International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Multilingual Hate Speech Detection

Tian Xiang Moy[1], Mafas Rahem[2], Rajasvaran Logeswaran[3]
[1,2,3]School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia, 57000
Email address: txmoy95@gmail.com, rmafas@gmail.com, loges@ieee.org

***Abstract*— *Due to recent reports of social media abuse, social media companies have been urged to address the issue of hate speech on social media. Advances in artificial intelligence and natural language processing have enabled the automation of hate speech detection. Despite that, challenges in the field of hate speech detection remain, as hate speech is highly context-dependent. This paper highlights the challenges of hate speech detection in multilingual communities and a solution for these challenges. This study adopts a hyperparameters fine-tuning approach on the pre-trained BERT model for the development of hate speech detection models in both the monolingual and multilingual scenarios. The findings of the research have revealed that the multilingual hate speech detection approximates or exceeds the performance of baseline monolingual hate speech detection models, achieving excellent performance on the English test data (Accuracy = 0.931, Precision = 0.877, Recall = 0.921, F-1 = 0.899) and the Malay test data (Accuracy = 0.872, Precision = 0.874, Recall = 0.868, F-1 = 0.871). The multilingual hate speech detection models can be applied to multilingual communities where members of the community use different languages interchangeably*.

***Keywords*— *Hate detection; multilingual model; dual language; social media; speech detection*.

I. INTRODUCTION

In the modern era of technology, social media is becoming increasingly integral, to the extent that it has become interwoven with many aspects of any individual's daily life. It is no exception for a developing country such as Malaysia, whereby a 24% increase in social media users have been observed in the Malaysian population from the year 2016 to 2021 [1]. As it stands, 86% of the Malaysian population are active social media users. Undoubtedly, social media has brought upon a lot of benefits and convenience to its users due to its accessibility and ease of use.

Unfortunately, there is a negative side to social media that warrants further work. One such example is the racial abuse of a professional footballer on social media, following poor performances in football matches [2]. Social media abuse has also been suspected to be the cause for the deterioration in the mental health of a Korean singer and entertainer, which ultimately led to her committing suicide [3]. Reportedly, the social media abuse received by these public figures contained sexist and racist ideals that can be loosely associated to hate speech, which is defined in [4] as *"any act which incites violence, spreads hatred, or threatens the safety, dignity and freedom of an individual based on their protected characteristics, such as gender, race, or sexual orientation*."

Hate speech can bring about detrimental effects to an individual's mental and physical well-being, as illustrated through the examples above. Hence, it is imperative to address the issue of hate speech on social media. Resultant of the reports, social media companies have been urged to tackle the issue of hate speech on their respective platforms. However, this remains a monumental challenge for these social media companies, as it is virtually impossible to manually monitor the vast number of exchanges on these social media platforms. Fortunately, the automation of the hate speech detection task has been made possible through recent advances in the field of artificial intelligence (AI) and natural language processing (NLP). A review of the literature has demonstrated that automated hate speech detection can be accomplished using machine learning [5] or deep learning methods [6], in conjunction with NLP techniques. However, previous researchers have highlighted the fact that most of the research focused only on European languages, particularly English [7]. Due to the emergence of hate speech-related concerns in non-English speaking countries, the need for automated hate speech detection in non-English languages is warranted.

To perform hate speech detection, a corpus in the targeted language is required. In essence, a corpus is a body of text acquired from various sources, with labels assigned to them across one or more dimensions [8]. The corpus plays a vital role in the model building process, as it is used to train models to perform specific tasks, which in this case, is to classify the body of texts into hate speech or non-hate speech. As the literature on hate speech detection is mostly focused on the English language, most of the corpora available in the literature are in English [9]. This results in a major impediment in the field of multilingual hate speech detection, as corpora for non-English languages are limited or unavailable. One prime example of such communities is Malaysia, a multi-cultural country where most members of the community are multilingual. As Malay is the national language, most Malaysians use at least some amount of it in their conversations, irrespective of their varying ethnic backgrounds [10]. However, there is a lack of publicly available corpus in the Malay language for NLP tasks [11]. Furthermore, there is no research in the literature on detecting hate speech in the Malay language to date. This research will attempt to address the challenges of hate speech detection in Malaysia.

The issue of a lack of corpus in non-English languages could potentially be circumvented by conducting hate speech detection from a multilingual standpoint, through the principles of transfer learning, where knowledge can be transferred from a high resource language (such as English) to a low resource language. Multilingual hate speech detection involves building a classifier using corpora from multiple languages and applying it to detect hate speech in a target

Tian Xiang Moy, Mafas Rahem, and Rajasvaran Logeswaran, "Multilingual Hate Speech Detection," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 4, Issue 10, pp. 19-28, 2022.

language. Extant research has demonstrated promising results from multilingual hate speech detection models, where the performance of multilingual hate speech detection models has been found to approximate or exceed the performance of monolingual models [9]. Additionally, previous research has also shown that it is possible to detect hate speech in a target language using a model trained with multiple languages aside from the target language, which indicates that it is possible to build a hate speech detection classifier without a corpus from the target language [9]. Thus, multilingual hate speech detection appears to be a promising solution to the issue of a lack of corpus in non-English languages. Thus, the main aim of this research is to build a multilingual hate speech detection model to detect hate speech in the Malaysian community.

II. Related Works

Different groups of researchers have attempted the classification task of hate speech detection through a variety of classification methods. Figure 1 displays an overview of the different classification methods employed by previous researchers in recent years.

