Journal of Theoretical and Applied Information Technology
15th October 2024. Vol.102. No. 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRANSFORMER-BASED MODEL WITH CNN AND CAPSNETS TO IMPROVE MALAY HATE SPEECH DETECTION IN TWEETS

NUR UMAIRA ABD RAHIM[1] , NORWATI MUSTAPHA[2]

[12]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia

E-mail: [1]umai127rahim@gmail.com, [2]norwati@upm.edu.my

ID 55229 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
01-08-2024	03-08-2024	04-10-2024	13-10-2024

ABSTRACT

With the rise of social media, the spread of hate speech poses a significant threat to online harmony, especially within the Malay-speaking community. Existing research mainly focuses on high-resource languages like English, leaving a gap in effective HSD for low-resource languages like Malay. Even with a study done in previous research on Malay HSD, there is some room for improvement, and the lack of diverse datasets may significantly affect the system's overall performance and generalization. Thus, this study proposes a model that uses a transformer-based model named RoBERTa integrated with CNNs and Capsule Networks. RoBERTa is very effective in handling contextual information in bidirectional ways. Experimental results demonstrate that the proposed models, which are RoBERTa, outperform other models in a new dataset in terms of F1-score and accuracy, which are 84.54% and 84.45%, respectively and also outperform the existing dataset, which is 77.67% and 77.45%, respectively. By offering an extensive architecture, this research not only advances the technological area but also tackles social problems by enabling safer online environments for Malay speaker's communities. Additionally, this research contributes a valuable new Malay Hate Speech dataset, enriching resources for low-resource languages. The results underscore the importance of dataset diversity and advanced NLP techniques in generalizing well across different datasets, making this model practical for real-world applications. Furthermore, this study highlights the global potential of these techniques for improving HSD in other low-resource languages.

Keywords: *Hate Speech Detection, Transformer, Natural Language Processing, XLNet, BERT, RoBERTa*

1. INTRODUCTION

A few years back, the arrival of social media has transformed communication, enabling diverse interactions among individuals and communities. However, these platform's widespread and continuous use has also given rise to a significant challenge: the growth of online hate speech. The uncontrolled spread of harmful and offensive content poses a significant threat to online harmony and social well-being [2]. Any statement that mocks, discriminates against, or promotes violence against individuals or groups based on characteristics, including race, religion, ethnic origin, sexual orientation, handicap, or gender, is considered hate speech [7].

Within the large number of social media users, the Malay-speaking community has emerged as an active participant, particularly in issues that happened within Malaysia. As individuals engage their daily lives through digital spaces, the community has experienced a concerning increase in the occurrence rate of online hate [3]. A statistic from IPSOS, cited in [8] article, shows the concern of high percentages of Malaysian cyberbullying compared to the global average percentage [9] highlighting Malaysia's concerning rank in cyberbullying incidents on social media, emphasizing the urgency of addressing this issue to protect online users.

The harmful effects of hate speech can further affect individual ability, posing a significant threat to multicultural harmony. Recognizing the seriousness of the situation, it becomes essential to actively

address the issue of hate speech. Hate Speech Detection (HSD) is one of the important solutions to counter these hate speech problems, which can automatically identify and mitigate any hateful or offensive characteristics directed at particular groups or individuals.

According to a survey, [2] stated that creating good and efficient HSD systems is vital to reduce the negative impacts of online hate speech and give all users better and safer digital environments. On social media and other websites, hate speech has spread very quickly. Any content made by users should be blocked by good HSD systems so that other users don't see any dangerous content. It is crucial for vulnerable people who are often hated and targeted as a victim of hate speech. Platforms like Twitter, Facebook, and YouTube have increased their HSD technology investments to promote inclusive and courteous online communities [4].

While most HSD has demonstrated its effectiveness in making the internet a safer and more inviting space, the existing research is more focused on high-resource languages, especially English. This leaves a clear deficiency and need for addressing hate speech in low-resource languages [5], like Malay. Remarkable research on this problem was done by [1], who contributed to creating Malay HSD and the benchmark 'HateM' dataset. This dataset, compiled from Malay language tweets on Twitter, and their two-channel deep learning HSD models demonstrate a good result.

Due to limited research in this area, developing an effective system is challenging, as more work is needed to understand the complex words, slang, and language structures used by Malay speakers. For instance, Malay speakers on social media often blend formal and informal language and mix sentences with multiple languages. Even with previous research that was done by [1], the performance and the results of the model have not reached a satisfactory result, and the current models are unable to give high performance to say that the model is strong enough. This shows that Malay HSD has room for improvement and advancement.

Another problem in this research is that there aren't many diverse and extensive hate speech datasets, especially for languages like Malay. There is only one Malay Hate Speech Dataset (HateM) available, which is a benchmark dataset created by [1], and this lack of diversity is affecting the ability of HSD models to have a generalized system [21]. Using just one dataset makes it harder for models to handle the wide range of hate speech that happens in real life, especially when it comes to the Malay language. Creating and using new datasets in the proposed model is, therefore, an important step to improve the training and testing processes. Adding more datasets can help the model perform better overall and be more useful in various [6].

