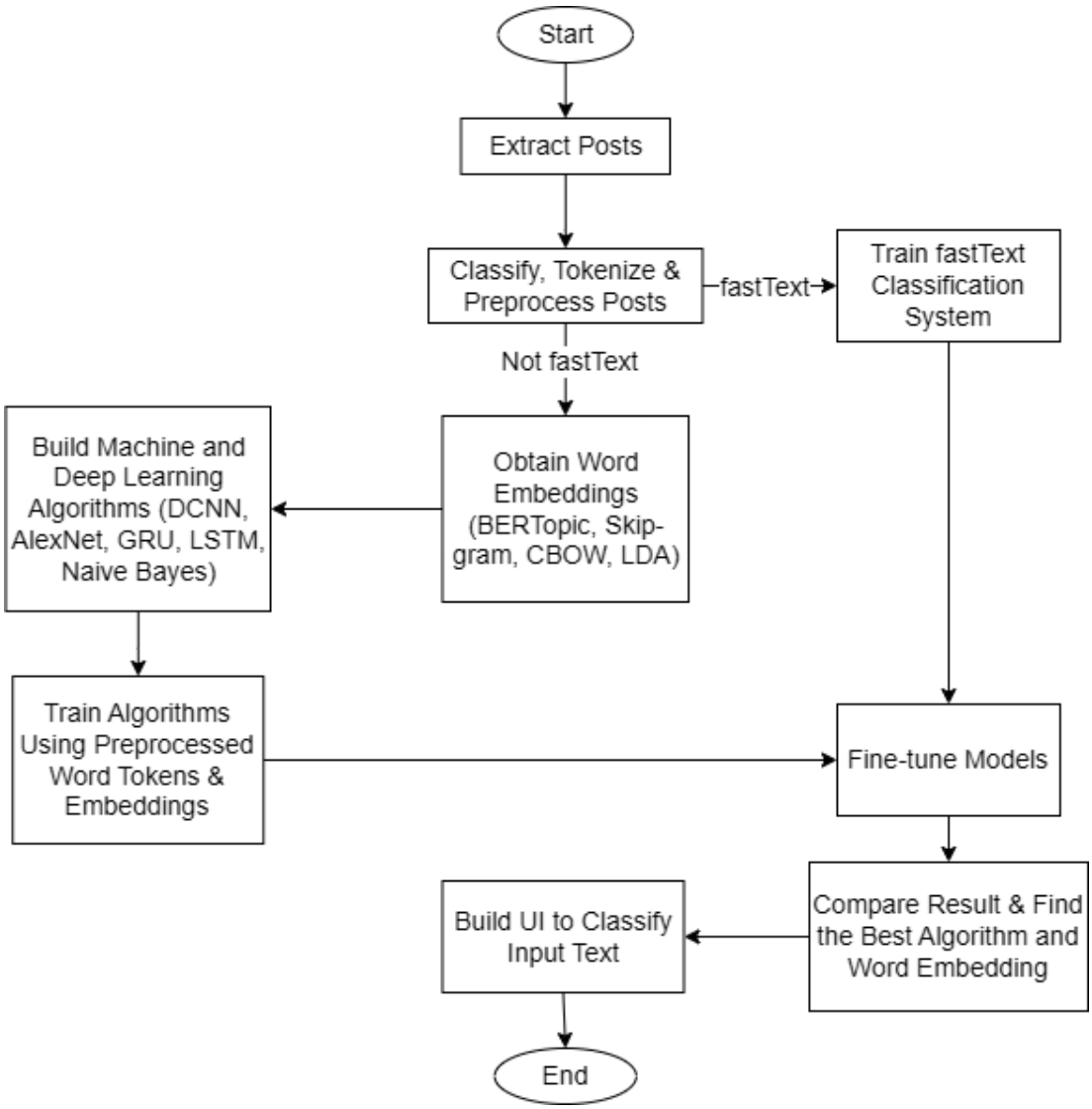


Figure 3.1: The Flow of Creating Hate Speech Detection System

This research has applied the definition of hate speech from Fortuna & Nunes (2019) to classify whether a sentence is spreading hateful ideas or not. The literature reviews have clearly shown that deep learning model performs better than machine learning model on most occasions (Roy et al., 2020; Al-Hassan & Al-Dossari, 2021). This combines with the fact that a detection system trained with more than one deep learning algorithm generally has higher accuracy. Thus, it is necessary to train corpus text with different deep learning models to obtain accuracy as high as possible. Multiple reviews have stated that BERTopic word embeddings performs better than other embeddings (CountVec, FastText) in correctly labelling a sentence to be hate speech or not (Muslim et al., 2021). The review stated that this advantages does not change no matter what kind of machine learning or deep learning algorithm is used in the training phase.

However, there exists literature papers that are in contrary to the points above have been found (Poletto et al., 2020; Ariwibowo, 2022). Even though there is a possibility of them being outliers, this thesis study will not exclude them from the final research methodology. Even though a deep learning detection system with BERTopic word embeddings still performs the best in most research, there is a chance it does not bring great result in this study. Therefore, this study also prepares detection system with alternative machine learning algorithms and word embeddings. As this research novelty is about the creation of a Malaysian hate speech dataset, it is actually not necessary to find or create an algorithm that can absolutely give the highest accuracy. The target is to ensure that detection algorithms can generally provide accuracy of over 70%. If the target succeeds, it is certain the data model has enough hateful and non-hateful posts in Malaysian context, and can be used to train detection system.

In the end, the main focus in the methodology still lies on creating a data model after preprocessing the Malaysian hateful social media posts and training deep learning algorithms with BERTopic embeddings. Other deep learning and machine learning

models have also been trained, to act as potential alternatives of the BERTopic embeddings or control cases. The flow of creating hate speech detection system is shown in Figure 3.1, and the following sections will discuss more details regarding the creation steps.

Validation accuracy and loss are the measurements used to see which model is the overall best model among every single model built in this thesis study. However, manual checking is still needed to see if the data model will go wrong with a specific context. The way to solve it is to build a UI for any user to test the hate speech detection system out. Different posts will be inputted into the UI to observe if the detection system has any bias or errors on any specific community, slang or semantic contexts in social media posts with more than one sentences. These results will determine whether this hate speech dataset in Malaysian context needs to be improved.

3.2 Data Collection

As this research focuses on the novelty of building a completely new dataset model of hate speech in Malaysian context, 1352 posts of Malaysians discussing race, religion or gender have been collected from websites such as Facebook, Twitter and Lowyat.net. For ones that do not know, Lowyat.net is a forum site that is popular among Malaysians to discuss any kind of topics. Nevertheless, most social media posts have been extracted either from discussions about a certain community in Malaysia, or discussions of local political affairs. After that, these data will be added with a binary "HateSpeech" label, with 1 and 0 being hate speech and non-hate speech respectively.

Even though Fortuna & Nunes (2019) has already given clear information on how to determine sentences to be hate speech, the thesis study still need to understand how Malaysians determine a sentence to be hateful. Thus, a survey is created for Malaysians from different race, religion and gender to fill it up. This survey first provide participants 21 sentences and its related context. Then, the survey will then ask participant to classify each sentence to be hate speech or non-hate speech. From the participants' responses, the thesis study can set up a few ground rules to determine any sentence being hateful or not. The survey form is available at https://forms.gle/q4QPtEwWvi8tPEpX6.

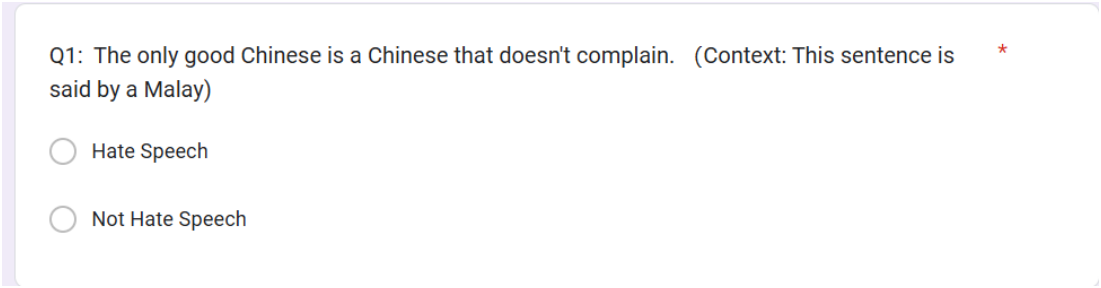


Figure 3.2: Example of the Questions in a Survey

Table 3.1 Sentence Examples inside Survey and Their Results

Sentence Examples	Observations	Ground Rules
Q2: Feminism is all about hating males.	These sentences either talk about certain group of people in a negative and irrational manner or advocate actions that negatively impact them. Overwhelming amount of participants (76%, 82%, 70%) think these sentences are hate speech.	Definition of hate speech: language that attacks and incite violences against specific groups, including promoting stereotypes, misleading information or subtle humour (Fortuna & Nunes, 2019) (Waseem & Hovy, 2016)
Q10: We must boycott non-Muslim products to take a stance for Palestine.		
Q19: Foreigner fart smell nice. (in a sarcastic manner)		
Q7: No matter what a Muslim is, he is better than non-Muslims.	This sentence argues that a group of people is better than another group. 88% of participants believe that this is a hateful argument.	Comparison on different demographics to be better or worse is considered as hate speech
Q12: Many Chinese are also gangsters. (This sentence is asked once without context, and once with the context of denying that Malays are gang members)	These sentences does not describe every single member in a particular group to be related to negative stereotypes. Instead, it only refers to certain members inside the group. However, 65% of participants refer both Q12 (with and without context) and Q18 as hate speech.	Criticism of certain demographic or its related ideology is considered as hate speech if without solid argument, even if the criticism is explcitly targetted only against certain people (Waseem & Hovy, 2016)
Q18: Malay historians and lying. Name a more iconic duo. (Context: This is a reply to a news about a Malay historian faking data for his thesis)		

Q3: Non-Muslim should keep quiet about anything related to Islam. (Context: This sentence is said by a Muslim)	This sentence argues that people outside of a group should not have any opinion on the group. 59% of participants agree that this sentence is a hate speech.	The banning of constructive criticism, communication, discussion and support on certain demographic is considered as hate speech
Q4: Fake Christian who didn't read the bible should not give his opinions on God here. (Context: This sentence is said by a Christian)	This sentence argues that people inside a group should understand that specific knowledge and culture to be deemed as "one of us". 53% of participants agree that this argument is valid and is not purposely spreading hatred against specific people.	People are allowed to criticise their own demographics and isolate people who they deem not qualified enough to be "one of them"
Q20: Malaysian can only be united if the minorities are assimilated.	This sentence call for at least one group of people to assimilate and abandon their initial identity and culture. 59% of participants believe that this sentence is hateful.	Advocation of a demographic to assimilate and lose its own identity is considered as hate speech

After ground rules have been set, the author can manually categorise social media posts into hate speech and non-hate speech. Thesis study also manually categorises hate speech into "Category" label with four possible values, which are none, race, religion and gender. While hate speech of certain aspects such as sexual orientation is prominent in Western world, the general consensus of Malaysia believe that criticizing these sexual orientations is actually good and morale (Jerome et al, 2021). Therefore, most social media posts in the data model mentions race, religion or gender. The data model also has an extra "SubCategory" label that states the exact race, religion or gender a post is talking about. One thing to note that there are only male and female tags in this label, since almost nobody in Malaysia is actually non-binary.