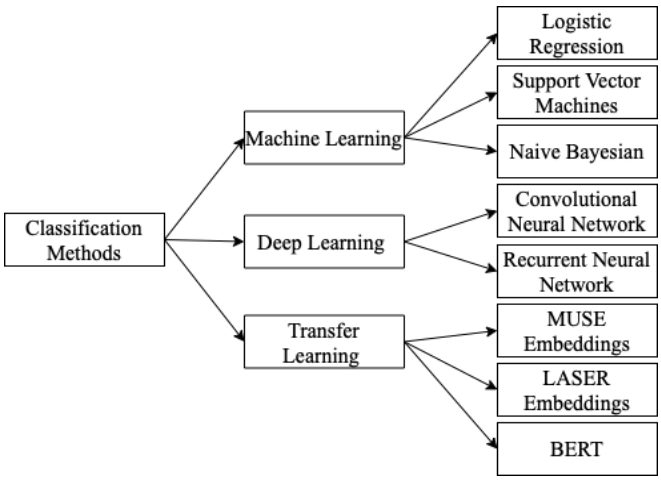

Fig. 1. Classification methods for hate speech detection

A. Machine learning methods

Machine learning methods emphasize manual feature engineering, in conjunction with statistical algorithms to perform the classification task [12]. In the context of hate speech detection, it is important to extract features from texts which can effectively distinguish between hate speech and non-hate speech. An exhaustive list of features for the classification task of hate speech detection has been proposed in [13], includes (1) simple surface features, such as bag-of-words and n-grams; (2) word generalizations, where a cluster of words with similar semantics are assigned a generalized feature; (3) negative sentiments, where sentiment analysis is used to detect the negative sentiments of hate speech; (4) lexical resources, which encompasses lists of words associated to various domains of hate speech; (5) linguistic features, such as type-dependent relationships between words; (6) knowledge-based features, which mainly refers to contextual information; (7) meta-information, which refers to information of users such as gender, activity level or the number of followers; and (8) multi-modal information, where information from other modes such as images or audio-visual content is emphasized.

Previous researchers have opted to employ various combinations of the features listed above and classification methods to perform the classification task. As hate speech in social media is embedded in texts, text-based features have traditionally been heavily emphasized. However, other features such as negative sentiments, meta-information and knowledge-based features have only been considered by researchers in recent years as text-based features cannot fully capture the different dimensions of hate speech.

In [14], a list of features such as word n-grams, character n-grams and negative sentiments was used in conjunction with four machine learning methods (Naïve Bayes, Bayesian Logistic Regression, Random Forest Decision Tree, and Support Vector Machine) to detect hate speech in the Indonesian language. They concluded that word n-grams were the most predictive feature, while negative sentiments were the least predictive feature in their models. Furthermore, they found that Random Forest, Decision Tree and Bayesian Logistic Regression models outperformed the Naïve Bayes and Support Vector Machine models in the hate speech detection task. In [5], the researchers employed word n-grams, character n-grams and user features (also referred to as meta-information) as features in a Logistic Regression model to detect hate speech in three different languages, namely, English, Portuguese, and German. They found that the inclusion of certain user features was able to consistently improve the performance of the hate speech detection models over the baseline models. However, they also noted that other user features did not demonstrate any effect on the hate speech detection models, which led to the conclusion that the effect of user features are highly dependent on the dataset.

The authors in [15] attempted to investigate the effect of social and cultural features (also referred to as knowledge-based features) on the performance of machine learning and deep learning models in the hate speech detection task. Furthermore, they compared the performance of machine learning models and deep learning models in the hate speech detection task. The findings indicated that the inclusion of social and cultural features was able to elevate the performance of the models over models that only included text-based features. They also demonstrated the importance of social and cultural context in distinguishing the different domains of hate speech in their research. The research showed that deep learning methods outperformed machine learning methods in the hate speech detection task. Another notable research in [16] sees two hate speech detection models being built in the Italian language using textual features (also known as simple surface features), lexical features, syntactic features (also referred to as linguistic features), as well as sentiment polarity. In this research, they recruited a machine learning model – Support Vector Machine and a deep learning model – Long Short-Term Memory (LSTM), for the hate speech detection task. Their findings demonstrated that the performance of the machine learning models approximated the

Tian Xiang Moy, Mafas Rahem, and Rajasvaran Logeswaran, "Multilingual Hate Speech Detection," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 4, Issue 10, pp. 19-28, 2022.

performance of the deep learning models in the hate speech detection task.

As illustrated above, some of the commonly used machine learning methods for the hate speech detection task include Naïve Bayesian, Logistic Regression and Support Vector Machines and Random Forest Decision Tree. Furthermore, most of the recent research have included text-based features and different non-text-based features for the detection of hate speech. Given all these research findings, it can be deduced that non-text-based features such as negative sentiments, knowledge-based features and meta-information has a relatively small influence on the performance of the model, compared to text-based features. However, non-text-based features were still shown to improve the performance of machine learning models over models that only employed text-based features. Hence, the inclusion of non-text-based features may prove to be beneficial for the overall performance of the model as it captures a different dimension of hate speech. Moreover, the comparison between machine learning and deep learning models have led to inconclusive findings, requiring further research in this area.

B. Deep learning methods

Deep learning methods place less emphasis on manual feature extraction, as the architectures are capable of automatically extracting multiple layers of features from the input provided [12]. Some of the more popular deep learning methods in the hate speech detection literature include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The CNN is a feed-forward neural network, where the input is fed through multiple hidden layers to produce an output. The main feature of CNN that distinguishes itself from other neural networks is the convolutional layers, which allows for the automation of feature extraction from the input data [12]. Hence, CNN is renowned as a "feature extractor" in the hate speech detection literature. On the other hand, the RNN is a neural network with a feedback loop that allows an output of the network to be fed back into the network [12]. In the context of hate speech detection, the feedback loop grants RNN the ability to analyze and store the semantics of a specific word in a sequence in one of its many hidden layers, to be used for the analysis of the next word in the sequence. Thus, it allows the RNN to learn dependency-relations between words.

In [17] and [18], the hate speech detection models in English were built using CNN, RNN and Gated Recurrent Unit (GRU) architectures. LSTM was used in [18] and [19] for models only in the English language, whilst the same was used in [20] for English, German and Italian. Another notable research in [6] used deep learning architectures to extract word embeddings from the input data to be fed into other machine learning models. The findings of the research have shown that the performance of machine learning models with features extracted using deep learning architecture superseded the performance of the deep learning models. Thus, the research demonstrated the potency of deep learning architectures as feature extractors.

