DETECTING CYBERBULLYING IN SOCIAL MEDIA TEXT USING NATURAL LANGUAGE PROCESSING

MUHAMMAD FIRDAUS AIMAN BIN MOHD YASIN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: [DETECTING CYBERBULLYING ON SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING]

SESI PENGAJIAN: 2022/2023

Saya: MUHAMMAD FIRDAUS AIMAN BIN MOHD YASIN

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
____✓____	TIDAK TERHAD	

(TANDATANGAN PELAJAR)

Alamat tetap: Lorong Sidang Hassan 2,
Jalan Sidang Hasan ,Bukit Katil 75450
Melaka

Tarikh:20/9/2023

(TANDATANGAN PENYELIA)

Dr. Nur Zareen Zulkarnain

Tarikh:20/9/2023

DETECTING CYBERBULLYING IN SOCIAL MEDIA TEXT USING NATURAL LANGUAGE PROCESSING

MUHAMMAD FIRDAUS AIMAN BIN MOHD YASIN

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Artificial Intelligence) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I hereby declare that this project report entitled

DETECTING CYBERBULLYING ON SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : _____ Date : 20/9/2023
(MUHAMMAD FIRDAUS AIMAN BIN MOHD YASIN)

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Artificial Intelligence) with Honours.

SUPERVISOR : _____ Date : 20/9/2023
(DR. NUR ZAREEN ZULKARNAIN)

DEDICATION

Thank to Allah,that I have a ability and enough motivation to go through this project and complete the report. With this report for my Final Year Project, I hope every people who read this will know what are the activity that have been done during 14 weeks.

I would like to give appreciation to parents and also to my beloved family for give me support from beginning and keep believe in me.I also would like to thank to all my friends in Univeristy Teknikal Malaysia Melaka including my supervisor of this project for helping me a lot and share all the knowledge needed.

Hope this report could give a benefit and idea to all people who read it. Maybe next time it could be meaningful knowledge in the future.

ACKNOWLEDGEMENTS

With all my might I praise to Allah,for given me this opportunity. As I am given a chance to put an end my Final Year Project for my Bachelor Computer Science (Artificial Intelligence) with honours as student successfully. I am thankful that I am in good health and have ability to complete all process in this project very well.

Thank you to my supervisor from Faculty of Information and Communication Technology, Dr. Nur Zareen Zulkarnain for helping me on the project that need to be develop. Next, I want to thank to my family and my parents for giving me support and motivation to go through this project.Also I want to appreciate al lecturer from my faculty that also give me idea and support along this journey.

I hope that this report could give a guide and knowledge to people that read this and all activity that I have done during this 14 weeks.

ABSTRACT

The goal of this project is to find the method that can solve on detecting cyberbullying on social media using natural language processing. Because the selected language is Malay hence there are a lot challenges to overcome. To get the classify the cyberbullying text and determine the best model, we will use Support Vector Machine, Naïve Bayes, Long Short Term Memory, Convolution Neural Network and Bidirectional Encoder Representation from Transformer. This project will make use of 1383 tweet data. The data is separated into two part: training, and testing. The model's results will evaluate using the confusion matrix such as accuracy, precision, recall and f1 score. In addition the model with highest accuracy can be selected to use for detecting the cyberbullying tweet. Last but not least the model deployment phase will be deploy on more interactive interface or graphical representation using Streamlit in order to test the best model produce in detecting cyberbullying tweet.

ABSTRAK

Matlamat project ini adalah untuk mencari kaedah yang sesuai bagi mengesan jenis-jenis tweet yang mengandungi unsur-unsur jenayah siber di media social menggunakan *Natural Language Processing.* Disebabkan Bahasa yang dipilih di dalam projek ini adalah Bahasa melayu maka terdapat banyak faktor dan kekangan yang menggangu dalam kajian projek ini. Bagi menentukan model yang terbaik untuk mengenalpasti teks yang mengandungi jenayah siber, kami akan menggunakan *Support Vector Machine, Naïve Bayes, Long Short Term Memory , Convolution Neural Network dan Bidirectional Encoder Representation from Transformer.* Dalam projek ini, sebanyak 1383 tweet digunakan. Data yang dikumpul akan di bahagikan kepada dua bahagian iaitu *training* dan *testing.* Keputusan model yang di hasilkan akan di nilai berdasarkan *confusion matrix* seperti *accuracy*, *precision, recall*, dan *f1 score*. Selanjutnya, model yang mempunyai ketepatan yang paling tinggi akan dipilih untuk mengenalpasti tweet yang mengandungi unsur jenayah siber. Akhir sekali, adalah fasa untuk membina sebuah laman web yang mengguna pakai model yang telah dibina untuk tujuan menguji dan penggunaan dalam mengenalpasti tweet yang mengandungi unsur jenayah siber.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