Even though the “SubCategory” label for race mainly consists of the races of local Malaysians, the thesis study does gather social media posts talking about foreign tourist and workers, as long as the poster is a local Malaysian. The reason to stop Malaysians from spreading unreasonably negative impression on foreigners and advocate hatred on them in the social media is simple. According to the broken window theory (Green, 2023), physical signs of minor crimes being neglected can lead to worse and more serious crime taken place. This same theory can be implied here as well. If Malaysians are allowed to spread hate speech against foreigners, this will encourage them to take a step further and advocate for hatred against local Malaysians. Not to mention that spreading hate speech against foreigners cause physical and mental harm towards them already.

When more than one group of people is mentioned in a social media post, the tag “Multiple” is then used for the “SubCategory” label. The tag “Unknown” works exactly the same as the previous one, except this time it is about a group of people with unknown race, religion or gender. If someone is talking about the general idea of race, religion or gender, the tag “None” is then used in the “SubCategory” label. If the tag “None” is used in both “Category” and “SubCategory” labels, that means the social media post does not even mention anything about race, religion or gender.

No	Category	SubCatego	Post	HateSpeech
1	Race	Chinese	Chinese are pigs!	1
2	Race	Indian	I don't like pajeets they smell like curry	1
3	None	None	Israel is defending itself from Hamas	0
4	Race	Indian	Indians always smell stinky and weird, what is up with that	1
5	Gender	Female	Women are so emotional they can't handle anything ar	1

Figure 3.3: Snippet of Data Model

3.3 Data Preprocessing

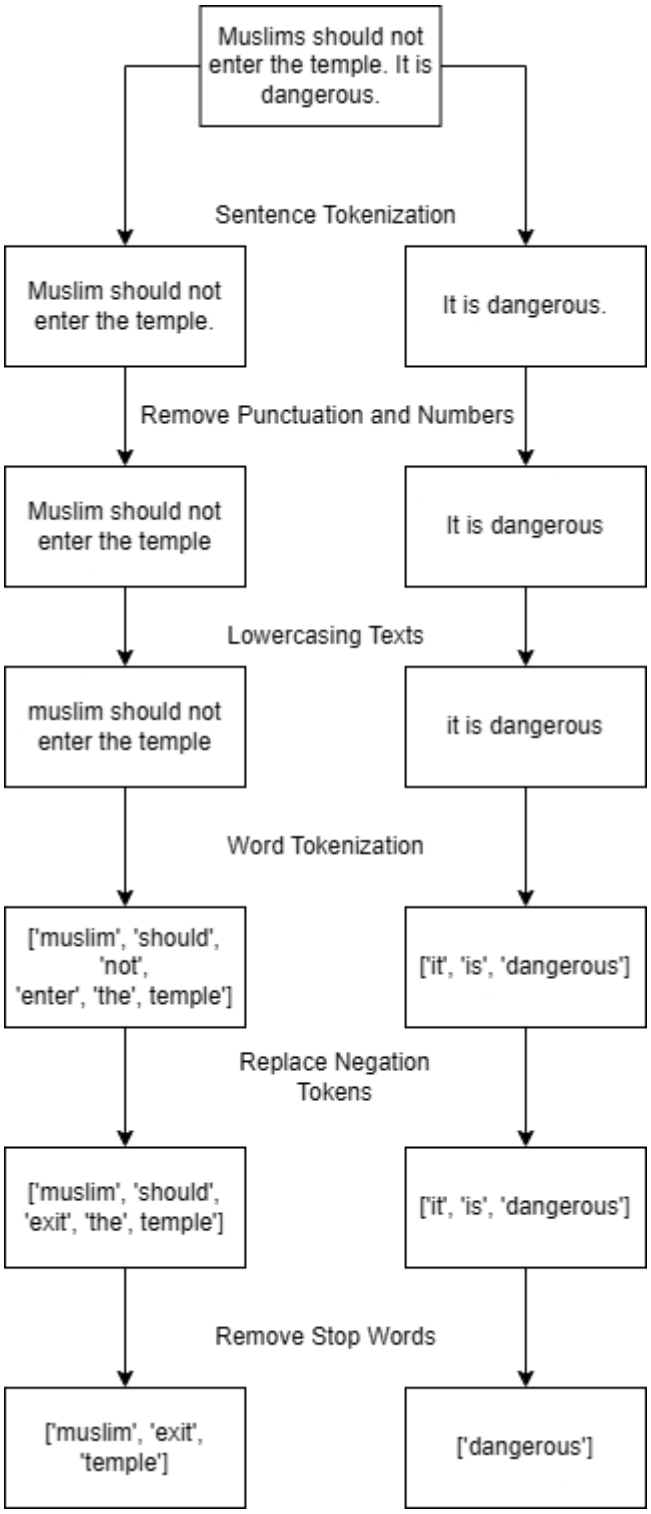


Figure 3.4: Preprocessing Steps of a Social Media Post

After manually collecting social media posts and labelling them with their categories and hate speech values, the data are now going through a preprocessing phase. Since each social media post may have sentences that potentially contradicting each other, every sentence inside post will first be tokenized into independent instances. The original dataset had 1352 posts, and the tokenized dataset had 1977 sentences after the tokenization. This will ensure that the machine learning algorithms are able to train with every single sentence in the data thoroughly.

After that, each instance will go through a list of functions to be processed into word tokens. Punctuations and numbers may affect the learning algorithms from actually recognizing the relationships between different words. Therefore, they are removed from the data. After lowercasing the tokens, the text data in each instance are tokenized into word tokens. After that, the process remove the negation tokens such as “not” and “neither” from the data. The words exactly behind the negative tokens are also replaced with its antonyms. Stopwords are words that commonly appears in the dataset but has no actual meaning. Therefore, stopwords are also removed from the text data to simplify the data and shorten its length. Lemmatization and stemming are actually not performed in the data preprocessing phase. This is because the data model contains many specific terms that are not actually included in the library of these two modules. They might misunderstand the meaning of these terms and change them into unrelated words.

One important thing to note is that this preprocessing has been performed twice here. This is because the hate speech dataset model has been extracted and pre-processed twice. One additional dataset models has been created with every instance not related to race, religion or gender removed. This new dataset with size of 1620 has its ratio of hate speech and non-hate speech to be almost 1:1 for each category. Therefore, this will be the standard dataset to evaluate the accuracy of every machine learning algorithms.

	No	Category	SubCategory	Post	HateSpeech
2	3	NaN	NaN	Israel is defending itself from Hamas	0
24	25	NaN	NaN	Malaysia truly Asia.	0
25	26	NaN	NaN	I love cars	0
27	28	NaN	NaN	Malaysia has better food than Singapore	0
28	29	NaN	NaN	Support Hamas and you support terrorist	0
...	...	...	...	...	...
986	987	NaN	NaN	I hate him so much i want him to die. He stole...	0
993	994	NaN	NaN	Accessibility in Malaysia in a picture:	0
995	996	NaN	NaN	He seems mentally unwell	0
996	997	NaN	NaN	I don't want to meet this psychopath	0
997	998	NaN	NaN	i still wonder why can't the royal families in...	0

233 rows × 5 columns

Figure 3.5: Every Instance That is Not Related to Race, Religion or Gender in Pre-processed Dataset

```
#check which dataset has NaN

none1 = df1.isna().any()

none2 = df2.isna().any()

print(f"Original dataset:\n{none1}\n")
print(f"Standard Dataset:\n{none2}")
```

```
Original dataset:
No             False
Category        True
SubCategory     True
Post           False
HateSpeech     False
dtype: bool

Standard Dataset:
No             False
Category       False
SubCategory    False
Post           False
HateSpeech     False
dtype: bool
```

Figure 3.6: Comparison of Pre-processed and Standard Dataset

```
df2['HateSpeech'].value_counts(normalize=True, dropna=False)
```

```
HateSpeech
1    0.501852
0    0.498148
Name: proportion, dtype: float64
```

```
race = df2[df2['Category']=='Race']
religion = df2[df2['Category']=='Religion']
gender = df2[df2['Category']=='Gender']
```

```
race['HateSpeech'].value_counts(normalize=True, dropna=False)
```

```
HateSpeech
1    0.502098
0    0.497902
Name: proportion, dtype: float64
```

```
religion['HateSpeech'].value_counts(normalize=True, dropna=False)
```

```
HateSpeech
1    0.500838
0    0.499162
Name: proportion, dtype: float64
```

```
gender['HateSpeech'].value_counts(normalize=True, dropna=False)
```

```
HateSpeech
1    0.503247
0    0.496753
Name: proportion, dtype: float64
```

Figure 3.7: Proof that Ratio of Hate Speech over Non-hate Speech Standard Data Model is Almost 1:1

3.4 Topic Embedding Models

Since BERTopic often gives the best result according to researches from the literature review, this thesis study has converted the texts from the custom made hate speech data model into BERTopic embeddings.