The authors in [21] proposed a robust hate speech detection framework known as the Sub-word Enriched and Significant Word Emphasized (SWE2), which uses sub-word information to provide resistance against adversarial attacks. The SWE2 framework incorporated CNN and LSTM as feature extractors, where CNN was used to extract the sub-word information such as character level information and phonetic level information, while the LSTM was used to extract word-level information such as general content semantic information. The work in [22] used deep learning architectures for the detection of hate speech in multi-modal publications. The researchers attempted to build a hate speech detection model that leverages both textual and visual information, which is achieved through an ensemble of CNN and RNN architectures. The findings of the research indicated that the visual information contributed to the detection of hate speech in the multi-modal model. However, the multi-modal hate speech detection model did not outperform the hate speech detection model that only leveraged textual information.

In summary, there has been a multitude of research applying deep learning architectures in a variety of ways for hate speech detection. They have also been shown to be effective in extracting features from the data, such as word embeddings and sub-word embeddings. Due to the complexity of the deep learning architectures, the capabilities of these architectures extend beyond textual information and can be applied to visual information as well.

C. Transfer learning methods

Transfer learning is a new approach to deep learning to improve learning in a new task by transferring knowledge from a similar task [23]. In the field of hate speech detection, transfer learning is often seen as a solution for the lack of corpora in non-English languages. As there are considerably fewer resources for hate speech detection in non-English languages, transfer learning has been used to transfer knowledge from the high resource languages (e.g., English) to low resource languages. According to the authors in [24], transfer learning using pre-trained language representations can be achieved in one of two ways, which include: (1) feature-based approach, using pre-trained vector representations of words or embeddings as features for the training of hate speech detection models; and (2) fine-tuning approach, using pre-trained language models for the classification task of hate speech detection.

Previous researchers have achieved varying levels of success using transfer learning methods to detect hate speech in low resource languages. The authors in [25] attempted to build a model to detect hate speech in English and Spanish, where the English dataset was considerably larger than the Spanish dataset. Hence, the researchers opted to use pre-trained word embeddings such as Multilingual Unsupervised and Supervised Embeddings (MUSE) and Embeddings from Language Model (ELMo) with adversarial learning for knowledge transfer. The resultant word representations from the pre-trained embeddings were then fed into variations of CNN and LSTM architectures to perform the hate speech

detection classification task. Different variations of the deep learning architectures produced varying results, but the researchers concluded that the pre-trained word embeddings used had a significant impact on the performance of the model.

In [9], the authors conducted a large-scale research to detect hate speech in nine languages, namely, Arabic, English, German, Indonesian, Italian, Polish, Portuguese, Spanish and French. The researchers cited that there was a lack of resources in certain languages, hence, they used pre-trained word embeddings such as MUSE and Language Agnostic Sentence Representations (LASER), as well as pre-trained language models such as the Bidirectional Encoder Representations Transformer (BERT), to build different variations of monolingual and multilingual hate speech detection models. This research significantly contributed to the literature, as it identified the best variation of the hate speech detection model for each of the nine languages. In [26], the researchers leveraged on the knowledge in a pre-trained BERT model by fine-tuning it to build a hate speech detection model for the English language, different fine-tuning strategies for the pre-trained BERT model were proposed, including (1) BERT-based fine-tuning; (2) insertion of non-linear layers; (3) insertion of bidirectional LSTM layers; and (4) insertion of CNN layers. They found that the hate speech detection model built using the latter two fine-tuning strategies were able to outperform the model built using the BERT-based fine-tuning. This research expanded upon previous research and identified alternative fine-tuning strategies to improve the performance of BERT-based models. Additionally, the authors in [23] also used the transfer learning method to build three different hate speech detection models, which includes an English, Chinese and multilingual model. These models were built by fine-tuning a pre-trained BERT model using different datasets in the respective languages. The researchers also included an ensemble model by combining all three BERT-based models to form a multichannel BERT model. The research findings showed that the multichannel BERT model outperformed all the other models in two out of three datasets. Furthermore, the researchers also noted that all models in this research performed as well or better compared to previous state-of-the-art models. Hence, this research has shown the potency of BERT-based models in the hate speech detection task.

Another notable research that uses transfer learning methods is [27], where the authors attempted to build a generalizable hate speech detection model that can be applied to multiple languages. The dataset used for this research is a combination of English, Hindi, and code-mixed datasets. This research utilized various machine learning and deep learning methods, with pre-trained word embedding using the BERT model. The findings indicated that the multilingual model outperformed multiple baseline monolingual models. This research also demonstrated that the BERT model can be used to extract features such as word embeddings to be fed into other machine learning and deep learning models.

In short, these research findings have shown that transfer learning methods can be used to overcome the lack of resources in certain languages. It has also made it possible to build multilingual hate speech detection models using pre-trained word embeddings such as MUSE and LASER, as well as pre-trained language models such as the BERT model. Aside from that, transfer learning methods have also shown promising results concerning hate speech detection, as previous research has shown those transfer learning models approximate or even exceed the performance of state-of-the-art models. Hence, transfer learning methods have proven to be important for the expansion of the hate speech detection literature to non-English languages.

D. Challenges

A review of the literature has also revealed some of the major challenges faced by researchers in the field of hate speech detection. One of the major challenges includes the generalizability of the models. Previous researchers have revealed that existing models do not perform as well when applied on datasets other than the one with which the model was trained [18]. These findings suggest that the models are overfitted to the training dataset. In [18], the dataset had a large proportion of the hate speech originating from the same user. Thus, a model trained with such a dataset would be prone to user overfitting. In [28], the low generalizability was attributed to the fact that datasets used to train the models generally do not represent the full range of hate speech on social media, which covers a variety of domains that include sexuality, gender, race, and religion, while most of the datasets used by the researchers potentially emphasize certain domains more than others. Hence, the implementation of research methods to ensure that the dataset covers a wide variety of hate speech domains is vital to improving the generalizability of the model.