FYP	**Final Year Project**
TF-IDF	**Term Frequency-Inverse Document Frequency**
LSTM	**Long Short Term Memory**
CNN	**Convolution Neural Network**
BERT	**Bidirectional Encoder Representation from Transformers**
TP	**True Positive**
FP	**False Positive**
TN	**True Negative**
FN	**False Negative**
BOW	**Bag of Word**
ML	**Machine Learning**
DL	**Deep Learning**
LLM	**Large Language Model**
AI	**Artificial Intelligence**

CHAPTER 1: INTRODUCTION

1.1 Introduction

Cyberbullying is refer to bully using digital technologies where it can be in various platform such as social media, communication platforms and many more. According to UNICEF cyberbullying is repeated behavior as the purpose is to scare, or shaming certain people and furthermore in 2020, Malaysia was ranked second for cyberbullying among youth (UNICEF, 2023). According to another studies (Malaysian Institute for Youth Development Research) in 2013, cyberbullying is increase by 55.6% compared to previous year (Arsad. M, 2020). In addition according to UReport in Selangor the cyberbullying most occurs on social media around 82% (UReport Malaysia, 2019). As this indicates that cyberbullying is a serious problem in our country. The objectives is to use Natrual Language Processing approach to analyze cyberbullying on social media.Using sentiment analysis is to identify every tweets whether the tweet is negative, neutral, or positive where usually cyberbullying tweets contain negative words. By using sentiment analysis it helps to show the expression of the text and recognize the patterns and trends in the language used by cyberbullies. Furthermore in order to get more better result of identify cyberbullying text is to identify mentions included in the text. This mention can sometimes help to identify cyberbullying but it does not include in every text. Last but not least the expected outcomes is to detect cyberbullying at the early stage by filters the identifying harmful content.

1.2 Problem Statement

As mention in the previous section, In 2020 Malaysia was reported as second ranks in Asia for cyberbullying among the youths(UNICEF, 2020). This indicates that cyberbullying is become more serious problem in our country. Social media is one of popular platform for this crime to be likely happen as it is not only among the youth but including adults as well. According to Aisyah (2022), due to the ease of access to information and communication technologies, cyberbullying has emerged as a new means for young people to vent their discontent, frequently violently. Meanwhile according to the Malaysian Institute for Youth Development Research (2017), there were 389 incidents of cyberbullying with an average victimisation rate of 62.3% among teens and a control rate of 37.7%. Mostly cyberbullying contain a negative word or something that is intended to harm certain people or organization. Cyberbullying is different from negative comment classification classification or tweet as there are many aspect that differentiate between these two category. In order to classify cyberbullying text is to know what are the characteristic and pattern of the word used. With machine learning approach, it is easier to identify the pattern of those text as it can discover the pattern contain in the word and predict the next sentences whether it falls under cyberbullying or not. In current situation, there is not much of tool that can help to prevent the cyberbullying tweet in malay language as building this kind of tool could decrease the amount of cyberbullying tweet on social media.

1.3 Objectives

1. To investigate related words that lead to cyberbullying.

2. To evaluate Machine Learning model using confusion matrix.

3. To classify which is the best Machine Learning algorithm to determine cyberbullying

1.4 Project Scope

The scope of this project is focus on Malay language and tweets in Malaysia only. For the dataset is collected from Twitter and GitHub as these platform are easier to gain and access.

1.5 Conclusion

As a conclusion Chapter 1 sums up what are the title is about and what are the problem statement with the objective that need to be achieve at the end of this project including the scope of the project. This section also tells what are the requirement specification for this project such as hardware and software.

CHAPTER 2: LITERATURE REVIEW AND PROJECT METHODOLOGY

2.1 Introduction

Literature review is a crucial step in any research or studies where it helps to get a general idea what have been done, what can be improve, where is the solution takes place for solving existing problem, and idea for future works. Literature review also includes some techniques or methodology that can be implemented on similar problem. Since this project title is about detect cyberbullying on social media using Natural Language Processing the similar research paper based on the title will be used in the project. The project methodology will be slightly different from most of the research paper since the selected language and scope for this project focus on Malay Language only. The general idea is to use the Machine Learning methodology which includes six important phase.