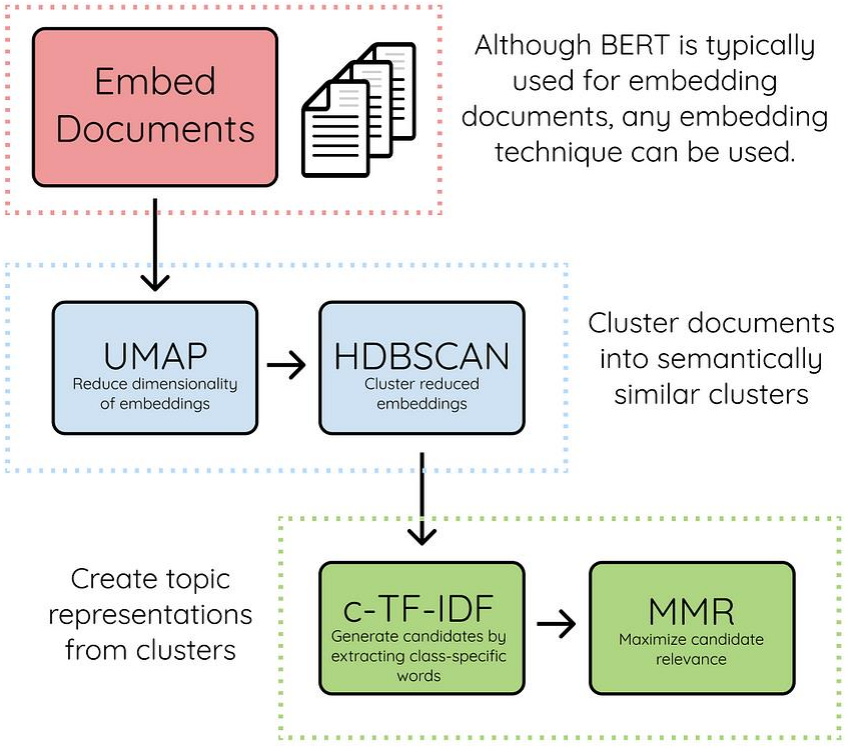


Figure 3.8: Steps of a BERTopic Algorithm (Distante, 2022)

According to (Distante, 2022) and (Grootendorst, 2022), BERTopic embedding algorithm first converts the data into BERT embedding. After that, the dimensionality of the BERT embeddings have been greatly reduced to prevent overfitting. UMAP algorithm is used for dimensionality reduction as it can preserve the local and global embedding structures. HDBSCAN data clustering algorithm is then applied to create topic clusters. Each topic clusters contains instances that are semantically similar to each other. Finally, BERTopic algorithm generates topic representations by finding the most important words with highest importance scores. The importance scores of words within the clusters are calculated TF-IDF algorithms.

PreprocessedToken	Topic_Probabilities	Topic	Representative_Docs
['chinese', 'pigs']	[4, 0.353802893618629]	[chinese, china, people, pig, go, indians, ric...	[chinese pig, pig chinese, chinese people chin...
['unalike', 'pajeets', 'smell', 'like', 'curry']	[7, 0.21967016813460452]	[indians, indian, pajeets, smell, pajeet, fuck...	[like indians, indians, indians]
['indians', 'smell', 'stinky', 'weird']	[7, 0.20338674496166917]	[indians, indian, pajeets, smell, pajeet, fuck...	[like indians, indians, indians]

Figure 3.9: Hate Speech Data Model after Going Through BERTopic Word Embedding

Figure 3.9 shows that BERTopic algorithm had clustered different instances into the same topic and representative document. Topic is the most significant word tokens inside the topic while representative document is about the most significant text documents. Furthermore, each instance also has the probability of it being in that specific probabilities ("Topic_Probabilites").

CBOW, fastText and Latent Dirichlet Allocation (LDA) word embeddings are the word embedding alternatives that have been used in this thesis study. CBOW and fastText are implemented here because Ariwibowo et al. (2022) reported that they managed to build a hate speech detection model with highest accuracy with fastText + CBOW model. Meanwhile, LDA is merely implemented as a control group since no literature paper has mentioned it to have great accuracy classifying the result. Their main purpose is to confirm if the accuracy of models with BERTopic embedding are the highest. If BERTopic failed to create word embeddings that provide great result, these alternatives can be used to train machine learning algorithms.

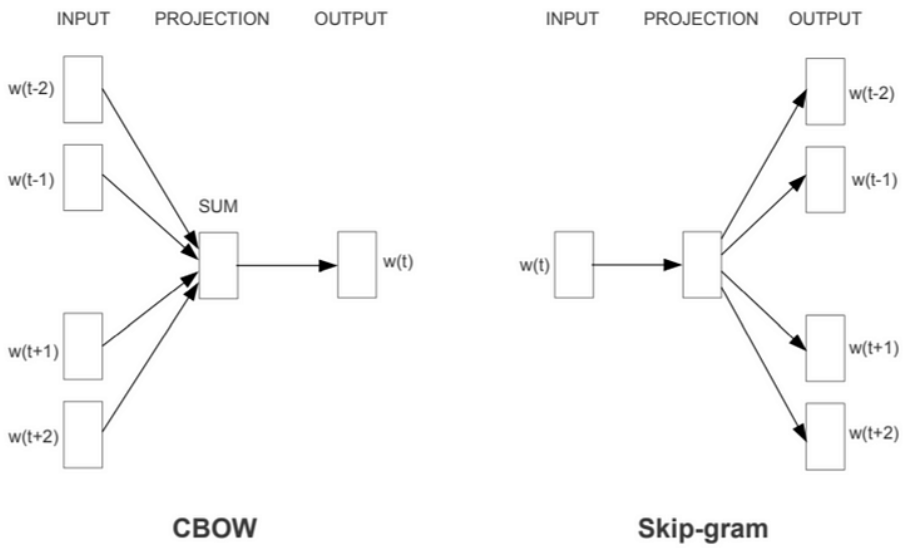


Figure 3.10: Steps of Projecting Embeddings in CBOW and Skip-gram (Kulshrestha, 2019a)

Kulshrestha (2019a) states that CBOW has a metric called window size. It is basically the size of the context words surrounding both sides of the target word. CBOW first list down the target words that can has its window completely filled with context words. Then, these context words will go through layers to predict what is the target word. After CBOW embeddings (target words) have been found, the author manually remove word tokens in the data model that do not have their corresponding CBOW embeddings. Now every word token in the data model instances can be converted into CBOW without any error or containing empty token or mojibake. This data model is now ready to be used in upcoming machine learning algorithms. An additional Skip-gram algorithm is also implemented in this thesis study. Unlike CBOW, Skip-gram predicts the context words based on the representation of input word.

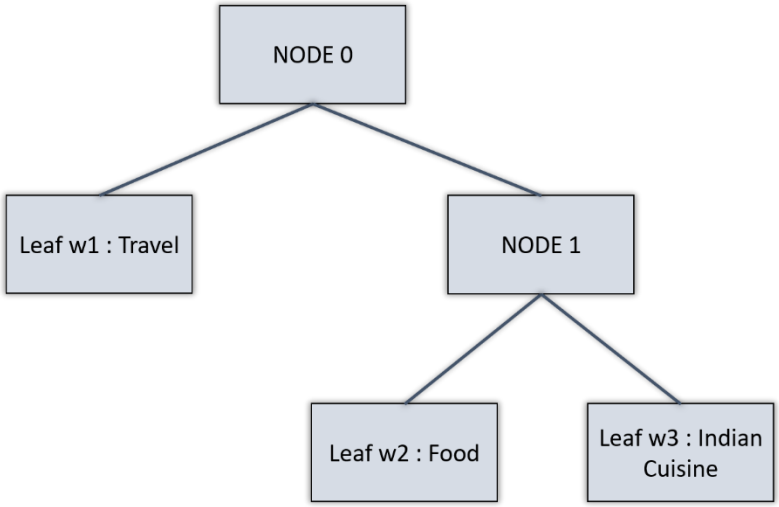


Figure 3.11: Binary Tree on fastText's labels

(*FastText Working and Implementation*, 2024)

fastText is a library to create word representation and classification. Therefore, it can directly classify instances in the data model as hate speech or non-hate speech. Normally, word embedding model search through every single label and compute their score. However, the architecture of fastText resembles a hierarchical tree. The fastText algorithm only has to search through the nodes inside the tree to find the target label. It does not need to actually go through other labels. This heavily reduces the time complexity as fastText algorithm only needs to compute once when it reaches target label (*FastText Working and Implementation*, 2024).

Given the number of topics is already known, LDA randomly assign word tokens in one of the topics. Then, LDA finds out two proportions for every single word and for every token respectively. They are the proportion of a specific word token belongs to a certain topic for a given instance, and also the proportion of a specific word from this certain topic all over the instances. These two proportion can be used to find the probability for a word belonging to a specific topic. Each instance finds the topic it belongs to, and what are the word tokens inside the topic. The work tokens inside the topic are added into the data model and used for future machine learning algorithms (Kulshrestha, 2019b).

3.5 Naïve Bayes & DCNN Models

After preprocessing and converting corpus data into word embeddings, machine learning and deep learning algorithms are now started to get built. The machine learning algorithms used in the thesis study are different types of Naives Bayes model. Meanwhile, the deep learning algorithms used in building hate speech detection models are Deep Convoluted Neural Network (DCNN), Long Short-Term Memory (LSTM) and GateRecurrent Unit (GRU) model. This study initially plan to copy what Al-Hassan & Al-Dossari (2021) had done, which is to build a combination of deep learning algorithms, such as LSTM + DCNN. However, this does not come true due to the time constraint given to complete this thesis study. Besides, the complexity of a model combined with different algorithms may cause uncertainties. The accuracy of combinations of two deep learning algorithms may be low and it will be hard to figure out what caused the problem.

