Journal of Positive School Psychology
2022, Vol. 6, No. 3, 8548–8560

Detecting Offensive Malay Language Comments on YouTube using Support Vector Machine (SVM) and Naive Bayes (NB) Model

Ameliyana Mohd Isa[1]; Suzana Ahmad[2]; Norizan Mat Diah[3]

[1,2,3] *Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,40450 Shah Alam, Selangor, Malaysia*

[1] *ameliyanamohd@gmail.com,*[2] *suzana@fskm.uitm.edu.my,* [3]*norizan@fskm.uitm.edu.my*

Abstract

Social media, such as YouTube, Twitter, and Facebook, have become a new way of communication allowing many users to interact and obtain information. Nowadays, many users on social media write and post using offensive language. Offensive language is an expression consisting of offensive words, either oral or text, including abusive, racial, and sexual content, and it can be in multiple languages. Offensive language may jeopardize user engagement. Users can manually control the offensive language; however, the colossal amount of unstructured data is challenging. Thus, this study addresses the issue by identifying the offensive words used in YouTube comments, focusing on the Malay language, based on the list of offensive words obtained from the Malaysian Communications and Multimedia Commission (MCMC). This study also builds an experiment for offensive YouTube comments detection using Term Frequency - Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) features. This study employed the Random undersampling and Random oversampling techniques to treat the imbalanced data. Support Vector Machine (SVM) and Naïve Bayes (NB) were used to identify whether the comment is offensive. The results showed that the SVM model and TF-IDF, as a weighting feature, are the best approach for this study, with Recall results of 98.70%. Both models are effective in this study, with NB produced slightly lower results than SVM. Results can improve by further data preprocessing and adjustment of the classifiers

Keywords: Classification Model, Language Detection, Offensive Malay Language, Random Undersampling Technique

I. INTRODUCTION

According to Digital 2020 Malaysia, out of 32.16 million of the total population, 26 million Malaysians are active on social media. Daily, users spend 2 hours and 45 minutes on average using social media. As the number of participants on social media increases, the risk of being exposed to offensive content on social media also increases (Yazdanifard et al., 2012). Apart from that, offensive words or inappropriate phrases also intensify on social media platforms, such as Twitter, Facebook, Instagram, and YouTube. YouTube has become the most popular video streaming application in today's era (Maadi et al., 2016). The number of YouTube users has reached more than 2 billion in over 100 countries, and almost 5 billion videos had been watched on YouTube every day(Maadi et al., 2016).

Comments containing offensive words may harm content creators and other users reading them (Yazdanifard et al., 2012). According to Fortuna et al., 2021, offensive language can be classified into nine categories: Sexual, Religion, Class, Ethnicity, Gender, Physical, Race, Disability, Behaviour and others. Since users of all ages can access and watch YouTube, some users might get offended by these offensive comments (Yazdanifard et al., 2012). Offensive language also includes sexual and profane words that people should avoid using in society since it is harmful and offensive (Dul et al., 1996). This phenomenon has been widely

observed in social media, and it would also affect the accessibility of other users while using these platforms.

Nowadays, text analytics is an active area of research to discover knowledge from text. Some of the techniques used are natural language processing (NLP), data mining (DM), and machine learning (ML) techniques (Singh et al., 2016). The basic framework for text analytics requires data acquisition, preprocessing, representation, and knowledge discovery techniques (Wan et al,. 2016). However, there is a lack of study on detecting offensive comments on YouTube focusing on the Malay language to the researchers' knowledge.

Therefore, this study addresses this issue by developing a classification model to detect offensive comments in the Malay language on YouTube.

II. THEORETICAL BACKGROUND

The Malay language is the National Language of Malaysia originating from the Austronesian family (Bianco et al., 2010). It has been widely used by most Malaysian and public education systems (Bianco et al., 2010). According to Khalifa, Ahmad, and Gunawan (Patel et al., 2011), Dewan Bahasa dan Pustaka (DBP) officially standardises the spelling and language structure of the formal Malay language in Malaysia. However, the social media evolution makes the language a whole different story. The language usages in social media appear to deviate from the standard language usages (e.g., emoticons, short forms, undefined terms, broken grammars, and incomplete sentence) (Purnama 2015; Ismail et al., 2011). In addition, the limit put on the character length in social media, such as Twitter (140 characters per tweet), causes users to minimise the characters to fit into the space provided (Hermandez et al., 2017). The judgement of the sentence with character minimisation depends on the commenter's understanding, and the interpretation of these messages depends on the reader's perspective (Singh et al., 2016). These constraints make it challenging for this study to apply the formal Malay language rules.