Besides that, the authors in [17] also raised concerns with regards to models that overfit certain frequently occurring words in the dataset, as this could result in certain biases in the model. In their study, they demonstrated that gender bias can be a consequence of the hate speech detection model overfitting certain words in the dataset. Moreover, [29] also revealed that certain corpora commonly adopted by researchers in the field propagate racial bias against African American English (AAE). They found that models trained using these corpora are twice as likely to classify AAE as offensive. Different groups of researchers have proposed innovative methods to reduce model bias. The authors in [30] proposed numerous methods to treat the training sample to reduce the amount of information available to the model. These methods are based upon the assumption that a hate speech detection model should be able to classify a body of text as hate speech or non-hate speech based on the text and not drawing upon confounding information from the text. The methods proposed include replacing bias-sensitive words with (1) part-of-speech tags; (2) named-entity tags; or (3) using K-Nearest Neighbor, as well as (4) knowledge generalization using lexical databases. Other techniques to reduce gender bias include [17]: (1) debiased word embedding; (2) gender swap data augmentation; and (3) fine-tuning with a larger

corpus. Limiting the number of texts from users could also potentially counteract user overfitting [18].

As illustrated from the findings, the training dataset has important implications for the generalizability of the hate speech detection model. Datasets that originate from various sources have different characteristics and demographics, which could result in an emphasis on different domains of hate speech [13]. The implications of the annotation process on the potential bias propagated by the training samples were demonstrated in [29], where racial and dialect priming could reduce the racial bias of the annotators in the annotation process. Hence, the identification of the best practices with regards to building a corpus or dataset for hate speech detection should be thoroughly investigated by future researchers.

III. METHODOLOGY

The main aim of this research is to build a multilingual hate speech detection model for a multilingual community using a transfer learning approach. The scope covers English and Malay, the two most widely used languages by the Malaysian population [31]. As hate speech is particularly rampant on social media, facilitated by the allowed anonymity, this research will focus on the detection of hate speech on social media. In light of the challenges faced by researchers in the field (as highlighted in the related works section), the datasets are chosen based on the criteria of the domains of hate speech involved, as well as the quality of the collection and annotation processes employed to build the dataset. For comparison purposes, this research will also build baseline monolingual models using both the Malay and English datasets respectively, in addition to building a multilingual model using a combination of both datasets.

A. Dataset

The English dataset used for this research is a combination of the Twitter datasets from [32] and [33]. The dataset[1] in [32] is a three-class (hate speech, offensive and neither) dataset built for multi-class classification. However, this research is only interested in distinguishing between hate speech and non-hate speech. Hence, the "offensive" and "neither" class from the dataset is assimilated. This is following the work of the authors in [28] who also assimilated the "offensive and "neither" classes for the dataset in [32]. Following the assimilation, the number of observations in each class is explored. The data exploration has revealed that this dataset has a class imbalance issue, where only 5% of the dataset is hateful tweets, while the other 95% are non-hateful tweets. Previous researchers have also reported the dataset in [32] to be highly imbalanced [18,28]. To counteract the class imbalance issue, the upsample approach has been adopted where hateful tweets from the dataset[2] in [33] were added to the former. Approximately 10,000 hateful tweets were added, resulting in a final dataset with approximately 33% hateful tweets and 67% non-hateful tweets.

The Malay dataset[3] is of Wikipedia comments built for the shared task, "Toxic Comment Classification Challenge" on the Kaggle platform [34]. The original dataset contained a total of 151,118 Wikipedia comments and is labelled by human annotators who distinguished them into six different classes: toxic, severe toxic, obscene, threat, insult, and identity hate. All six classes were assimilated into one class, as this research is only interested in a binary classification to distinguish hate speech and non-hate speech. After the assimilation, there was still class imbalance as the toxic comments only made up approximately 11% of the dataset. Downsampling by randomly removing non-toxic comments from this dataset resulted in the final data set of 30,000 comments, with 50% toxic comments and 50% non-toxic comments.

B. Data preprocessing

Each dataset required different preprocessing. Initial data exploration revealed that the English dataset contained HTML entities, such as "&#57361" that were decoded as emojis and converted into a word equivalent (e.g., the "😃" emoji is converted into "grinning_face_with_big_eyes"). User mentions, URLs and punctuations were removed. The English dataset also contained reserved words typical of Twitter datasets, such as RT (which represents retweet) and FAV (which represents favourite), which were consequently removed. Normalization was performed for words with different case types but the same vector representation (e.g., flower and Flower), by changing all to lowercase. Finally, stop words specified by the Natural Language Toolkit (NLTK) Python library were removed, except for the negation terms as they are commonly used to convey hate speech.

The Malay dataset had 261 observations with missing textual data that were removed as they did not provide additional information for model development. The dataset also contained URLs and punctuations, which were removed. As with the English dataset, the words were converted to lowercase and stop words (except for negation) were removed. The list of stop words in the Malay language was obtained from the Malaya Python Library.

C. Model development

In [9], it was concluded that a feature-based approach works better in a low resource setting, while a fine-tuning approach works better for a high resource setting. As the datasets identified in this research are much larger than what the authors in [9] defined as a high resource setting in their research, a fine-tuning approach was adopted.