In this study, the main objective is to propose a model that integrates XLNet [24], BERT [22], and RoBERTa [23] algorithms with CNN [2] and Capsule Network [25] to improve the accuracy of Malay HSD. A study by [10] highlights the better performance of transformer-based models over other deep learning models in various HSD benchmarks, demonstrating their potential to tackle the complicated way of hate speech. Another objective is to analyze the generalization of the proposed model across diverse hate speech data by experimenting with it on the existing dataset and the newly collected dataset.

The proposed architecture may not only advance the field of HSD for low-resource languages, especially in Malay, but also provide a scalable framework that could potentially be adapted to other linguistics. Also contributes in newly created dataset that adds more diversity by capturing recent socio-political events. This work represents a critical contribution to the growing field of NLP, addressing both technical challenges and societal needs by enabling safer digital environments for the Malay-speaking community. The rest for the article is structured as follows. In section 2, we discuss on related works to on HSD. In section 3, we focus on dataset used and proposed method. In section 4 and 5 we discuss on results and findings. And lastly, conclusion and future works are highlighted in section 6.

2. RELATED WORKS

In this section, a detailed critical assessment of the works by previous researchers that implemented HSD in several contexts and languages is presented. The section also highlights the method, datasets, results, and limitations of those author's respective systems. The previous research, which was carried out between 2020 and 2024, had the primary objective of developing methods for making the HSD system more effective.

Significant research has been conducted on HSD across various research areas. For instance, multilingual or cross-lingual HSD has been explored

by [11] and [12], where models capable of detecting hate speech across multiple languages were developed using diverse language datasets. However, a significant limitation of such research is the necessity to train models with numerous languages due to variations in hate terms across different linguistic contexts.

Other areas of focus lie in contextual-based Hate Speech Detection (HSD), as demonstrated by [13], who developed a model to identify the contextual variables influencing hate speech. Additionally, [14] presented another HSD approach utilizing parallelized ensemble learning models, with a specific emphasis on the parallelized bagging method for hate speech detection.

All studies above have explored different techniques applied to HSD models. However, most of these studies have utilized general or standard datasets comprising high-resource languages such as English, Chinese, and German. Some research has shifted attention to low-resource languages like Dravidian, as seen in the work of [15], focusing on Tamil and Malayalam languages and employing the Deep Ensemble method of BERT+DNN+MuRIL. Although it achieved a high F1 score of 93.3%, the model is limited to shorter sentences and may not perform well with longer, more complex texts. Our work overcomes this by focusing on tweets, which vary in length and complexity, ensuring the model can handle diverse text inputs. Similarly, research on Bengali datasets by [16] utilized a transformer-based model of G-BERT (BERT+GRU), resulting in a high F1-measure of 90%.

HSD has been a wide range of models applied from standard Machine Learning to an advanced Deep Learning model in various kinds of languages. So far, there is only one research study that focuses on Malay HSD, which was conducted by [1]. The author proposed a model called XLCaps model with two channel of deep learning model and the effectiveness of every single channel was also assessed. The first channel, which utilizes XLNet, combined with Capsule Network, achieved an F1 score of 77.48%. The second channel, which employs FastText embeddings adding to Bi-GRU with attention, achieved an F1-score of 74.72%. But for this study, we will focus on the first channel, which integrates with the transformer-based model.

In a recent research venture, HSD has undergone comprehensive and in-depth studies, focusing on transformer-based models and different methodologies across various linguistic contexts. [17] focused on research on Cyberbullying Detection, particularly emphasizing emotion-based analysis. Utilizing models such as XLNet and BERT on various cyberbullying datasets, they achieved remarkable results, attaining an F1 measure of 96%. Despite this success, the study identified a limitation stemming from the need for well-annotated data, posing potential biases within the dataset.

[18] applied a BERT+CNN method to English datasets, achieving an F-measure of 73%. While their approach demonstrated effectiveness in capturing contextual nuances, it was limited by the use of standard CNN layers, which may not capture complex spatial hierarchies in text as effectively as Capsule Networks. Our study extends this line of research by integrating Capsule Networks, which excel in capturing hierarchical relationships in text data, leading to better generalization across diverse datasets.

Recent advances in hate speech detection (HSD) have focused heavily on transformer models due to their superior ability to capture contextual information. For example, [19] utilized BERT and RoBERTa for multiclass and multilabel classification, achieving a notable F1-score of 79.59%. However, their study focused on high-resource languages and an unbalanced dataset, limiting its generalizability to low-resource languages like Malay. In contrast, our work improves on these models by integrating RoBERTa with CNN and Capsule Networks, which enhances its ability to capture both local patterns and complex word relationships in Malay text.

[20], in their research on HSD, use the DistilBERT model for the multiclass hate speech and offensive (HSO) dataset. The results obtained are an accuracy of 92% and an F1 measure of 75%, respectively, which outperform the other transformer-based models. Overall, these studies collectively contribute to advancing the understanding of HSD across different linguistic and social media contexts, shedding light on both successes and challenges in the pursuit of more effective and nuanced automated moderation systems. In this study, a new model is proposed to improve the Malay HSD by integrating it with transformer-based model and a new dataset is created to have dataset diversity for better model performance.