According to (Singh et al., 2016), there are seven types of Malay social media text: spelling style variations, mixed sentence referring to the mixture of language used, English words spelt using Malay phonology, Malaysian region-based slang, no-vowel spelling or short forms, numerical suffixes representing words that are duplicated, and conveying expressions indicating the multiple sequences of character in the word. In Malaysia, the Malay language in the social media context is somewhat complex and varies due to the mixed culture and race that this country has (Singh et al., 2016).

A. Offensive words in the Malay Language

Offensive words in the Malay language are defined based on the list of prohibited words in the media context, obtained from the Malaysian Communications and Multimedia Commission (MCMC) website. MCMC is the regulator for the converging communications and multimedia industry in Malaysia. The list of words was manifested in October 2017 with the cooperation of DBP, Film Censorship Board of Malaysia (LPF), Communications and Multimedia Content Forum of Malaysia (CMCF), and Commercial Radio Malaysia (CRM). The usage of these prohibited words is highly forbidden in the media context as it will degenerate the quality of language in both spoken and written forms. Prohibited language is also defined as rough language, abusive language, and obscene language, indicating impolite language, language used for speech or painful writing, and language with bad, abusive, or vile words. In general, these languages are also considered offensive. Based on MCMC, prohibited language is classified into seven categories consisting of 43 words in total. The examples of words according to the categories are listed in Table 1.

Table 1 - Examples of prohibited words based on categories

Category	Example of words
Rude title refering to individual (leader/special individual)	"Mat Mongol", "Mat Sembab", Si Gagap", "Mat Komedi", "Si Puaka", "Idi Amin Malaysia", "Maha Firaun", "Mat Sabun", "Mak Lampir"
Rude title referring to institution or organisation	"DAPig", "fuckatan", "Pascai", "Umngok", "Umporno", "TV tiga suku", "Utusex", "Parti Kencing Rakyat", "PANDap"
Equens human with animals or bad creatures	"babi", "anjing", "lembu", "hantu", "syaitan", "iblis"
Human body parts	"pukimak", "butuh", "tahi"
Human behaviour	"bodoh", "dungu", "celaka", "haram jadah", "betina", "sundal", "mampus", "tonggong"
Equens human with objects	"bangkai", "dedak", "kayu", "hampas"
Religion and racism	"kafir", "laknat", "tongsan", "laknatullah", "murtad

Based on the study by Purnama (2015), the authors treated messages containing rude, figurative, offensive, and dirty words as abusive. Meanwhile, Ibrohim and Budi (Ibrohim et al., 2018) explained that words expressing unpleasant conditions in a conversation are categorised as offensive words. Therefore, in this study, the researchers refer to the MCMC list of prohibited words as the reference for offensive language.

B. Offensive Language in Social Media

Nowadays, the usage of offensive words in social media has been emerging as never before. Many existing studies in the broader literature have examined the use of offensive words in social media. According to Goel and Sharma (2019), inappropriate words, especially those extremely rude or offensive, are considered online harassment. Meanwhile, in the USA, it is determined as an online threat. The authors also mentioned several reported cases due to abusive comments on social media. Moreover, social media benefits the content creators and viewers, but there is always a risk of being exposed to inappropriate content (Fang et al., 2014). Kaushal, Saha, Bajaj, and Kumaraguru (Majid et al., 2015) stated that unsafe content on the YouTube platform affects the video content and the comment section.

C. Related Work on Offensive Language in Social Media

Previous researchers have conducted numerous studies to detect abusive language in social media using various approaches. For example, Purnama (2015) explored abusive language detection in the Thai language. They extracted Facebook posts and comments as a dataset. Their study used several classifiers, such as SVM, NB, k-Nearest Neighbour (kNN), Random Forest Decision Tree (RFDT), and others with several weighting features, such as word n-grams and TF-IDF. The experimental results showed that Discriminative Multinomial Naive Bayes produce the best results compared to other classifiers. Furthermore, Ibrohim and Budi (2018) discussed abusive language detection in the Indonesian language using Twitter as a dataset. They used three classifiers: NB, SVM, and RFDT, with term weighting features, such as simple word n-gram and character n-gram. The authors classified abusive Indonesian language into eight categories: condition, animals, astral beings, object, part of the body, family member, activity, and profession. According to the authors, words expressing unpleasant condition and bad characteristics are considered offensive. The results from this experiment depicted that NB performs better than SVM and RFDT for classifying abusive language.