In the literature, the BERT model was commonly fine-tuned for the construction of multilingual hate speech detection models [9,23,26]. It is a language model pre-trained using a large corpus from various languages. It encompasses a stack of transformer encoder layers, each consisting of two sub-layers: multi-head attention mechanism, and fully connected feed-forward neural network sub-layers [35]. The multi-head attention mechanism works by computing a key-value and query vector for every input token in a sequence,

[1] https://github.com/t-davidson/hate-speech-and-offensive-language
[2] https://github.com/mayelsherif/hate_speech_icwsm18
[3] https://github.com/huseinzol05/Malay-Dataset#toxicity-small

Tian Xiang Moy, Mafas Rahem, and Rajasvaran Logeswaran, "Multilingual Hate Speech Detection," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 4, Issue 10, pp. 19-28, 2022.

which will be further used to create a weighted representation. This allows the BERT model to learn contextual relations between words. Subsequently, the output of the multi-head attention mechanism will then be fed to the fully connected feed-forward neural network. Each sub-layer in the transformer encoder is connected through residual connection and the output of each sub-layer is normalized. The architecture of the transformer encoder is displayed in Figure 2.

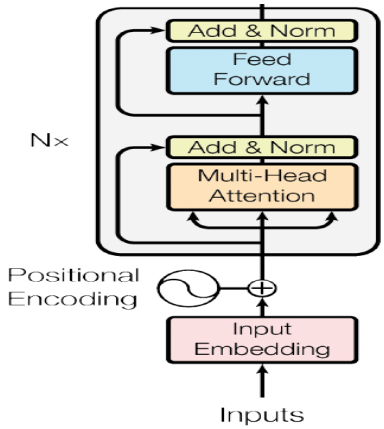

Fig. 2. Transformer encoder architecture [35]

The BERT model also distinguishes itself from previous transformer-based language models due to its bidirectionality, achieved using the pre-training objectives of "masked language model" (MLM) and "next sentence prediction" (NSP). The MLM task involves pre-training the model to predict a randomly selected masked word from the input sentence, while the NSP task pre-trains the model to predict whether a pair of input sentence is consecutive. The use of these pre-training objectives allows the BERT model to process text from both left to right and right to left, enabling it to gain more contextual information from the surrounding words in a sequence. This renders the BERT model to be ideal for this research, as the hate speech detection task is highly context-dependent [23,28,29]. Furthermore, previous researchers have also found that the inclusion of contextual information improves the performance of hate speech detection models [15]. Hence, this work employs BERT to build the monolingual and multilingual hate speech detection models.

The development of classification models in this research was achieved using the Hugging Face's transformers Python library. The BERT model is a pre-trained language model, but it can also be fine-tuned for classification tasks. This study opted to use the "BertForSquenceClassification" model in the transformers Python library, which adds a single linear layer on top of BERT's pooled output. Pre-trained weights of different BERT models were then loaded into the model for fine-tuning. Three different classification models (Malay, English and multilingual) were developed to detect hate speech by fine-tuning pre-trained BERT models. Hence, different weights were used for the English, Malay, and multilingual scenarios.

According to [36], learning rate and batch size are amongst the hyperparameters with the highest importance for the performance of the BERT models. The learning rate is an important hyperparameter for neural networks and it defines the rate of change of the model in response to the estimated errors whenever the model weights are updated. It is heavily dependent upon the data, as well as the models used. Furthermore, the batch size is defined as the number of samples processed before the model weights are updated and it is highly dependent on the size of the dataset. As different scenarios employ different datasets for the fine-tuning task, this study attempts to identify the optimal learning rate and batch size for the different scenarios (Malay, English and multilingual). The values of learning rate and batch size used in this study are referenced from the original BERT paper [24].

For model setup, the weights from "bert-based-uncased" pre-trained by the authors in [24] were loaded into the "BertForSquenceClassification" for the fine-tuning task of the English model. The "bert-base-bahasa-cased", which is a language model pre-trained with Malay text, was used for the Malay model, whilst "bert-base-multilingual-uncased" was used for the multilingual model. The evaluation strategy used for the fine-tuning task is "steps", which indicates that the evaluation metrics will be generated after a specified number of steps (in this case 500 steps) and 3 epochs were chosen.

The performance of the hate speech detection models was evaluated using a series of evaluation metrics commonly employed for classification tasks, namely, Accuracy (Acc), Precision (P), Recall (R) and F-1 scores (F-1). The performance of the English BERT model using the different combinations of learning rate and batch size is summarized in Table I, whilst the results for the Malay BERT and multilingual BERT are given in Tables II and III, respectively. From all three tables, the combination of hyperparameters in Model 3 demonstrated the lowest validation loss, hence, the best performing combination of hyperparameters for all three BERT models was with a learning rate of 2e-5 and batch size of 32. Therefore, Model 3 of each was chosen as the final BERT model for the respective English, Malay, and multilingual models.

TABLE I. PERFORMANCE OF ENGLISH BERT MODELS WITH DIFFERENT COMBINATIONS OF HYPERPARAMETERS

Model	Hyperparameters		Evaluation Metrics				
	BS	LR	VL	Acc	P	R	F-1
1	16	2e-5	0.240	0.935	0.907	0.896	0.902
2	16	3e-5	0.244	0.933	0.900	0.896	0.898
3	32	2e-5	0.208	0.936	0.904	0.903	0.904
4	32	3e-5	0.215	0.935	0.905	0.898	0.901

BS = Batch Size, LR = Learning Rate, VL = Validation Loss, Acc = Accuracy, P = Precision, R = Recall, F-1 = F-1 Score

TABLE II. PERFORMANCE OF MALAY BERT MODELS WITH DIFFERENT COMBINATIONS OF HYPERPARAMETERS

Model	Hyperparameters		Evaluation Metrics				
	BS	LR	VL	Acc	P	R	F-1
1	16	2e-5	0.416	0.876	0.881	0.868	0.875
2	16	3e-5	0.417	0.876	0.878	0.874	0.876
3	32	2e-5	0.354	0.873	0.876	0.868	0.872
4	32	3e-5	0.372	0.872	0.876	0.867	0.872

TABLE III. PERFORMANCE OF MULTILINGUAL BERT MODELS WITH DIFFERENT COMBINATIONS OF HYPERPARAMETERS

Model	Hyperparameters		Evaluation Metrics				
	BS	LR	VL	Acc	P	R	F-1
1	16	2e-5	0.321	0.907	0.888	0.885	0.886
2	16	3e-5	0.320	0.906	0.880	0.893	0.886
3	32	2e-5	0.275	0.908	0.886	0.889	0.888
4	32	3e-5	0.279	0.906	0.885	0.886	0.886

IV. EXPERIMENTAL RESULTS

This study has developed three hate speech detection models, two monolingual (English and Malay) and one multilingual. The monolingual hate speech detection models were evaluated using the test set in the corresponding language, whilst the multilingual model was evaluated using test sets in both languages.