Naive Bayes models are a group of algorithms that relies on Bayes theorem and the probabilities of a certain value in different labels respectively. Bayes theorem basically calculate P (B | A) given that it already knows the values of P(A), P(B) and P (A | B) (Prabhakaran, 2022). In this scenario, Naïve Bayes model applies the probabilities of hate speech posts to contain a specific word embedding, to calculate the probability of instances with this word embeddings to be hate speech. These probability scores are saved as features in the hate speech detection system, so the system can classify new social media post as hateful or not, without further training. This thesis study trains Gaussian, Bernoulli and Multinomial Naïve Bayes models. While Gaussian and Multinomial models calculates the feature values in normal and multinominal distributions respectively, Bernoulli Naïve Bayes model calculates them as independent binary values (Prabhakaran, 2022). Each Naïve Bayes model is also trained with topics, topic probabilities and representative documents from BERTopic embedding model, and also the CBOW embeddings. This is to figure out which embeddings give the best accuracy to Naïve Bayes model.

Naïve Bayes model is considered as machine learning algorithm instead of deep learning algorithm due to the architecture of Native Bayes model being relatively simple and straight forward. Unlike other deep learning algorithms, Naïve Bayes does not need to train in multiple neural network layers (Oppermann, 2023). However, many assumptions have been made in the Naïve Bayes model so it can actually perform tasks. Such assumption includes the features of the model are always independent to each other and as important as others. It also assume the hate speech data model does not lost any of its data (Prabhakaran, 2022). These assumptions may decrease the accuracy of hate speech detection system as the detection system has limited ability to handle continuous features that has huge correlation between each other.

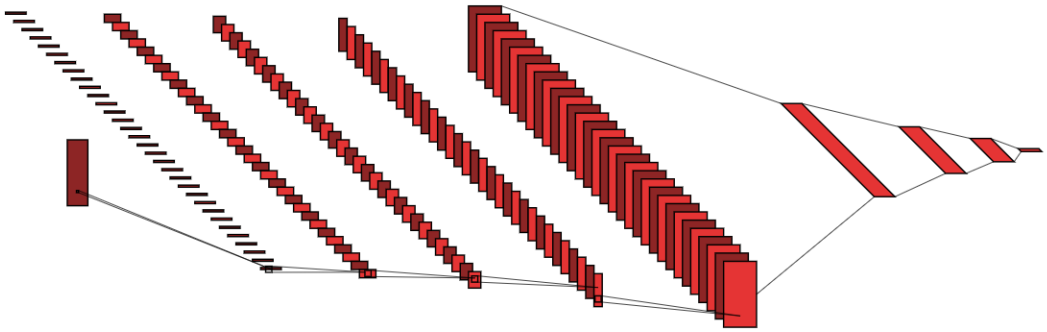


Figure 3.12: Custom Made DCNN Model
with Topics as Input

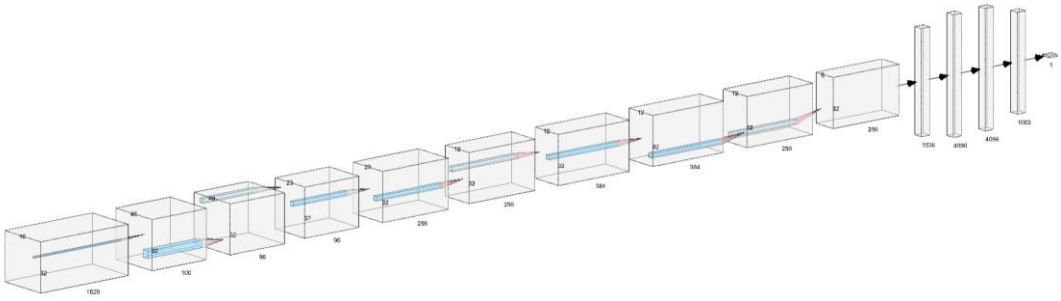


Figure 3.13: AlexNet Model

A custom made DCNN model and AlexNet model are created and trained for hate speech detection. Initially, AlexNet model was not planned to be used in the thesis study. However, there is a risk that the custom model may not perform well. Thus, the AlexNet model serves as a control group. According to Moolayil (2021), the difference between CNN and DCNN model is only the number of layers used in the model, and the exact number of layers to differentiate CNN and DCNN models are always different to different engineers and scientists. As the original AlexNet model only has 5 convolutional layers and 3 fully connected layers, it is safe to assume that the custom model is a DCNN model since this custom DCNN model has 4 one-dimension convolution layers and 3 fully connected layers (Krizhevsky et al., 2017).

One additional embedding layer has added into the AlexNet model before the first convolution layer. This is because the original embedding layer cannot compress the original size of 1600 instances to no less than 100. One Dense layer with size of 1 has also added in the end because the final result of every model instance should be only one integer value. This value determines whether the instance is hate speech or non-hate speech. This custom DCNN model is trained 5 times with topics, topic probabilities and representative documents from BERTopic embedding model, CBOW embeddings and the original pre-processed texts. As the input data only has size of 2 dimensions (batch size, length of word embedding), they are first converted into three dimensions (batch size, number of channels, length of word embedding). One conv1D layer created inside is immediately followed by one ReLU and one dropout layer. This is to prevent vanishing gradient problem and overfitting respectively.

After resizing the word embeddings back to two dimensions (batch size, number of channels * length of word embeddings), the word embeddings go through the fully connected layers. Every time they go through a fully connected layer, they will also go through a dropout layer.

Meanwhile, only topic embeddings from BERTopic model and pre-processed tokens go through AlexNet model. The reason is due to AlexNet model only accept inputs that has its element only a word token. The elements inside representative document is a string consists of multiple word tokens, while the elements inside topic probability is an integer tuple. In the beginning, the dimension of word embeddings / word tokens has been condensed from 1620 (the number of instances in data model) to 100. As text data generally is not as condensed as music or image data, the condensed CBOW embeddings can now provide more semantic context and more detailed relationship between different word tokens to the AlexNet model. After that, the word embeddings goes through several Conv1D and MaxPooling layers before getting flattened into one dimension vectors.

3.6 LSTM and GRU Models

These two models are initially built to verify the findings of Al-Hassan & Al-Dossari (2021). However due to time constraint, the author is only able to train LSTM and GRU models, without combining it with an additional DCNN model. However, it is still a crucial step in this thesis study to determine whether LSTM, GRU or DCNN will have higher accuracy in categorizing social media posts in Malaysian context into hate speech and non-hate speech.

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_3 (Embedding)     (None, 46, 100)           162000

 lstm_6 (LSTM)               (None, 46, 100)           80400

 lstm_7 (LSTM)               (None, 100)               80400

 dropout_6 (Dropout)         (None, 100)               0

 dense_6 (Dense)             (None, 50)                5050

 dropout_7 (Dropout)         (None, 50)                0

 dense_7 (Dense)             (None, 1)                 51

=================================================================
Total params: 327901 (1.25 MB)
Trainable params: 327901 (1.25 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 3.14: The Composition of LSTM Model

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 46, 128)           640000

 gru_2 (GRU)                 (None, 46, 128)           99072

 gru_3 (GRU)                 (None, 128)               99072

 dropout_2 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 64)                8256

 dropout_3 (Dropout)         (None, 64)                0

 dense_3 (Dense)             (None, 1)                 65

=================================================================
Total params: 846465 (3.23 MB)
Trainable params: 846465 (3.23 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 3.15: The Composition of GRU Model

These two models are trained with the topics, topic probabilities and representative document of BERTopic models and the original preprocessed word tokens. After embedding the dimension of input data model from 1620 to 100 and 128 respectively, the input data will go into two different LSTM/GRU layers. The reason there are two LSTM/GRU layers instead of one in their respective models, is because running them twice technically increases the number of epochs being ran in both of these models. Therefore, these models will have more experiences on understanding the semantic contexts of the input texts. After going through a few dropout and dense layers, the whole model has done their training processes.

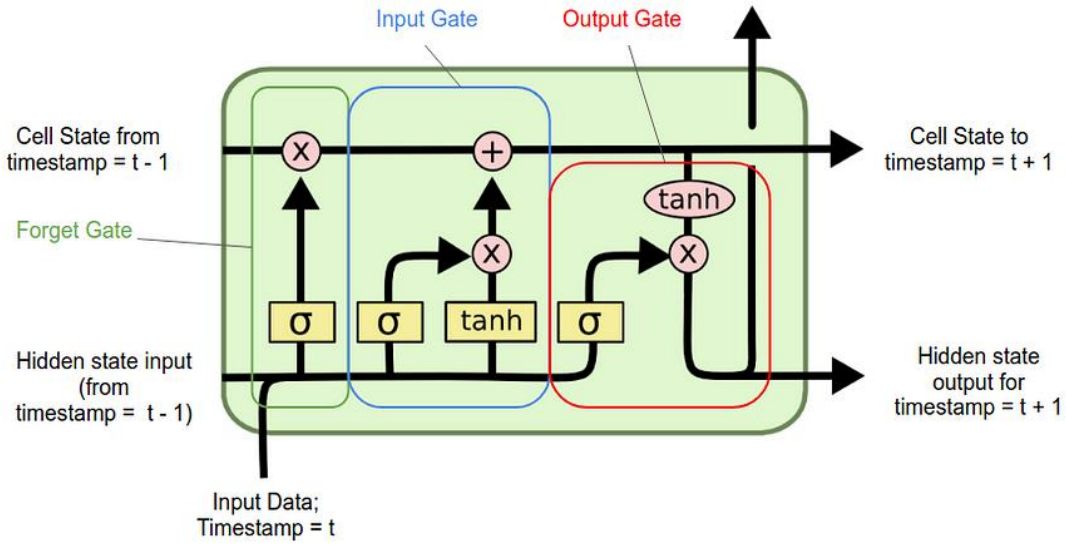


Figure 3.16: Structure of LSTM Cell (Vincent, 2020)

Basically, one LSTM or GRU layer consists of multiple LSTM / GRU cells. Each cell receives text data from outside and from previous cells (if it has any). After processing the data, this cell can pass these data to the next cell or release it out. Of course, LSTM and GRU cells have the ability to choose the amount of data they want to receive and release. Vincent (2020) states that one LSTM cell consists of forget gate, input gate and output gate. Forget gate controls the amount of data from previous state is being

transferred into the current cell. Meanwhile, input gate determines how much data can be received from the outside. Lastly, output gate decides what kind of data to be released out of the LSTM layer and send to the next LSTM cell.