The study conducted by Castano (2021), aimed to identify racist social media comments in the Sinhala language. They used Facebook comments as their dataset. The study used binary classes in which the comments were labelled as "racist" or "non-racist". A Two-Class Support Vector Machine was used as a classifier and n-grams features as a weighting feature. The accuracy showed a promising result indicating that the classifier was relevant

for the study. Apart from that, Goel and Sharma (2019) focused on detecting insulting comments on a Twitter dataset. They combined NLP with ML techniques in their studies. The weighting features used were count vectors, TF-IDF vectors, and n-gram sequencing, while the classifier used were Logistic Regression (LR), SVM, Random Forest (RF), and Gradient Boost (GB). The results showed that LR and RF produced higher accuracy compared to the other two classifiers.

On the other hand, Jha et al., (2020) studied the tweets classification based on different categories. They classified tweets into the categories of sexism: ' Hostile', 'Benevolent', or 'Others'. They used SVM, sequence-to-sequence model, and FastText as a classifier, and the results showed that the FasrText classifier performed better than the SVM and sequence-to-sequence model. Chen et al., (2012) designed a prototype to detect the offensive language in social media using Lexical Syntactic Feature (LSF) architecture. They obtained the data from YouTube comments. LSF was compared with other classifiers, SVM and NB, resulting in LSF performing better than the other two classifiers.

Previous researchers, Alfina, Mulia, Fanany, and Ekanata (2017) conducted a study on hate speech detection in the Indonesian Language. They used Indonesian tweets as a dataset and labelled them into two categories. The study used four classification models: NB, SVM, Bayesian Logistic Regression (BLR), and RFDT. The results showed that RFDT produced better results compared to the other models.

Vandersmissen (2012) used a data corpus extracted from the Dutch distribution of Netlog's social media platform. The study selected SVM and NB as classifiers to detect offensive language. The results showed that SVM achieved reasonable precision, as opposed to NB. Furthermore, Nobata, Tetreault, Thomas, Mehdad, and Chang (2016) differentiated abusive comments in Yahoo! Finance and News. They labelled the data as clean or abusive. The authors implemented Vowpal Wabbit's regression model to classify comments.

Park and Fung (2017) classified comments into two categories, sexism and racism, using the Twitter dataset. The authors used three different Convolutional Neural Network (CNN): CharCNN, WordCNN, and Hybrid CNN. The results showed that HybridCNN produced a higher score than CharCNN and WordCNN. A study by Anagnostou et al. (2018) classified YouTube comments into different categories, comments rating on various areas and according to those areas' classification was done. Then, the Precision-Recall curve was constructed based on categories that used SVM and NB methods. The results for this experiment showed that produced a good F1-score.

Based on this review, it has been noted that SVM and NB classifiers are applied and tested on many datasets to detect abusive/offensive/negative comments on social media. In the past experiments testing on different datasets, these two classifiers produced reasonable accuracy, and several studies showed that they performed better than other classifiers. Hence for this research, the researchers employed SVM and NB as a classifier.

III. CONSTRUCTION METHOD

A. Data Extraction

This phase focuses on extracting comments from selected videos from YouTube. YouTube is chosen for this study because it has more significant communities all over the world. According to Dinakar, Reichart, and Lieberman (2011), rude comments usually appear on videos related to controversial topics. Therefore, this study collected the data from 60 controversial YouTube videos posted by Malaysian that are highly likely to have rude comments, as shown in Figure 1.

Figure 1: Sample of the dataset

B. Preprocessing Technique

All collected data in this study will undergo a preprocessing process. The first process of preprocessing is translation. Since social media language involves a mixture of language, the foreign words will be translated into Malay language using Google Translate. The following preprocessing step is de-noising, the process of removing all the punctuations from the data (DeLone et al., 1992). The punctuations include commas, question marks, exclamation marks, emoticons, and other symbols. Besides, numbers will be removed from the text. Then, all sentences will be converted into lowercase. Words containing an unimportant sequence of characters will also be manually identified and removed. For example, "seksiiiiiiiiiiiiii" will be converted into "seksi". Apart from that, common words in short forms will also be manually identified and transformed into the original word; for example, "xsuka" is converted to "tidak suka".

The following process is the removal of stopwords. Stopwords describe frequently used words in a sentence but do not have meaning in the sentence (DeLone et al., 2003). A machine does not understand the sentence meaning; it matches the patterns for text analysis. Therefore, the stopwords, such as 'a', 'an', 'the', 'some' and 'many more', can be considered words and removed (DeLone et al., 2003) . Since this study analyses the Malay language text, the researchers will be using the "stopwords" library from (Lin 2007), consisting of a stopwords collection in the Malay language. A few examples of stopwords in the Malay language are "di", "ialah", "lagi", and "ada". Subsequently, the tokenisation process will take place. Tokenisation is the process of splitting or breaking text into tokens by white space. This process will be conducted using the "Malaya" library package. For example, the word "saya suka tidur" will be tokenised into "saya", "suka", and "tidur".