A. Model evaluation

The developed models were evaluated in terms of the commonly used metrics, as in the previous section. The respective confusion matrices were generated to illustrate the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) predicted by the models. The confusion matrix for the English model is shown in Figure 3, with 2077 instances of TP, 4415 instances of TN, 256 instances of FP and 227 instances of FN.

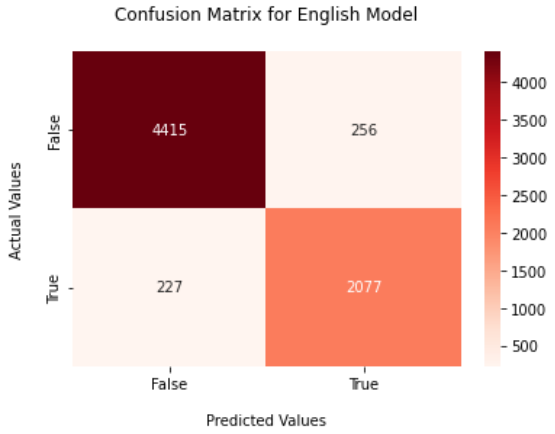

Fig. 3. Confusion matrix for English model

Fig. 4.

TABLE IV. EVALUATION METRICS FOR THE PROPOSED MODELS

Model	Test data	Accuracy (Acc)	Precision (P)	Recall (R)	F-1 score
English	English	0.931	0.890	0.901	0.896
Malay	Malay	0.873	0.874	0.872	0.873
Multilingual	English	0.931	0.877	0.921	0.899
	Malay	0.872	0.874	0.868	0.871

Using these, the evaluation metrics are calculated, as given in Table IV (see row for English model). At first glance, the accuracy of the English model is relatively high, which indicates that the English model is performing quite well. However, accuracy is often influenced by class imbalance, where the high accuracy is a result of correctly labelling most of the non-hate speech (dominant class) observations instead of intended purpose of a hate speech detection model. On the other hand, precision is a good evaluation metric when the cost of false positive (incorrectly labelling hate speech) is high. The precision score of the English model was also relatively high, which indicates that most of the observations predicted as hate speech by the model were actually hate speech. Recall is a good evaluation metric when the cost of false negative (incorrectly labelling non-hate speech) is high. From the table, the recall score of the English model is also relatively high, which indicates that the model can reliably detect hate speech in texts. The good performance of the model is further evidenced by the F-1 score, which is reflective of the high precision and recall score as the F-1 score is the result of the harmonic mean of precision and recall scores of the model.

Similarly, the confusion matrix for the Malay model is given in Figure 5, with 2684 instances of TP, 2692 instances of TN, 386 instances of FP and 395 instances of FN. The evaluation metrics for this model is also given in Table 4. where once again good scores were obtained for all the measures.

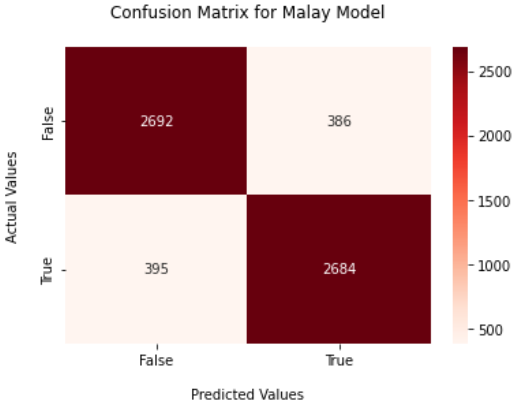

Fig. 5. Confusion matrix for Malay model

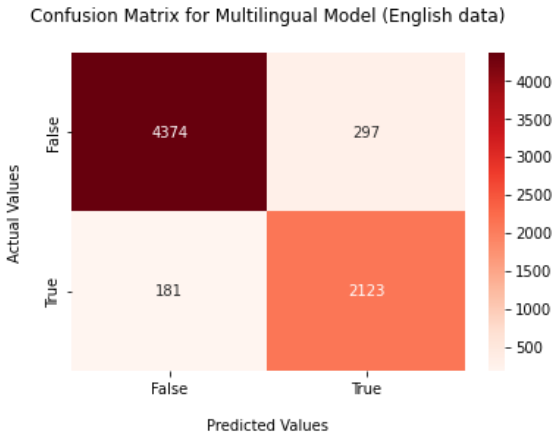

Fig. 6. Confusion matrix for multilingual model (English test dataset)

Two confusion matrices were generated for the multilingual model as it was tested on both datasets. Figure 5 illustrates the confusion matrix for the English test dataset, whilst the results for the Malay test dataset is given in Figure 6. From these, the evaluation metrics calculated were added to Table IV. From the results, it is found that the results were

Tian Xiang Moy, Mafas Rahem, and Rajasvaran Logeswaran, "Multilingual Hate Speech Detection," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 4, Issue 10, pp. 19-28, 2022.

high for both, although slightly higher for the test with the English dataset. This indicates that the multilingual model can reliably detect hate speech in both English and Malay text.