2.2 Facts and findings

According to Raj (2021) machine learning approach is one of the method that can help for detecting online cyberbullying but there are some drawbacks which was the accuracy of the model are low because due of limitation of the dataset for supervised classification. However with the help of neural network it can provide support and solution for the limitation. The researcher discuss the proposed methodology of classification framework using Wikipedia Attack dataset and Wikipedia Web Toxicity dataset. The early phase in machine learning methodology is the data preprocessing and cleaning as the author use few techniques which is remove empty rows, punctuation, special characters, numerical values, stopwords, lowercasing

text, tokenization and lastly stemming. This is an important step to make sure the raw data or input are cleaned as text classification could not retrieve raw data like a human understanding. According to Dewani (2023), data preprocessing is about to convert the low quality unstructured data into structured data as this is essential for detect a pattern. The reason is because text data are usually written in free form and full of abbreviation as this makes the data is not standardize hence inappropriate for process in the next phase. The author state that in this research , the data use is in Urdu language. The author begin with similar data preprocessing method which is remove punctuation, special characters, numeric , removal hashtag, extra whitespace, and convert to lowercase. Lowercase help to narrow down from huge dimensional space.Next, tokenization where each sentences will be split into single unit which is word as this makes each text more manageable. The process continue by remove stop word which was the most occur word in sentence. Usually stopword are removed since stop word does not contain any value. Lastly, the author proceed by expanding contraction using Urdu language. The method to execute this is by mapping the contraction word in Urdu with their original word.

According to Rajpara (2022), data cleaning and preprocessing is important to eliminate the unwanted and unnecessary data as this help for the next steps. The element that usually affect the performance of model is noise that contain in the data such as punctuation, special character, numeric value, stopword , and contraction.

Therefore after the second phase is finish, the next step is to convert the cleaned data into numerical value as this help machine to understand the data the way human understand. There are several techniques to use for transforming the text, which is Counter Vectorization where it is a straightforward statistical technique for producing embedded vectors of input text. According to Raj (2021) using the frequency of the occurrence of a term in document help to generate its embedding vector. For the full collection of documents, a matrix is formed, with each document's rows corresponding to its columns as words. The values of a term's frequency of occurrence in a document are contained in the cells. As this matrix help for training phase for machine learning approach. Eventhough the Counter Vectorization is simple and straightforward statistical method it was not able to identify the important and less important of word including identify the relationship between one word to another word as usually the

word that occurs many time will overshadow the less word that occurs. However, there are other method which Term Frequency Inverse Document Frequency(TF-IDF) where it focus on capture the high value of word contain in the corpus. TF-IDF is a opposite method than Counter Vectorizer, which mean that the most occur word have the low value as this help to identify the rare word that appear in the corpus. Gada (2023) employed TF-IDF on basic classification algorithms as a first step to obtain the baseline model. TF-IDF can be use in various way which mean that it can combine with N-gram and Char which can give different approach toward getting a solution for any text classification. According to Cheng (2019) TF-IDF is a simple and proven method in text classification.

For more powerful feature engineering, that can relate previous word to next word is to use word embedding, as this is an advanced feature engineering. Which can capture the semantic of text as this makes help to understand the text and easier for text classification. There several word embedding example such as GloVe, FastText, Paragram as this is common word embedding use for text classification. According to Pennington (2014) ,GloVe is a method for obtaining word embeddings from text input that is unsupervised. The utlizie an co-occurrence matrix to gain representation help to identify the semantic relationship between term. GloVe is very powerful word vector as it group all similar word in same cluster. In order to create word vectors, GloVe incorporates global statistics (word co-occurrence) in addition to local statistics (local context information of words). Another word encoding technique intended to better capture contextual similarities is the paragram. It applies attract-repel, counter-fitting, and fine-tuning techniques to the ParaPhrase DataBase (PPDB) in order to inject synonym and antonym features as a vectorization constraint. According to Raj (2021), Paragram is comparably strong because of a superior knowledge of context. Next, FastText is also another word embedding example where it was introduced by Facebook’s AI Research Lab (FAIR), FastText is a skip gram based model to make word representation better Mikolov (2013). This method's efficacy is due to the fact that it takes a language's word morphology into account. Some embedding methods represent each word as a separate vector. FastText are able to understand the context

of unknown words by splitting the text into smaller part and match the similarity of word in the corpus.