The final process in the preprocessing workflow is the stemming process. Stemming is the process of reducing words into their root (DeLone et al., 1992). For example, "terduduk" can be reduced into "duduk", while "larian" and "berlari" will be reduced to their root word, which is "lari". Upon the completion of all these preprocessing techniques, the data is ready for the following process. The original or raw dataset has a total of 191368 words. Table 2 shows the number of words after the preprocessing procedure. The number of words decreases to 189967 upon removing the noisy data. Then, the total number of words increases to 190561after the word transformation process. Lastly, removing the stopwords decreases the

number of words to 94507. The table shows the summary of the total number of words for every preprocessing step.

Table 2 - Summary of the total number of words for every preprocessing step

Steps	Total Number of Words
Original	191368
Removal of noisy data	189967
Word transformation	190561
Removal of stopwords	94507

- ***Data Labelling***

Based on the data obtained, 1346 comments had been collected and labelled. Comments containing at least one identified Malay offensive words are labelled as '1' indicating 'offensive comment', and comments without identified Malay offensive words are labelled as '0'. The number of labelled comments is shown in Figure 2.

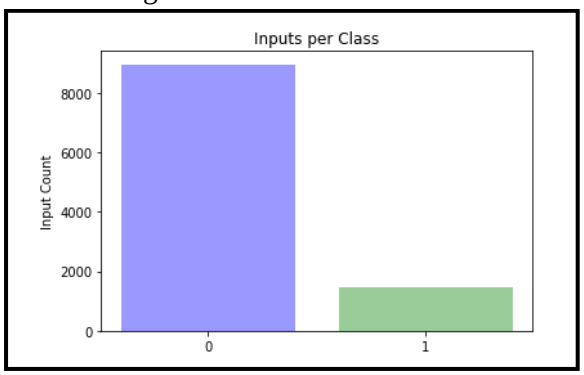

Figure 2: The number of labelled comments

A string data type is the predictor variable, while a Class is the target variable represented by integer data type. The study's findings demonstrated 1346 records of Class "1" and 9045 records of Class "0".

- ***Data Splitting***

After data preparation, the datasets are divided into two datasets, the training set and the testing set. The training dataset is used to fit the model, and the predictions are performed using the testing dataset. In this study, five-fold cross-validation with stratified sampling is used. The dataset is divided into five parts in which 8/10 will become the training data, and the remaining 2/10 will become the testing data. The training-testing process using this technique will be carried out five times to become both training and testing data simultaneously. In this study, 8346 comments are used as a training set, while 2086 comments are used as a testing set.

- ***Data Balancing***

It is considered a major problem for the data used, as the target variable is heavily imbalanced, and there are much fewer records of Class 1. This study compares multiple techniques to see which technique performs well for the model and increases the model performance to solve this problem. RandomOverSampling and RandomUnderSampling are the techniques used to balance the training data for better learning for the model. Oversampling and undersampling can lead to an increase or decrease in performance. These data balancing techniques help the model learns correctly as each class's data is passed with equal frequency, leading to no bias in learning the data. The majority class is randomly eliminated for undersampling. On the other hand, oversampling causes the comments to replicate randomly until both classes have the same number of comments.

C. Feature Selection

This study applies two feature selections: TF-IDF and BoW. TF-IDF is used as a weighing feature since it can capture word importance. TF represents a word frequency score in the current document, while IDF represents a score of how rare the word is across documents. The TF-IDF formula is as follows:

t f idf (w) = f (t, d) x log N/(|{dєD: tєd}|)

f(t, d) refers to the term frequency (TF), in which the number of times a word (w) appears in a document is divided by the total number of words in the document (d). Meanwhile, the right-hand side of the formula indicates the

calculation of IDF. First, the number of documents is divided by the number of documents containing the word w (Hassenzahl 2013). Then, TF will be multiplied by IDF to obtain the weight. Meanwhile, BoW is the most common vector space representational model for unstructured text (Hassenzahl 2013). A vector space model is simply a mathematical model representing unstructured text (or any other data) as numeric vectors. Each dimension of the vector is a specific feature/attribute (Hassenzahl 2013). It is called "Bag of Words" because each document represents a bag of its own words, disregarding word order, sequences, and grammar. The value frequency of the word in the document and each of the words corresponds to a feature (Al-Kilidar et al., 2005).