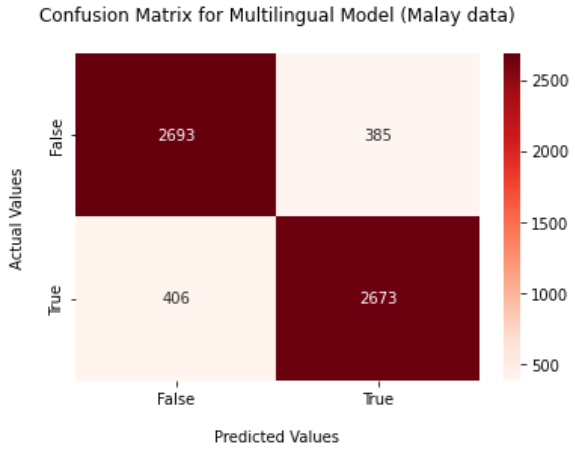

Fig. 7. Confusion matrix for multilingual model (Malay test dataset)

B. *Model comparison*

One of the main objectives of this study is to compare the performance of baseline monolingual hate speech detection models and multilingual hate speech detection models. From Table IV, it is observed that overall, the performance of multilingual and monolingual models is largely similar, except for small differences in certain evaluation metrics.

In the classification of the English test data, the multilingual and English models have comparable accuracy scores. However, the multilingual model scored lower in terms of precision, but higher in the recall. As the overall F-1 score of the multilingual model was higher, it can be deduced as the better model. In classifying the Malay test data, the multilingual model obtained similar accuracy and precision scores compared to the monolingual model but performed more poorly in terms of recall and F-1 scores. Thus, the Malay monolingual model can be seen as a better classification model for the Malay test data for hate speech detection.

V. DISCUSSION

The BERT model was chosen for this study due to its excellent capability in extracting contextual information in a body of text. Hence, this study opted to develop two BERT models in the monolingual scenario and one BERT model in the multilingual scenario for the task of hate speech detection. The hyperparameters of the BERT models were fine-tuned during the model development phase and the best combination of hyperparameters for each scenario were identified. Interestingly, the same combination of hyperparameters (learning rate of 2e-5 and batch size of 32) produced the best performance in all three BERT models. Hence, this may be used as a reference of the optimum combination of hyperparameters when fine-tuning the BERT model for hate speech detection tasks.

Following that, the performance of each BERT model was evaluated and critically analyzed. Various insights can be extracted from the evaluation of the BERT models. This study has confirmed that all three BERT models were able to demonstrate good performance in the classification task of hate speech detection. The performance of all three BERT models is largely comparable, except for minor differences in specific evaluation metrics. In line with the findings of previous researchers [9,25,27], a comparative analysis of the three BERT models has revealed that the multilingual model approximates or even exceeds the performance of baseline monolingual models. This finding is significant as it opens up an avenue for the development of a universal hate speech detection model, provided a large enough corpus or dataset can be obtained. Upon closer inspection, the multilingual model was found to be more suitable for the task of hate speech detection in the English language due to the high cost associated with false negatives. Conversely, the baseline monolingual model was found to be more suitable for the task of hate speech detection in the Malay language for the same reason mentioned above. Hence, different BERT models can be employed for the task of hate speech detection based on the need of the community. For example, a multilingual community might be better off with a multilingual model as members of the community tend to converse in multiple languages. On the other hand, monolingual communities can employ either the monolingual or multilingual models as the performance of these models are comparable.

It should also be noted that analysis of the losses in the training and validation test sets lead to suspect that the models may suffer from signs of overfitting. This could potentially reduce the generalizability of the model. Overcoming this issue in future work, through better training sample selection, could spur on better performance results.

Another limitation in this study that could be consider for future work is to expand on the range of hyperparameters of the BERT models that could be tuned, besides the two discussed in this work. Some options, subject to available resources, include weight decay and training epochs. A grid search could be employed for the purpose of tuning the hyperparameters of the BERT model to further improve the performance of the hate speech detection models.

To prevent overfitting of the BERT models, some potential strategies could include employing an early stopping call back approach while fine-tuning the BERT models. This approach allows the termination of model fine-tuning when the validation loss does not improve over a specified number of steps or epochs. By doing so, it can prevent the model from overfitting to the train data and result in a more generalizable model. Aside from that, future research could also alter the model architecture by adding kernel regularizers, such as L1 or L2 regularizers, batch normalization layers or dropout layers, to reduce model overfitting. This study employed a pre-built "BertForSequenceClassification" model where the model architecture cannot be altered. Hence, this study did not attempt to add any of the mentioned layers to prevent model overfitting.

Finally, this study only employed two languages (English and Malay) in the development of multilingual hate speech detection model as these are amongst the most commonly used languages in Malaysia. However, this potentially limits the

utility of the hate speech detection model as it can only detect hate speeches in these two languages. Future research can extend the work to include other popular languages used in the community, such as Mandarin and Tamil, for the development of a more comprehensive multilingual hate speech detection. By doing so, this would allow the hate speech detection model to be extended to other communities in Malaysia.

VI. Conclusion

In conclusion, this study has extended the hate speech detection literature to the Malay language. To the knowledge of the authors, there are no prior research for hate speech detection in the Malay language. Hence, this study has contributed to the body of knowledge in that aspect. Future research can build upon the findings of this research and improve the hate speech detection task in the multilingual scenario.

References

[1] Nurhayati-Wolff H. Active social media users as percentage of the total population in Malaysia from 2016 to 2021. Statista 2021. https://www.statista.com/statistics/883712/malaysia-social-media-penetration/ (accessed March 13, 2021).

[2] Sports S. Anthony Martial: Man Utd forward receives racist abuse online after West Brom draw 2021. https://www.skysports.com/football/news/11667/12218327/anthony-martial-man-utd-forward-receives-racist-abuse-online-after-west-brom-draw (accessed March 13, 2021).

[3] The Star. Dead K-pop star Sulli was target of cyber bullies 2019. https://www.thestar.com.my/lifestyle/entertainment/2019/10/14/sulli-target-cyber-bullies (accessed March 13, 2021).

[4] Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M. An Italian twitter corpus of hate speech against immigrants. 11th International Conference on Language Resources and Evaluation (LREC 2018), 2018, p. 2798–805.