Furthermore after the data are cleaned, and already in numerical representation either using traditional feature engineering or advanced feature engineering, we are ready to train our model to solve our problem. There are several way or method to implement and the most known method for text classification is using traditional machine learning approach. Inside machine learning approach there are many algorithm that can help to easily classify the text, such as Naïve Bayes and Support Vector Machine (SVM). Naïve Bayes use probabilistic to predict the next word by applying one of famous theorem which is Bayes Theorem. As stated by Sulzman (2007) Naïve Bayes classifier is widely used in many application as it was effective that can reducing the computational costs. Naïve Bayes also suitable in handling large dataset and still able to produce high accuracy. Meanwhile, Support Vector Machine is a discriminative classifier that formally define by separating hyperlane. SVM use the separation margin between data point of classes. There will be a plane usually depends on the dimension of the data, the plane will get differentiate between two classes. Support vector are very crucial as this will determine the hyperplane as the target is to make sure that the margin of hyperplane is maximize. According to Sarkar (2015) SVM is a supervised algorithm that classifies data points by the distance between classes separation margins. The use of machine learning approach are good when you have an domain expert to identify the important feature but the deep learning comes in hand to help solve problem end to end. Deep learning example is neural network approach that can help in solving text classification such as Convolution Neural Network (CNN), Long Short Term Memory (LSTM), Recurrent Neural Network (RNN). A neural network type called a convolutional neural network, or CNN or ConvNet, is particularly adept at processing input with a grid-like architecture, like an image. A binary representation of visual data is a digital image. It has a number of pixels that are arranged like a grid and are each assigned a value to indicate how bright and what colour each pixel should be (Dumane, 2021). Long Short Term Memory (LSTM) is a type of Recurrent Neural Network which was intended to handle sequential data such as speech and text. LSTM have large capacity to hold sequential data compare to Recurrent Neural Network which makes it suitable for task such as

language translation and speech recognition. As state by (Raj, 2021) LSTM can solve problem of vanishing gradient descent which usually face in RNN.

In addition, we could also use pretrained model such as Bidirectional Encoder Representations from Transformers (BERT) for text classification.BERT is published by researchers at Google AI Language. Transformer, an attention mechanism used by BERT, is used to understand the contextual relationships between words (or subwords) in a text (Horev, 2018). According to Andrade-Segarra,(2021) BERT produce an efficient prediction accuracy but consumes a lot of system resources to run.

2.2.1 Technique

1.Support Vector Machine

Support Vector Machine is one of classification algorithm, where it is a discriminative classifier formally defined by separating hyperplane. The purpose of hyperplane is to separate between classes and each hyperplane is different depends on the dimension space use. The support vector machine technique seeks to locate an N-dimensional space hyperplane that clearly categorises the data points (Gandhi, 2022).

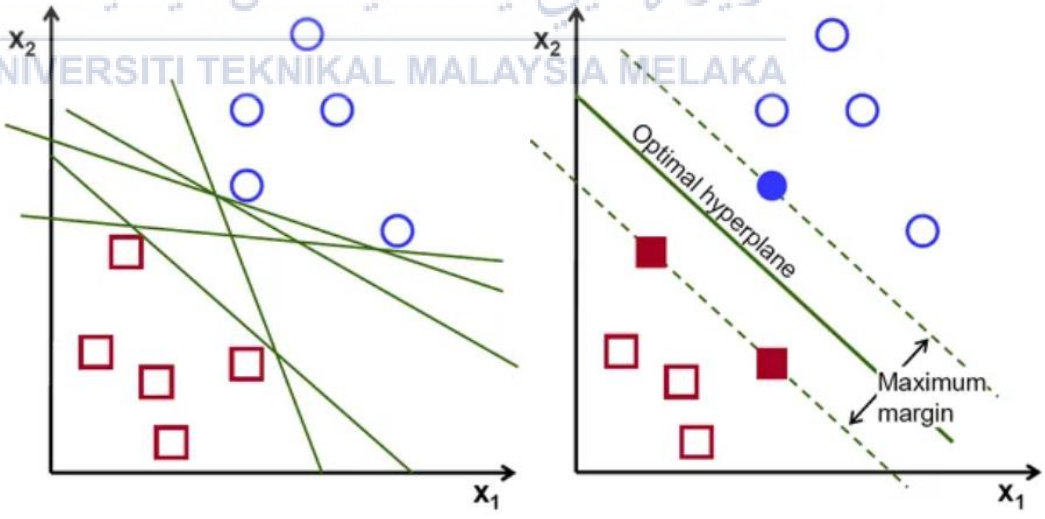

Figure 1: Optimal Hyperplane (Gandhi, 2022)

Figure 1 shows that there are many possible hyperplane that can be construct. However the objective is to find the optimal hyperplane that have maximum margin