The text data is then converted into numerical feature vectors using TF-IDF and BoW. TF-IDF builds a vocabulary of words, which it has learned from the dataset, and it will assign a unique integer number to each of these words. The training dataset and testing dataset are then transformed to TF-IDF vectorised sets. Figure 3 shows an example of the TF-IDF vectorised data.

```
(0, 2467)     0.2555431839843941
(0, 2153)     0.1614756438485371
(0, 2040)     0.20541915470673686
(0, 1473)     0.18231239867589472
(0, 1444)     0.20245083253285093
(0, 1120)     0.23395593957832267
(0, 675)      0.22132821110686837
(0, 574)      0.24291545551293978
(0, 94)       0.22700639911280823
(1, 4943)     0.5127429306471287
(1, 2046)     0.28778409228395774
(1, 1487)     0.5323788008275363
```

Figure 3 - Example of TF-IDF vectorised data

The output for (0, 2467) 0.2555431839843941describes:

- 0: Row number of vectorised data in the training set
- 2467: Unique Integer number of each word in the first row
- 0.2555431839843941: Score calculated by TF-IDF Vectoriser

Meanwhile, BoW represents each text document as a numeric vector. Each dimension is a specific word from the dataset. The value can be its frequency in the document, occurrence (denoted by 1 or 0), or even weighted values. The output of the BoW model is a frequency vector. Figure 4 shows the results of non-zero feature positions in the sparse matrix.

```
(0, 6137)     3
(0, 979)      1
(0, 4293)     1
(0, 6568)     1
(0, 1666)     1
(0, 1564)     2
(0, 9293)     1
(1, 405)      2
(1, 5200)     1
(1, 9777)     1
(1, 5278)     2
(1, 8737)     2
```

Figure 4 - Non-zero feature positions in the sparse matrix

The feature matrix is traditionally represented as a sparse matrix since the number of features increases phenomenally with each document considering each distinct word becomes a feature. The preceding output tells us about the total count for each (x, y) pair. Here, x represents a document and y represents a specific word/feature, and the value is the number of times y occurs in x. All the unique words are then stored in a vocabulary as document feature vectors before being fed into the classifier.

D. Model Development and Evaluation

In this stage, the researchers use two machine learning algorithms, SVMand NB, to classify the YouTube comments into offensive or non-offensive comments. The researchers then analyse and compare the performance of these algorithms. This study measures the performance of the model based on accuracy, precision (P), recall (R), and F1-score (F), as follows:

- Accuracy: Correctly predicted instances.
- Precision: Correctly predicted instances over the total prediction of positive instances

- Recall: Correctly predicted positive instances over the total positive instances
- F1-score: The weighted average score of precision and recall

The overall accuracy is measured as shown in the following formula:

Accuracy= (TP+FN)/(TP+FN+FP+FN)

On the other hand, the equations for precision, recall, and F1-score are as follows:

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

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

Table 3 describes the confusion matrix:

Table 3 Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Where,

- True Positives (TP): Model correctly predicted positive instances
- True Negative (TN): Model correctly predicted negative instances
- False Positive (FP): Model incorrectly predicted positive instances
- False Negative (FN): Model incorrectly predicted negative instances
- m: number of class
- S: Support (The number of accurately predicted target instances)

These sampling techniques are performed on the imbalance dataset. Dataset has been transformed into the vector using weighting features before being fed into two machine learning algorithms, SVM and NB. The model performance is compared to achieve the best evaluation model.

IV. RESULT AND ANALYSIS

From the total of the comments consisting of identified offensive words, 1197 comments consisted of only one offensive word; while, the remaining 149 comments consisted of more than one offensive words. Figure 5 shows the results.

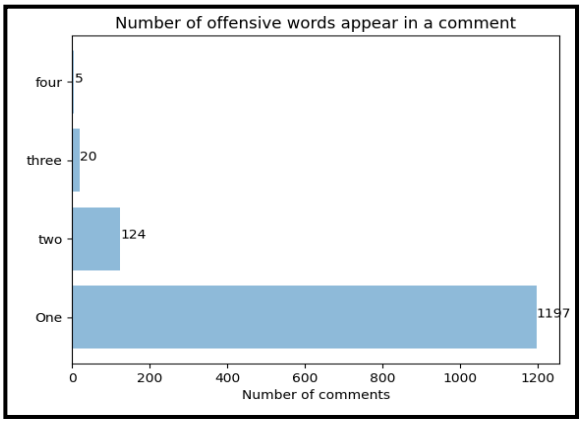

Figure 5: Number of offensive words appearing in comments

There were 43 Malay offensive words identified in this study, and only 23 of the words appeared in the raw YouTube dataset. Figure 6 shows the Malay offensive words appearing in the YouTube comments dataset. It shows that those 855 comments consist of the word "Bodoh", contributing most to the Class 1 category. It is followed by the word "mampus", "pukimak", and "anjing", appearing in 74, 73, and 69 comments, respectively.