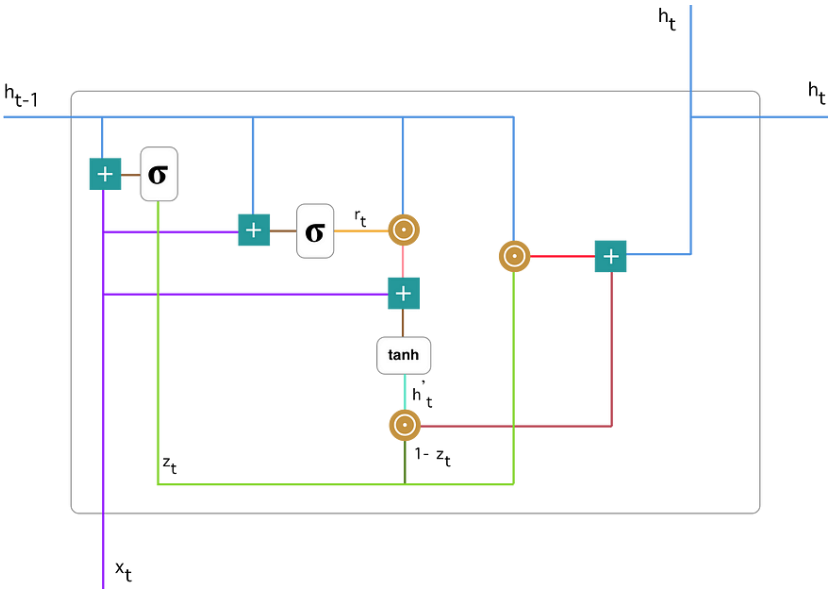


Figure 3.17: Structure of GRU Cell (Kostadinov, 2017)

Based on Kostadinov (2017), GRU is an algorithm that is created to be a better and simpler alternative of LSTM. Thus, it uses gating algorithms that is similar to LSTM. When input word embeddings and value obtained from previous hidden layer flow into a GRU hidden layer, they first go through an update gate so only certain amount of previous information and current input will be used. These variables also go through a reset gate. The reset gate decides how much past information should be forgot and reset to zero. The equations above determine value of both gates and the x and h value have not been reduced by the gates yet. The formulas for them are the same and the only difference is the weight values in each gate.

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1})$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1})$$

Figure 3.18: Formula for Update & Reset Gate (Kostadinov, 2017)

From the equation below, x and h values now go through both gates. The formula above showed that how each gate will affect the final output of h value. The reset gate first removes some portion of the previous GRU layers. After that, the update gate then decides how much of the current input and previous GRU layers (after some portions of it has been removed) will be pass through itself. This portion updates the information of previous GRU layers, so it includes the input of the current layer. The newly obtained h value will then go through the next GRU layer until every GRU layer has been visited. GRU algorithm will eventually output a h value at the end of GRU algorithms.

$$h_t^{'} = \tanh(Wx_t + r_t \odot Uh_{t-1})$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t^{'}$$

Figure 3.19: Formula for Information of Previous GRU Cells (Kostadinov, 2017)

3.7 Message Detection Program

After machine learning and deep learning algorithm have been completely trained, a message detection program is created using the algorithm that gives the best validation accuracy. It can detect whether what user have inputted is a hate speech or not. Figure 3.20 shows how the user interface looks like when the user has first opened the software. The software asks for users to input any sentence.

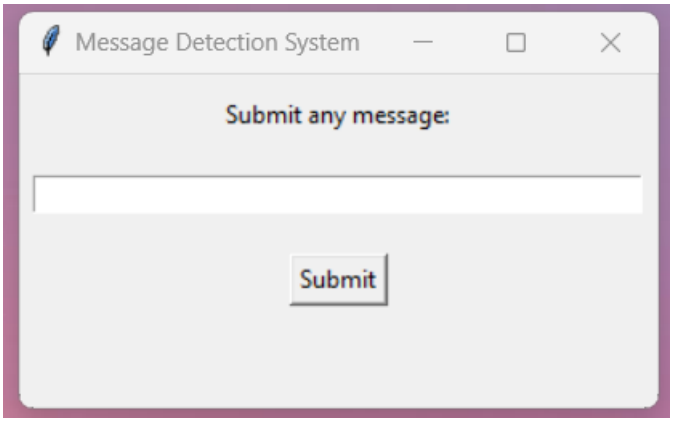


Figure 3.20: Message Detection Program before Submitting Any Input

Figure 3.21 describes the moment when a user has sent text messages to the program. The program will inform the user whether these messages are hate speech or not. Users can repeatedly submit as many messages as they want to this program, without any interruption.

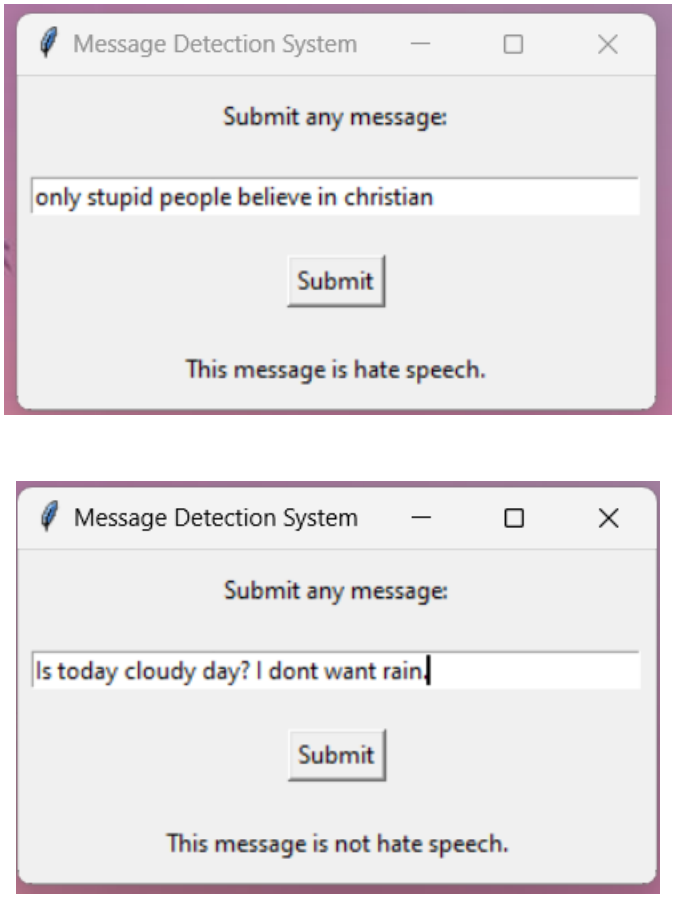


Figure 3.21: Message Detection Program after Submitting Any Input

This hate speech detection program is programmed using the Tkinter module. When the program has received a user input, it will first perform preprocessing to convert it into word tokens. If the input has more than one sentences, the program will actually combine them into one single string first before tokenizing them. This is different from the training process, which sentences are separated into different instances instead. After that, the word token will be sent into a local GRU hate speech detection model that is trained with pre-processed token without any word embeddings. This GRU model will give a value between 0 and 1. Any value that is over 0.5 is considered as hate speech.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Naïve Bayes Model

The images below show the accuracies of Naïve Bayes models while training with BERTopic word embeddings. The topic, topic probabilities and representative documents are word embeddings from the BERTopic model.

```
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
#TOPIC
```

```
Gaussian Naive Bayes model accuracy(in %): 52.674897119341566
```

```
from sklearn import metrics
print("Multinomial Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
#TOPIC
```

```
Multinomial Naive Bayes model accuracy(in %): 53.49794238683128
```

```
from sklearn import metrics
print("Bernoulli Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
#TOPIC
```

```
Bernoulli Naive Bayes model accuracy(in %): 51.028806584362144
```

Figure 4.1: Accuracy of Naive Bayes Model Training with Topics

```
from sklearn.naive_bayes import MultinomialNB
# TOPIC PROBABILITY
nb_classifier = MultinomialNB()
nb_classifier.fit(x_train, y_train)

accuracy = nb_classifier.score(x_test, y_test)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.4547325102880658
```

```
from sklearn.naive_bayes import BernoulliNB
# TOPIC PROBABILITY
nb_classifier = BernoulliNB()
nb_classifier.fit(x_train, y_train)

accuracy = nb_classifier.score(x_test, y_test)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.5205761316872428
```

```
from sklearn.naive_bayes import GaussianNB
# TOPIC PROBABILITY
nb_classifier = GaussianNB()
nb_classifier.fit(x_train, y_train)

accuracy = nb_classifier.score(x_test, y_test)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.4403292181069959
```

Figure 4.2: Accuracy of Naive Bayes Model Training with Topic Probabilities

```
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test2, y_pred2)*100)
#REPRESENTATIVE DOCUMENT
```

```
Gaussian Naive Bayes model accuracy(in %): 48.971193415637856
```

```
from sklearn.naive_bayes import MultinomialNB
gnb2 = MultinomialNB()
gnb2.fit(x_train2, y_train2)
y_pred2 = gnb2.predict(x_test2)
print("Multinomial Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test2, y_pred2)*100)
#REPRESENTATIVE DOCUMENT
```

```
Multinomial Naive Bayes model accuracy(in %): 47.53086419753087
```

```
from sklearn.naive_bayes import BernoulliNB
gnb3 = BernoulliNB()
gnb3.fit(x_train2, y_train2)
y_pred2 = gnb3.predict(x_test2)
print("Bernoulli Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test2, y_pred2)*100)
# REPRESENTATIVE DOCUMENT
```

```
Bernoulli Naive Bayes model accuracy(in %): 47.53086419753087
```