[5] Fehn Unsvåg E, Gambäck B. The Effects of User Features on Twitter Hate Speech Detection. Proceedings of the second workshop on abusive language online (ALW2), Brussels: 2018, p. 75–85. https://doi.org/10.18653/v1/w18-5110.

[6] Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. Proceedings of the 26th international conference on World Wide Web companion, 2017, p. 759–60. https://doi.org/10.1145/3041021.3054223.

[7] Ombui E, Muchemi L, Wagacha P. Hate Speech Detection in Code-switched Text Messages. 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE; 2019, p. 1–6. https://doi.org/10.1109/ISMSIT.2019.8932845.

[8] Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. Resources and benchmark corpora for hate speech detection : a systematic review. Language Resources and Evaluation 2021;55:477–523. https://doi.org/10.1007/s10579-020-09502-8.

[9] Aluru SS, Mathew B, Saha P, Mukherjee A. Deep learning models for multilingual hate speech detection. ArXiv 2020:1–16.

[10] Adelaar KA. Malay - The National Language of Malaysia. Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas 2011:729–34.

[11] Lan TS, Logeswaran R. Challenges and development in Malay natural language processing. Journal of Critical Reviews 2020;7:61–5. https://doi.org/10.31838/jcr.07.03.10.

[12] Biere S. Hate Speech Detection Using Natural Language Processing Techniques. 2018.

[13] Schmidt A, Wiegand M. A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the fifth international workshop on natural language processing for social media, 2017, p. 1–10. https://doi.org/10.18653/v1/w17-1101.

[14] Alfina I, Mulia R, Fanany MI, Ekanata Y. Hate speech detection in the Indonesian language: A dataset and preliminary study. 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE; 2018, p. 233–8. https://doi.org/10.1109/ICACSIS.2017.8355039.

[15] Vijayaraghavan P, Larochelle H, Roy D. Interpretable Multi-Modal Hate Speech Detection. ArXiv 2021.

[16] del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. Hate me, hate me not: Hate speech detection on Facebook. Italian Conference on Cybersecurity (ITASEC17), Venice: 2017, p. 86–95.

[17] Park JH, Shin J, Fung P. Reducing gender bias in abusive language detection. ArXiv 2018:2799–804.

[18] Arango A, Pérez J, Poblete B. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, 2020, p. 45–54. https://doi.org/10.1016/j.is.2020.101584.

[19] Xia M, Field A, Tsvetkov Y. Demoting Racial Bias in Hate Speech Detection. ArXiv 2020:1145–54. https://doi.org/10.18653/v1/2020.socialnlp-1.2.

[20] Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. A Multilingual Evaluation for Online Hate Speech Detection To cite this version : HAL Id : hal-02972184 A Multilingual Evaluation for Online Hate Speech Detection. ACM Transactions on Internet Technology (TOIT) 2020;20:1–22. https://doi.org/10.1145/3377323.

[21] Mou G, Ye P, Lee K. SWE2: SubWord Enriched and Significant Word Emphasized Framework for Hate Speech Detection. International Conference on Information and Knowledge Management, Proceedings, 2020, p. 1145–54. https://doi.org/10.1145/3340531.3411990.

[22] Gomez R, Gibert J, Gomez L, Karatzas D. Exploring hate speech detection in multimodal publications. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2019, p. 1470–8.

[23] Sohn H, Lee H. MC-BERT4HATE: Hate speech detection using multi-channel bert for different languages and translations. IEEE International Conference on Data Mining Workshops (ICDMW), IEEE; 2019, p. 551–9. https://doi.org/10.1109/ICDMW.2019.00084.

[24] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv 2018.

[25] Bojkovský M, Pikuliak M. STUFIIT at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter with MUSE and ELMo Embeddings. 13th International Workshop on Semantic Evaluation (SemEval-2019), Minnesota, USA: Association for Computational Linguistics; 2019, p. 464–8. https://doi.org/10.18653/v1/s19-2082.

[26] Mozafari M, Farahbakhsh R, Crespi N. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. International Conference on Complex Networks and Their Applications, Springer, Cham; 2019, p. 928–40. https://doi.org/10.1007/978-3-030-36687-2_77.

[27] Vashistha N, Zubiaga A. Online multilingual hate speech detection: Experimenting with hindi and english social media. Information 2021;12:1–16. https://doi.org/10.3390/info12010005.

[28] Gröndahl T, Pajola L, Juuti M, Conti M, Asokan N. All you need is "love": Evading hate speech detection. Proceedings of the ACM Conference on Computer and Communications Security, 2018, p. 2–12. https://doi.org/10.1145/3270101.3270103.

[29] Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. Proceedings of 57th Annual Meeting of the Association for Computational Linguistics 2019:1668–78. https://doi.org/10.18653/v1/p19-1163.

[30] Badjatiya P, Gupta M, Varma V. Stereotypical Bias Removal for Hate Speech Detection Task. World Wide Web Conference (WWW '19), San Francisco: Creative Commons CC-BY 4.0; 2019, p. 49–59. https://doi.org/10.1145/3308558.3313504.

[31] Thirusanku J, Yunus MM. Status of English in Malaysia. Asian Social Science 2014;10:254–60. https://doi.org/10.5539/ass.v10n14p254.

[32] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. Proceedings of the

11th International Conference on Web and Social Media, ICWSM 2017, 2017, p. 512–5.

[33] ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding E. Hate lingo: A target-based linguistic analysis of hate speech in social media. 12th International AAAI Conference on Web and Social Media, ICWSM 2018, 2018, p. 42–51.

[34] Kaggle. Toxic Comment Classification Challenge | Kaggle. Kaggle 2018. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (accessed September 18, 2021).

[35] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, p. 5998–6008.

[36] Kamsetty A. Hyperparameter Optimization for Hugging Face Transformers . Distributed Computing with Ray 2020. https://medium.com/distributed-computing-with-ray/hyperparameter-optimization-for-transformers-a-guide-c4e32c6c989b (accessed December 29, 2021).