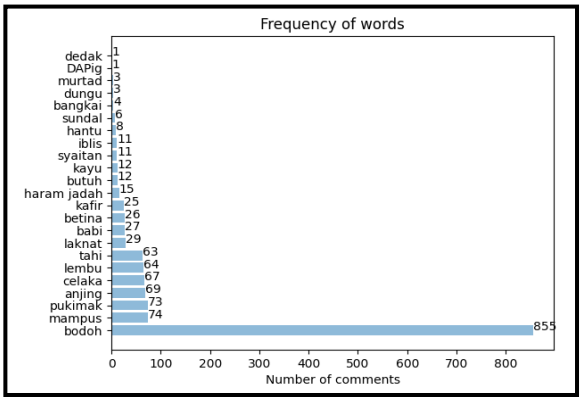

Figure 6: Number of offensive words appearing in comments

Based on the 94507 number of words in the whole dataset, Figure 7 shows the top 10 frequent words used. It shows that the most frequently used word in this dataset is "bodoh", followed by "tengok", "cakap", "suka", "kena", "bukan", "artis", "lagu", "kaya", and "diva". Out of the top 10 frequent words, only one word is listed in the Malay offensive words, which is "bodoh".

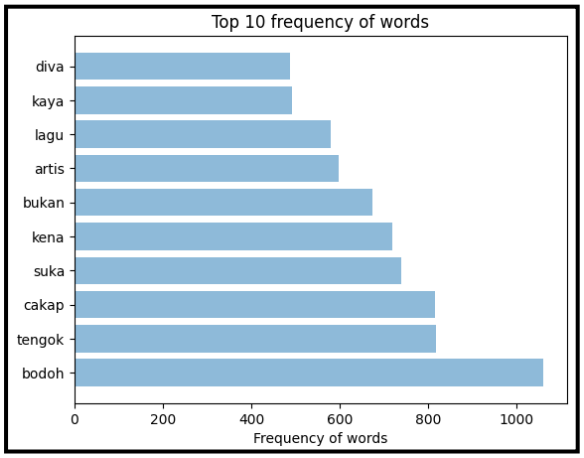

Figure 7: Frequency of Words

The last step in this study was to evaluate the classification model. This study employed Machine Learning Algorithm to measure the performance and accuracy of the model. In addition, this study used SVM and NB algorithm on TF-IDF and BoW weighting features. The execution of these classifiers was done using a library from Scikit-Learn. Scikit-Learn is a library providing some popular classification algorithms in Python language and often used to research text classification. One of them is research by Ibrohim and Budi (2018), using SVM, NB, and RFDT library from Scikit Learn. Then, the study compared the model performance for these classifiers. The model performance was tested based on their accuracy, precision, recall, and F1-score. This study chose recall as the main performance evaluation metric since it refers to the percentage of total relevant results correctly classified by the algorithm. The implementation of this process was done on both sampling dataset using the Scikit-Learn library.

Table 3 and Table 4 show the results for NB and SVM tested on different sampling techniques and weighting features. The results show that the models produced good results for both classifiers that is between 70% to 98%. Concerning the NB model, the accuracy, recall, and F1-score are high for the undersampling dataset compared to the oversampling dataset, using both weighting features.

The SVM model also produces similar results; the undersampling dataset produced better results than oversampling dataset. This study used sampling approaches as the dataset is imbalanced, and this played a significant role in defining the final model performance. As shown in the tables, both sampling techniques used for this study produced different results for different models. However, both models produced better results on the undersampling dataset, probably due to the majority classes being reduced and letting the models learn better.

Table 3: Results for the NB model for different sampling techniques and weighting features

Weighting Feature	Dataset	Accuracy	Precision	Recall	F1-Score
TF-IDF	Random undersampling	79.74	81.68	79.74	79.94
	Random oversampling	72.44	83.25	72.44	74.02
BoW	Random undersampling	81.04	85.03	81.04	81.43
	Random oversampling	70.42	87.66	70.43	73.21

Table 4: Results for the SVM model for different sampling techniques and weighting features

Weighting Feature	Dataset	Accuracy	Precision	Recall	F1-Score
TF-IDF	Random undersampling	98.70	98.73	98.70	98.70
	Random oversampling	94.23	94.82	94.23	94.25
BoW	Random undersampling	98.65	98.71	98.65	98.67
	Random oversampling	97.55	97.54	94.52	97.47

This analysis also found that models with the TF-IDF weighting feature are better than the BoW weighting feature when tested on SVM models instead of NB models. It is due to the TF-IDF characteristics containing information on the more important words and the less important ones than BoW just creating a set of vectors containing the count of word occurrences in the document.