Figure 4.3: Accuracy of Naïve Bayes Model Training with Representative Documents

It seems like the BERTopic word embeddings does not provide great accuracy on predicting social media posts to be hate speech or not. Before making any speculation, the thesis study checks how good Naïve Bayes models perform if they are trained with CBOW and Skip-gram representations. One thing to note is that both of these embedding models outputs both word embeddings with positive and negative values. Multinomial Naïve Bayes cannot accept negative value as input because multinomial distribution is used to find the probability of number of amount of events happen. An event can happen 0 times, or more than 0 times, but never less than 0 times.

```
# Split the data
X_train, X_test, y_train, y_test = train_test_split(doc_vectors, df['HateSpeech'], test_size=0.3)

# Initialize and train the Naive Bayes classifier
Gaussian_classifier = GaussianNB()
Gaussian_classifier.fit(X_train, y_train)

# Predict on the test set
y_pred = Gaussian_classifier.predict(X_test)

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.602880658436214
```

```
# Initialize and train the Naive Bayes classifier
nb_classifier = BernoulliNB()
nb_classifier.fit(X_train, y_train)

# Predict on the test set
y_pred = nb_classifier.predict(X_test)

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

#Cannot do MultinomialNB as it has negative value
```

```
Accuracy: 0.565843621399177
```

Figure 4.4: Accuracy of Naïve Bayes Model Training with CBOW Embeddings

```
# Split the data
X_train, X_test, y_train, y_test = train_test_split(skipgram_doc_vectors, df['HateSpeech'], test_size=0.3)

# Initialize and train the Naive Bayes classifier
Gaussian_classifier = GaussianNB()
Gaussian_classifier.fit(X_train, y_train)

# Predict on the test set
y_pred = Gaussian_classifier.predict(X_test)

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.5987654320987654
```

```
# Initialize and train the Naive Bayes classifier
nb_classifier = BernoulliNB()
nb_classifier.fit(X_train, y_train)

# Predict on the test set
y_pred = nb_classifier.predict(X_test)

# Evaluate the classifier
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

#Cannot do MultinomialNB as it has negative value
```

```
Accuracy: 0.5473251028806584
```

Figure 4.5: Accuracy of Naïve Bayes Model Training with Skip-gram Embeddings

Even though Naïve Bayes model with CBOW and Skip-gram embeddings only have accuracy of around 55 to 60%, they still perform slightly better than models with BERTopic embeddings. This does not match with the claims of literature papers, who stated that BERTopic embeddings gave the best result in classifying hate speech and non-hate speech. There are a few possible reasons for it to happen. However, it is better to look at the topic representations trained by BERTopic model first.

Figure 4.6: Representations of Top 10 Topics with Highest Frequencies

Most of the topics in BERTopic model actually has its “theme” focusing on race, religion and gender communities in Malaysia. The word representations also correlate with each other. Therefore, the word representations inside BERTopic is not the problem causing the Naïve Bayes model to have low accuracy. According to the author of BERTopic, BERTopic has a design flaw that each instance only belongs to a topic (Grootendorst, n.d.). If an instance is mentioning issues involving more than one topic, it will still show only one topic to the Naïve Bayes model. Thus, the model is unable to understand the whole instance context and classify the instance correctly.

Before looking at next model, the accuracy of Naïve Bayes model trained in LDA embeddings is evaluated. Since LDA is supposed to be a control group for other algorithms to compare, there is no surprise that the models’ accuracies are pretty average here. However, its accuracy is still higher than model trained with BERtopic representations.

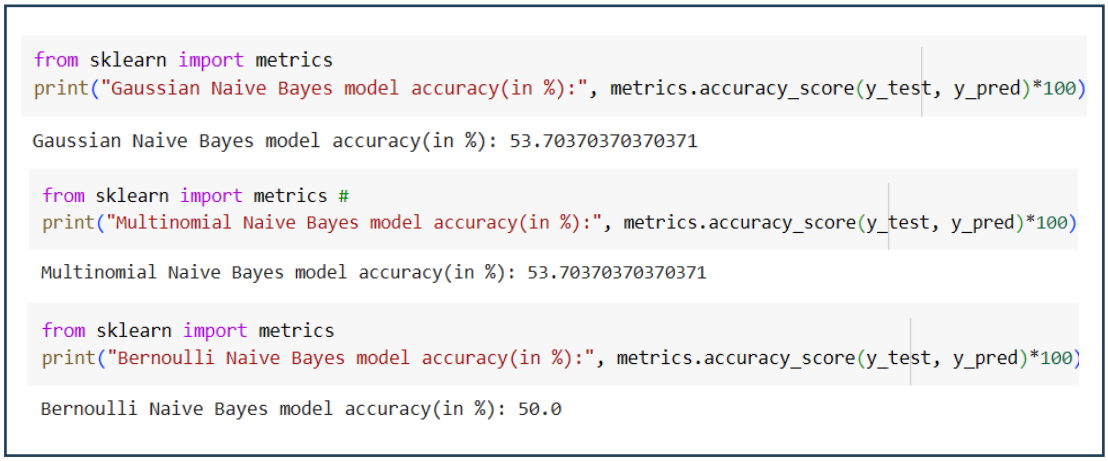


Figure 4.7: Accuracy of Naïve Bayes Model Training with LDA

4.2 FastText

As FastText already provides a function to classify full sentences of social media posts as hate speech or non-hate speech, no additional machine learning or deep learning is trained here. The accuracy of fastText can be surprisingly high and surprisingly low at same time. It all depends on the learning rate of fastText model.

Figure 4.8: Accuracy of fastText model with different learning rates

The tuple that is printed every time a fastText go through validation process is simply (number of test instances, precision, recall). The precision and recall values of the fastText model are the highest (91%) when learning rate is 0. The precision and recall values will drop in a very large degree with learning rate increasing, until learning rate reaches 1x10^-4. After that, the precision and recall values will gradually increase until they stop when they reach about 47%.

This is likely due to the speciality of hate speech data model in Malaysian context. As this dataset consists of real social media conversations extracted from the internet, the relationship between words and the overall context in each instance is already fixed. Learning rate basically lets fastText model changing a little portion of the word

embeddings to be random integers. Therefore, the pre-existing relationship and context may be messed up even if only a small portion of word data is changed. This may cause fastText model to unable recognise the original meaning of the instances in the hate speech data model. Nevertheless, fastText model is clearly not a great model to detect hate speech in social media if some noises in data can completely ruin its performance.

4.3 Deep Learning Algorithms

In the beginning, the results of custom built DCNN and AlexNet models are evaluated and compared to find which DCNN model can determine social media post as hate speech or non-hate speech more accurately.

```
# Evaluate the model (Topic Probabilities)
model.eval()
correct = 0
total = 0
with torch.no_grad():
    for inputs, labels in dataloader:
        outputs = model(inputs)
        max_value, predictions = torch.max(outputs, 1)
        correct += (predictions == labels).sum().item()
        total += labels.size(0)

accuracy = correct / total
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.51
```

```
# Evaluate the model (Topic)
model.eval()
correct = 0
total = 0
with torch.no_grad():
    for inputs, labels in dataloader:
        outputs = model(inputs)
        max_value, predictions = torch.max(outputs, 1)
        correct += (predictions == labels).sum().item()
        total += labels.size(0)

accuracy = correct / total
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.51
```

```python
# Evaluate the model (Representative Documents)
model.eval()
correct = 0
total = 0
with torch.no_grad():
    for inputs, labels in dataloader:
        outputs = model(inputs)
        max_value, predictions = torch.max(outputs, 1)
        correct += (predictions == labels).sum().item()
        total += labels.size(0)

accuracy = correct / total
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.49
```

```python
# Evaluate the model (PreprocessedToken)
model.eval()
correct = 0
total = 0
with torch.no_grad():
    for inputs, labels in dataloader:
        outputs = model(inputs)
        max_value, predictions = torch.max(outputs, 1)
        correct += (predictions == labels).sum().item()
        total += labels.size(0)

accuracy = correct / total
print(f'Accuracy: {accuracy:.2f}')
```

```
Accuracy: 0.54
```

Figure 4.9: Accuracy of Custom DCNN Model

```
[42] # Evaluate the model (Topic)
     loss, accuracy = model.evaluate(data, y_list)
     print(f'Loss: {loss}, Accuracy: {accuracy}')

     51/51 [==============================] - 5s 106ms/step - loss: 0.7316 - accuracy: 0.4981
     Loss: 0.7315887808799744, Accuracy: 0.4981481432914734

[25] # Evaluate the model (PreprocessedToken)
     loss, accuracy = model.evaluate(data, y_list)
     print(f'Loss: {loss}, Accuracy: {accuracy}')

     51/51 [==============================] - 4s 83ms/step - loss: 0.7418 - accuracy: 0.4981
     Loss: 0.7417582273483276, Accuracy: 0.4981481432914734
```

Figure 4.10: Accuracy of AlexNet Model

Figure 4.10 states that the accuracy of the DCNN model specifically built for this thesis study is on par with a widely recognised AlexNet model. However, both model still have relatively low accuracy. DCNN model is merely not good enough to be used as the hate speech detection system. Luckily, this changed when LSTM and GRU step in.

```
# TOPIC
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 30ms/step - loss: 0.6750 - accuracy: 0.5556
Validation Loss: 0.6749892234802246
Validation Accuracy: 0.5555555820465088
```

```
# TOPIC PROBABILITY
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 30ms/step - loss: 0.6830 - accuracy: 0.5463
Validation Loss: 0.6830440759658813
Validation Accuracy: 0.5462962985038757
```

```
# REPRESENTATIVE DOCUMENT
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 29ms/step - loss: 0.6880 - accuracy: 0.5586
Validation Loss: 0.6880491971969604
Validation Accuracy: 0.5586419701576233
```

```
# PREPROCESSED TOKEN
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 36ms/step - loss: 1.1712 - accuracy: 0.8025
Validation Loss: 1.1712394952774048
Validation Accuracy: 0.8024691343307495
```

Figure 4.11: Accuracy of LSTM Model

```
# Evaluate the model on the validation data of Topic
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 29ms/step - loss: 0.6638 - accuracy: 0.5864
Validation Loss: 0.6638393402099609
Validation Accuracy: 0.5864197611808777
```

```
# Evaluate the model on the validation data of Topic Probability
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 1s 55ms/step - loss: 0.6755 - accuracy: 0.5401
Validation Loss: 0.6754544973373413
Validation Accuracy: 0.540123462677002
```

```
# Evaluate the model on the validation data of Representative Document
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 28ms/step - loss: 0.6859 - accuracy: 0.5247
Validation Loss: 0.6858827471733093
Validation Accuracy: 0.5246913433074951
```

```
# Evaluate the model on the validation data of WordRepresentation
loss, accuracy = model.evaluate(X_val, y_val)
print(f"Validation Loss: {loss}")
print(f"Validation Accuracy: {accuracy}")
```

```
11/11 [==============================] - 0s 38ms/step - loss: 1.0015 - accuracy: 0.8519
Validation Loss: 1.0015413761138916
Validation Accuracy: 0.8518518805503845
```

Figure 4.12: Accuracy of GRU Model

Even though LSTM and GRU models still obtain average accuracy with BERTopic embeddings, both of them manage to obtain validation accuracy of higher than 80%. However, the validation loss for both models is unreasonably high, suggesting that there may be potential overfitting, inappropriate loss functions or learning rate being set too high. Nevertheless, these high accuracies of LSTM and GRU models are obtained after validating from test data that have not been used for training process earlier. Therefore, this thesis study has successfully found a great deep learning algorithm (GRU model with original pre-processed word tokens) to implement in the hate speech detection system.