Table 5: Recall results for NB and SVM models for different sampling techniques and weighting features

Weighting Feature	Dataset	NB	SVM
TF-IDF	Random undersampling	79.74	98.70
	Random oversampling	72.44	94.23
BoW	Random undersampling	81.04	98.65
	Random oversampling	70.43	94.52

Table 5 shows the compressed results of recall for both models with different weighting features and sampling dataset. For this study, high recall is better since the model developed does not want to lose the classes and predict wrongly. Based on the results, the findings showed that the SVM model is working better than NB when tested on both sampling dataset and weighting features. It is probably due to its ability to learn can be independent of the feature space dimensionality. On the other hand, NB models produced a slightly lesser percentage than SVM. Overall, SVM with TF-IDF using the Random undersampling technique is the best approach for this study.

V. CONCLUSION AND RECOMMENDATION

This study aims to detect offensive comments in the Malay language using the YouTube dataset. This paper discussed the offensive Malay language words, based on the MCMC list, on the YouTube platform, commonly coming from an unpleasant condition. Nowadays, the problem of offensive language is increasing like anything over social media. YouTube is a social media platform in which people have the freedom to say or express whatever they want. It is also a social platform for people of every age to join and find relevant content. Comments cannot be stopped but can be detected, and then it is the users who choose what they want to do with the comments, such as blocking or updating their content based on the comments. Text analytics is a great approach and an extensive area. Many approaches can be considered to do this analysis. This study employed ML and NLP to analyse offensive comments in the Malay Language on YouTube.

CRISP-DM is the approach for the study design. YouTube comments dataset was first understood, prepared, modelled, and later evaluated. In the initial stage, this study determined the offensive words in the Malay language context from the list of prohibited words taken from the MCMC website. These words were used in the data labelling stage. Different models were carried out when the data was preprocessed, and features were extracted from it. This study used SVM and NB as the classification models. The models were compared based on recall. The findings show that when recall is high, accuracy and F1-Score are also high. In conclusion, the SVM model

using both feature extractions and the undersampling technique performs better than the NB model. On the other hand, the NB model shows lower results probably due to its simplicity; it does not fully consider the actual message structure.

The overall performance results are consistent with a previous study by Tuarob and Mitrpanont [9] in which SVM produces higher results than NB in terms of accuracy, recall, and F1-score. Furthermore, Ibrohim and Budi [11] also obtained a similar results pattern, showing that all classifiers (SVM, NB, RFDT) produce higher results. However, when comparing our results with other previous studies, it must be pointed out that different dataset, weighting features, and data balancing techniques and labelling affect the results.

Based on the study's findings, there are some recommendations that future research can consider. Nevertheless, the main recommendation from this study is to test the classification model on different Malay language datasets collected from different social media platforms, such as Twitter, Facebook, and others, with massive samples. Furthermore, people often use informal Malay language on social media, as Maskat et al. (2019) described. However, many samples remain inconsistent and unstructured. Thus, further data cleaning and preprocessing technique using the Malay language as a domain are required. In addition, it is suggested that future research can develop a richer vocabulary of Malay words considering different forms of writing, such as short-forms, slang, Malay words written using English characters, and others. It is due to there are still many words that are missing out.

This study used common feature extractions. Therefore, this study suggests that future research can explore further using current feature extractions, such as n-grams, Latent Semantic Analysis (LTA), Predictive Word Embeddings, such as Word2Vec features, Doc2Vec features, and many others. In addition, future research may also consider punctuation, upper case, and laugh words dictionary for feature extractions to improve the classification results. More interestingly, future research can explore the current machine learning model, such as deep learning, and get more exciting results, such as a pattern of offensive words or producing new offensive words.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Computer and Mathematical Sciences for sponsoring this paper.

BIBLIOGRAPHY

1. ALFINA, I., MULIA, R., FANANY, M. I., & EKANATA, Y. (2017, October). Hate speech detection in the Indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 233-238). IEEE.
2. AL-KILIDAR, H., COX, K., & KITCHENHAM, B. (2005, November). The use and usefulness of the ISO/IEC 9126 quality standard. In 2005 International Symposium on Empirical Software Engineering, 2005. (pp. 7-pp). IEEE.
3. ANAGNOSTOU, A., MOLLAS, I., & TSOUMAKAS, G. (2018, January). Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech. In IJCAI (pp. 5796-5798).
4. BALANESCU, E., & DARIE, C. (2008). Beginning PHP and MySQL E-Commerce: From Novice to Professional, (Beginners/Beginning Guide).
5. BIANCO, F., & MICHELINO, F. (2010). The role of content management systems in publishing firms. International Journal of Information Management, 30(2), 117-124.
6. CASTANO-PULGARÍN, S. A., SUÁREZ-BETANCUR, N., VEGA,

L. M. T., & LÓPEZ, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. Aggression and Violent Behavior, 101608. Powell, T. (2002). Web design. McGraw-Hill Professional Publishing.