Given that BERTopic embeddings still remain low, it is recommended for upcoming researches to use the original , manually made hate speech data model in Malaysian context, until someone can find a way to solve this issue. As stated by Vishwanath (2023), BERTopic has a high chance of generating outliers during training phase. The author believes that if this problem has been solved, machine learning models trained with BERTopic embeddings, will be able to effectively recognise hate speech.

Another observation is that DCNN models never performs as good as LSTM and GRU models, even if all three of them are trained with the preprocessed tokens. This may be potentially caused by the nature of these models. LSTM and GRU are made specifically for data that are in a sequences, such as sound and text messages (Gomede, 2023). Meanwhile, DCNN is more suitable to be used to perform pattern or text recognition inside image, as these data are presented to DCNN directly and without sequence.

Table 4.1: Results of Algorithms Used in Thesis

Algorithm	Dataset	Accuracy
Gaussian Naïve Bayes	Topic (BERTopic)	52.7%
	Topic Probability (BERTopic)	44%
	Representative Document (BERTopic)	49%
	CBOW Embeddings	60.3%
	Skip-gram Embeddings	60%
	LDA Embeddings	53.7%
Bernoulli Naïve Bayes	Topic (BERTopic)	51%
	Topic Probability (BERTopic)	52.1%
	Representative Document (BERTopic)	47.5%
	CBOW Embeddings	56.6%
	Skip-gram Embeddings	54.7%
	LDA Embeddings	50%
Multinomial Naïve Bayes	Topic (BERTopic)	53.5%
	Topic Probability (BERTopic)	45.5%
	Representative Document (BERTopic)	47.5%
	CBOW Embeddings	N/A
	Skip-gram Embeddings	N/A
	LDA Embeddings	53.7%

fastText	Learning Rate of 0	91.4%
	Learning Rate of 0.0001	8.6%
	Learning Rate of 0.001	34.6%
	Learning Rate of 0.1	46.9%
	Learning Rate of 0.2	47.5%
	Learning Rate of 1	46.6%
Custom DCNN	Topic (BERTopic)	51%
	Topic Probability (BERTopic)	51%
	Representative Document (BERTopic)	49%
	Pre-processed Word Tokens	54%
AlexNet	Topic (BERTopic)	49.8%
	Topic Probability (BERTopic)	N/A
	Representative Document (BERTopic)	49.8%
	Pre-processed Word Tokens	N/A
LSTM	Topic (BERTopic)	55.6%
	Topic Probability (BERTopic)	54.6%
	Representative Document (BERTopic)	49.8%
	Pre-processed Word Tokens	80.2%
GRU	Topic (BERTopic)	58.6%
	Topic Probability (BERTopic)	54%
	Representative Document (BERTopic)	52.4%
	Pre-processed Word Tokens	85.2%

4.4 Message Detection Program

Even though this message detection program usually can correctly identify input messages to be hateful or not hateful, sometimes it still identify them incorrectly. In this section, the author will focus on the potential reasons causing the hate speech detection system to predict wrongly with real examples.

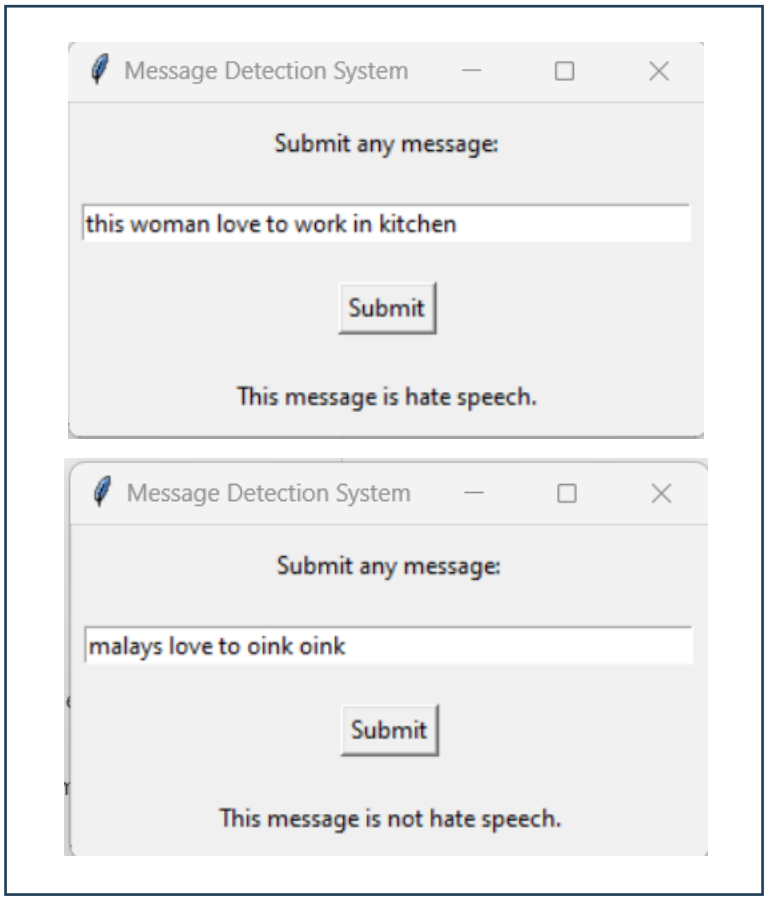


Figure 4.13: Examples of Message Detection Software Predicting Wrongly

It is certain that these incorrect predictions are caused by lack of instances in the hate speech dataset model. As this entire dataset is entirely built from nothing, the author had spent days searching for hate speech made by Malaysians in the social media. However, only small number of hate speech exists on social media without getting wiped out, and these hate speech do not hide in plain sight. Therefore, the author is aware that this dataset is too small for hate speech detection system to project everything correctly.

For instance, a few hate speech in the dataset is about men suggesting that women is only suitable to work in kitchen. This message detection system probably sees the tokens “women” and “kitchen” and assumes that this normal sentence is actually hate speech. The term “oink” does not appear in the data model once. Therefore, the message detection software is unable to relate the term “oink” with pigs.

4.5 Discussion

This thesis study manages to confirm that deep learning algorithms such as GRU and LSTM, performs better than machine learning algorithms. This aligns with the conclusion of most of the literatures. However, this thesis study actually finds that models train with pre-processed list of words has a better accuracy than the ones train with word embeddings, and none of the literature this thesis study have reviewed, agree with this point.

While this thesis study does not disprove the effectiveness of word embedding in training deep learning algorithms, it gives an alternative perspective for researchers to review. Most importantly, the author has successfully created a dataset filled with Malaysian social media posts by scratch, that can teach hate speech detection system to combat against hate speech.

CHAPTER 5

CONCLUSION

5.1 Conclusion

In this thesis study, a Malaysian hate speech data model has been created by obtaining social media posts from the internet. This hate speech data model has been proven to be able to train hate speech detection systems to detect hate speech in Malaysian context. Furthermore, this thesis study has found a deep learning algorithm that performs the best and has made a message detection program for users to input their messages and check whether it is hate speech or not. These data model and algorithm are used to create automatic hate speech detection system for social media. Social media can scam through the posts posted by Malaysia and delete the hateful ones before it gets widespread.

5.2 Future Works

This thesis study has achieved its overall plan. However, there are a couple of things anyone can done in the future to further improve this data model. This data model initially has all its instances that is not related to race, religion and gender removed to balance the dataset. However, most of the social media posts in real word does not involved with race, religion or gender. Therefore, one future work for this thesis study should be adding back these unused instances back to the dataset. The accuracy of the hate speech detection system should not be affected by the presence of these instances. Meanwhile. more Malaysian social media posts should be extracted from internet and added into the hate speech dataset. As the size of dataset currently only has around 1350 posts, it is advised to search more posts to discover more contexts and hidden meanings inside them. Detection system will learn these contexts and detect hate speech more effectively.

Additionally, category and sub-category labels in this data model have not been utilized in the study due to time constraint. Therefore, it is advised to train algorithms with these aforementioned labels in the future. This can find relationships between each label and the keywords that are correlated with the label. The detection system will determine the likelihood of social media posts being hateful, not only based on its context, but also on the category values. This shall improve the accuracy of the system classifying hate speech.

REFERENCES

Teoh, S. (2023, June 1). *Malaysian polls in November saw surge in hate speech on social media: Study*. Retrieved on July 20, 2024, from The Straits Times. https://www.straitstimes.com/asia/se-asia/malaysian-polls-in-november-saw-surge-of-hate-speech-on-social-media-study

Choy, N. Y. (2023, November 8). *More than 3,700 posts involving fake news and hate speech taken down over Jan-Oct, says Teo.* Retrieved on July 20, 2024, from The Edge Malaysia. https://theedgemalaysia.com/node/689282

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Advances in Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Vincent, J. (2020, November 13). *Facebook is now using AI to sort content for quicker moderation.* Retrieved on July 20, 2024, from The Verge. https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation

Paul, K., & Dang, S. (2022, December 22). *Exclusive: Twitter leans on automation to moderate content as harmful speech surges.* Retrieved on July 20, 2024, from Reuters. https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in *Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Jourová, V., Wigand, C., & Voin, M. (2017). Speech by Commissioner Jourová - 10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_17_403

Merriam-Webster. (2023a). *Race*. In Merriam-Webster Dictionary. https://www.merriamwebster.com/dictionary/race

Cambridge University Press. (2023). *Religion*. In Cambridge Advanced Learner's Dictionary & Thesaurus. https://dictionary.cambridge.org/dictionary/english/religion

Merriam-Webster. (2023b). *Gender*. In Merriam-Webster Dictionary. https://www.merriamwebster.com/dictionary/gender

Houghton Mifflin Harcourt. (2011). *Physique*. In American Heritage Dictionary of the English
Language, Fifth Edition. https://www.thefreedictionary.com/physique

Eldridge, A. (2023, September 19). *Sexual Orientation*. Encyclopedia Britannica. https://www.britannica.com/topic/sexual-orientation

Centers for Disease Control and Prevention. (2020, September 16). *Disability and health overview*. Retrieved on July 20, 2024, from Disability and Health Promotion. https://www.cdc.gov/ncbddd/disabilityandhealth/disability.html

Bahador, B. (2020, November 17). *Classifying and identifying the intensity of hate speech*. Retrieved on July 20, 2024, from Items. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

USAFacts. (n.d.). *How has the racial and ethnic makeup of the US changed?* Retrieved on July 20, 2024, from Our Changing Population: United States. https://usafacts.org/data/topics/people-society/population-and-demographics/our-changing-population/

Hinton, E., & Cook, D. (2021). The mass criminalization of Black Americans: A historical overview. *Annual Review of Criminology*, *4*(1), 261–286. https://doi.org/10.1146/annurev-criminol-060520-033306

Susman, T. (2011, September 27). *Fugitive in hijacking case caught after 40-year hunt.* Retrieved on July 20, 2024, from Los Angeles Times. https://www.latimes.com/world/la-xpm-2011-sep-27-la-na-new-jersey-fugitive-20110928-story.html

Gado, M. (n.d.). *The Brinks robbery of 1981*. Retrieved on July 20, 2024, from Crime Library. https://www.crimelibrary.org/terrorists_spies/terrorists/brinks/1.html

Morales, F. R., Nguyen-Finn, K. L., Haidar, M., & Mercado, A. (2022). Humanitarian

crisis on the US–Mexico border: Mental health needs of refugees and asylum seekers. *Current Opinion in Psychology, 48*, 101452. https://doi.org/10.1016/j.copsyc.2022.101452

Galemba, R. B. (2021). "They steal our work": Wage theft and the criminalization of immigrant day laborers in Colorado, USA. *European Journal on Criminal Policy and Research, 27*(1), 91–112. https://doi.org/10.1007/s10610-020-09474-z

Ngwainmbi, E. K. (2022). Hate speech and the re-emergence of Caucasian nationalism in the United States. *Dismantling Cultural Borders Through Social Media and Digital Communications*, 73–104. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-92212-2_4

Haque, A., Tubbs, C. Y., Kahumoku-Fessler, E., & Brown, M. (2019). Microaggressions and islamophobia: Experiences of Muslims across the United States and clinical implications. *Journal of Marital and Family Therapy, 45*(1), 76–91. https://doi.org/10.1111/jmft.12339

Duvall, J. (2022, December 5). *New statistics on growing persecution against Christians*. Retrieved on July 20, 2024, from Liberty Champion. https://www.liberty.edu/champion/2022/12/new-statistics-on-growing-persecution-against-christians/

Price, M., Hollinsaid, N. L., McKetta, S., Mellen, E., & Rakhilin, M. (2023). Structural transphobia is associated with psychological distress and suicidality in a large national sample of transgender adults. *Social Psychiatry and Psychiatric Epidemiology.* https://doi.org/10.1007/s00127-023-02482-4

HRC Foundation. (2023, May 31). *Glossary of terms.* Retrieved on July 20, 2024, from Human Rights Campaign. https://www.hrc.org/resources/glossary-of-terms

Zambon, V., & Kuehnle, F. (2023, October 16). *What are the different types of sexualities?* Retrieved on July 20, 2024, from MedicalNewsToday. https://www.medicalnewstoday.com/articles/types-of-sexuality#types

Mahidin, M. D. (2020, July 15). *Current population estimates, Malaysia, 2020*. Retrieved on July 20, 2024, from Department of Statistics Malaysia. https://dosm.gov.my/portal-main/release-content/current-population-estimates-malaysia-2020

Sani, M. A. M. (2011). Cultural arguments against offensive speech in Malaysia: Debates between liberalism and Asian values on pornography and hate speech. *Mediterranean Journal of Social Sciences, 2*(2), 122-134.

Zamri, N. a. K., Nasir, N. a. M., Hassim, M. N., & Ramli, S. M. (2023). Digital hate speech and othering: The construction of hate speech from Malaysian perspectives. *Cogent Arts & Humanities, 10*(1). https://doi.org/10.1080/23311983.2023.2229089

Fernandez, K. (2020). Three waves of hate speech spreading faster than the pandemic in Malaysia: An analyses of outgroup populist narratives and hate speech during the COVID-19. *Geografia, 16*(4). https://doi.org/10.17576/geo-2020-1604-21

Agarwal, R. (2018, July 26). *Twitter hate speech*. Retrieved on July 20, 2024, from Kaggle. https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, *8*, 204951–204962. https://doi.org/10.1109/access.2020.3037073

Al-Hassan, A., & Al-Dossari, H. (2021). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, *28*(6), 1963–1974. https://doi.org/10.1007/s00530-020-00742-w

Muslim, F., Purwarianti, A., & Ruskanda, F. Z. (2021). Cost-Sensitive Learning and Ensemble BERT for Identifying and Categorizing Offensive Language in Social Media. *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. https://doi.org/10.1109/icaicta53211.2021.9640280

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1415–1420. https://doi.org/10.18653/v1/n19-1144

Jahan, S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Ariwibowo, S., Girsang, A. S., & Diana. (2022). Hate Speech Text Classification Using Long Short-Term Memory (LSTM). 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), 1–6. https://doi.org/10.1109/icosnikom56551.2022.10034908

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation, 55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Aditya, A., Vinod, R., Kumar, A., Bhowmik, I., & Swaminathan, J. (2022). Classifying Speech into Offensive and Hate Categories along with Targeted Communities using Machine Learning. *2022 International Conference on Inventive Computation Technologies (ICICT)*, 291-295. https://doi.org/10.1109/icict54344.2022.9850944

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. *2018 IEEE/ACM International*
Conference on Advances in Social Networks Analysis and Mining (ASONAM). 69-76. https://doi.org/10.1109/asonam.2018.8508247

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of NAACL-HLT 2016*, 88–93. https://aclanthology.org/N16-2013.pdf

Jerome, C., Ting, S.-H., Yeo, J. J.-Y., & Hsin, N. L. (2021). Examining discrepant views of LGBT and non-LGBT individuals on societal receptivity towards the LGBT phenomenon in present-day Malaysia. *International Journal of Social Science Research*, *3*(1), 55–66. https://myjms.mohe.gov.my/index.php/ijssr/article/view/12817

Green, R. (2023, July 7). *Broken Windows Theory: How Environment Impacts Behavior*. Retrieved on July 20, 2024, from Verywell Mind. https://www.verywellmind.com/broken-windows-theory-7550632

Distante, E. (2022, October 20). *BERTopic: Topic modeling as you have never seen it before.* Retrieved on July 20, 2024, from Medium. https://medium.com/data-reply-it-datatech/bertopic-topic-modeling-as-you-have-never-seen-it-before-abb48bbab2b2

Grootendorst, M. (2022, March 11). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv.org. https://arxiv.org/abs/2203.05794

Kulshrestha, R. (2019a, November 25). *NLP 101: Word2VEC — Skip-Gram and CBOW*. Retrieved on July 20, 2024, from Medium. https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

fastText. (n.d.). *Text classification*. https://fasttext.cc/docs/en/supervised-tutorial.html

FastText working and implementation. (2024, May 24). Retrieved on July 20, 2024, from GeeksforGeeks. https://www.geeksforgeeks.org/fasttext-working-and-implementation/

Kulshrestha, R. (2019b, July 20). *A Beginner's Guide to Latent Dirichlet Allocation(LDA).* Retrieved on July 20, 2024, from Medium. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Prabhakaran, S. (2022, April 20). *How Naive Bayes algorithm works? (With example and full code)*. Machine Learning Plus. https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/

Prabhakaran, S. (2022, April 20). *How Naive Bayes algorithm works? (with example and full code)*. Retrieved on July 20, 2024, from Machine Learning Plus. https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/

Oppermann, A. (2023, December 12). *What is deep learning and how does it work?* Retrieved on July 20, 2024, from Built In. https://builtin.com/machine-learning/deep-learning

Moolayil, J. J. (2021, December 14). *A Layman's Guide to Deep Convolutional Neural networks.* Retrieved on July 20, 2024, from Medium. https://towardsdatascience.com/a-laymans-guide-to-deep-convolutional-neural-networks-7e937628605f

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

J., R. T. J. (2020, September 2). *LSTMS explained: A complete, technically accurate, conceptual guide with Keras.* Retrieved on July 20, 2024, from Medium.https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2

Kostadinov, S. (2017, December 16). *Understanding GRU networks.* Retrieved on July 20, 2024, from Medium. https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

Grootendorst, M. P. (n.d.). *Topic distributions*. BERTopic. Retrieved on July 20, 2024, from https://maartengr.github.io/BERTopic/getting_started/distribution/distribution.html

Vishwanath, Y. (2023, March 14). *BerTopic Modelling - Advanced Topic Modelling.* Retrieved on July 20, 2024, from Medium. https://medium.com/digital-engineering-centific/bertopic-modelling-advanced-topic-modelling-73af7697b7f3

Gomede, E. (2023, December 30). *Understanding Long Short-Term Memory (LSTM) Networks: A journey through time and memory.* Retrieved on July 20, 2024, from Medium. https://medium.com/the-modern-scientist/understanding-long-short-term-memory-lstm-networks-a-journey-through-time-and-memory-57e3c947fc63