7. CHEN, Y., ZHOU, Y., ZHU, S., & XU, H. (2012, September). Detecting offensive language in social media to protect adolescent online safety. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 71-80). IEEE.
8. DELONE, W. H., & MCLEAN, E. R. (1992). Information systems success: The quest for the dependent variable. Information systems research, 3(1), 60-95.
9. DELONE, W. H., & MCLEAN, E. R. (2003). The DeLone and McLean model of information systems success: a ten-year update. Journal of management information systems, 19(4), 9-30.
10. DINAKAR, K., REICHART, R., & LIEBERMAN, H. (2011, July). Modeling the detection of textual cyberbullying. In fifth international AAAI conference on weblogs and social media.
11. DUL, J., DE VLAMING, P. M., & MUNNIK, M. J. (1996). A review of ISO and CEN standards on ergonomics. International Journal of Industrial Ergonomics, 17(3), 291-297.
12. FORTUNA, P., SOLER-COMPANY, J., & WANNER, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?. Information Processing & Management, 58(3), 102524.
13. GOEL, B., & SHARMA, R. (2019, June). USF at SemEval-2019 Task 6: Offensive language detection using LSTM with word embeddings. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 796-800).
14. HASSENZAHL, M. (2013). User experience and experience design. The encyclopedia of human-computer interaction, 2.
15. HERNANDEZ, S., ALVAREZ, P., FABRA, J., & EZPELETA, J. (2017). Analysis of users' behavior in structured e-commerce websites. IEEE Access, 5, 11941-11958.
16. IBROHIM, M. O., & BUDI, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science, 135, 222-229.
17. ISMAIL, M., DIAH, N. M., AHMAD, S., KAMAL, N. A. M., & DAHARI, M. K. M. (2011, June). Measuring usability of educational computer games based on the user success rate. In 2011 International Symposium on Humanities, Science and Engineering Research (pp. 56-60). IEEE.
18. JHA, V. K., HRUDYA, P., VINU, P. N., VIJAYAN, V., & PRABAHARAN, P. (2020). Dhot-repository and classification of offensive tweets in the hindi language. Procedia Computer Science, 171, 2324-2333.
19. LIN, H. F. (2007). The impact of website quality dimensions on customer satisfaction in the B2C e-commerce context. Total Quality Management and Business Excellence, 18(4), 363-378.
20. MAADI, M., MAADI, M., & JAVIDNIA, M. (2016). Identification of factors influencing building initial trust in e-commerce.
21. MAJID, E. S. A., KAMARUDDIN, N., & MANSOR, Z. (2015, August). Adaptation of usability principles in responsive web design technique for e-commerce development. In 2015 International Conference on Electrical

Engineering and Informatics (ICEEI) (pp. 726-729). IEEE.
22. MASKAT, R., & MUNARKO, Y. (2019). A taxonomy of Malay social media text. Indonesian Journal of Electrical Engineering and Computer Science, 16(1), 465-472.
23. NOBATA, C., TETREAULT, J., THOMAS, A., MEHDAD, Y., & CHANG, Y. (2016, April). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).
24. PARK, J. H., & FUNG, P. (2017). One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206.
25. PATEL, S. K., RATHOD, V. R., & PRAJAPATI, J. B. (2011). Performance analysis of content management systems-joomla, drupal and wordpress. International Journal of Computer Applications, 21(4), 39-43.
26. PURNAMA, B. E. (2015). e-Commerce web development in wiga art.
27. SINGH, T., MALIK, S., & SARKAR, D. (2016, April). E-commerce website quality assessment based on usability. In 2016 international conference on computing, communication and automation (ICCCA) (pp. 101-105). IEEE.
28. TANGCHAIBURANA, S., & TECHAMETHEEKUL, K. W. (2017). Development model of web design element for clothing e-commerce based on the concept of mass customization. Kasetsart journal of social sciences, 38(3), 242-250.
29. VANDERSMISSEN, B. (2012). Automated detection of offensive language behavior on social networking sites. IEEE Transaction.
30. YAZDANIFARD, R., & ZARGAR, A. (2012). Today need of e-commerce management to e-skill trainings. International Journal of e-Education, e-Business, e-Management and e-Learning, 2(1), 52.