3. METHODOLOGY

This section will explain the methodology, and topics like data acquisition, model development, experimentation, and evaluation metrics.

3.1 Data Acquisition

This study will use both existing and newly collected datasets that will be scraped from Twitter and consist of Malay Language only. Most of the steps for new dataset collection are the same method used by [1] to collect their 'HateM' dataset. The steps involved are data collection, data cleaning, data annotation, and data statistics, which will be explained in the next part.

3.1.1 HateM Dataset

The HateM dataset, introduced by [1], consists of 4,892 annotated tweets specifically designed for hate speech detection in the Malay language. The dataset is an unbalanced mix of hate and non-hate speech, with total tweets of 1,890 and 3,002, respectively, with an average post length of around 22.35. The selection period for the previous research is from December 2022 until January 2023, and during that duration, there are a lot of things happened, like the Appointment of the Deputy Prime Minister of Malaysia and some other major events that may have had a significant impact on generalization and diversity on the hate speech dataset.

3.1.2 New Data Collection

The initial phase of this study involves gathering the new raw dataset that will be collected and compiled from tweets on Twitter based on specific 'keywords' that contain hate definitions in the Malay language that are used by [1]. Some of the specific keywords that will be used are 'Bodoh' – Stupid, 'Sial' – Damn, 'Gila' – Insane, 'Babi' – Pig, 'Haram' – Forbidden, 'Anjing' – Dog, 'Mati' – Dead, 'Setan' – Devil, 'Celaka' – Unfortunate, 'Bangsat' – Bastard, 'Jahat' – Evil, 'Hitam' – Black, 'Pendek' – Short, and 'Lembab' – Slow.

The new dataset that will be used in this experiment may have a different landscape from the previous one as the current major events are different from before. The newly collected dataset is taken from tweets from the period of October 2023 until June 2024, and currently, we have huge issues ongoing globally, like the Palestine issues and the boycott issues, as well as current Malaysian political issues. However, as mentioned before, the keywords that will be used for this new dataset are still the same as those used for the previous dataset.

3.1.3 Data Cleaning

Once the raw dataset is collected and combined, a crucial step is to clean the data to ensure its quality and relevance. The raw data will have a lot of irrelevant tweets, and we need to clean the raw data by creating some filter that will remove the following aspects:

- The repeated tweets.
- The tweets with only URLs.
- The tweets are in languages other than Malay such as English and Indonesian.
- The tweets that are less than ten characters.

3.1.4 Data Annotation

Annotation plays a crucial role in enhancing the quality of the collected dataset. It is a process where human annotators thoroughly analyze and mark specific attributes or characteristics within the data. In the context of HSD, annotation focuses on two fundamental aspects below:

- Define Hate Speech to annotators

Before going to the annotation process, we must grasp and understand the meaning of hate speech as contextual variation exists [1]. Hate speech is not always straightforward and can take on different meanings based on the context in which it is being used. The annotators will need to employ their understanding skills to identify and label hate speech based on these predefined definitions or criteria. Since hate speech in Malaysia lacks a national definition, the annotating guidelines will adopt the UN's definition and include the protected features that Twitter specified [1].

- Annotation Training

In the annotation training phase, a group of annotators work together throughout the process. This approach is a widely adopted method in survey done by [2]. The use of multiple annotators introduces diversity in perspectives, minimizing individual biases and contributing to a strengthened and reliable dataset. Each annotator brings their unique understanding and interpretation, enhancing the overall annotation process.

Building upon predefined definitions or criteria in step before, hate speech will be clearly defined to the annotators. This step provides a clear guideline for annotators to identify and label instances of hate speech within the dataset. The definitions act as a reference point, ensuring a consistent and standardized approach to identifying hate speech across different annotators. This clarity is crucial for maintaining the dataset's integrity and reducing ambiguity in the annotation process.

To annotate the new dataset for Malay Hate Speech Detection (HSD), we selected three diverse Malay-speaking annotators with different backgrounds who are familiar with and use social media platforms, especially Twitter. Using Online Forms, each annotator needed to label each tweet as either hate speech or no ("Yes" or "No").

Clear guidelines and examples were provided to standardize the annotation process. Regular reviews and discussions resolved any disagreements, ensuring consistency and quality. As a token of appreciation, annotators received a small gift upon completing their tasks.

3.1.5 Data Statistic for new dataset

The newly collected Malay Hate Speech dataset comprises 2,822 tweets, which are labeled as either "Yes" or "No", which represents Hate and Non-Hate, respectively. Specifically, 1,806 tweets are labeled as "Yes", while the remaining 1,016 tweets are marked as "No". The average post length in this dataset is approximately 22 words. Figure 1 shows the bar chart for each total hate and non-hate speech, while Figure 2 shows the distribution of the bar chart for the total count of each keyword in the new dataset. Differ from the existing dataset, this newly collected dataset has more hate tweets than non-hate tweets, but both datasets still have this imbalanced data that contributes more to the majority class. This will be solved when weight is applied to minority classes during classification.

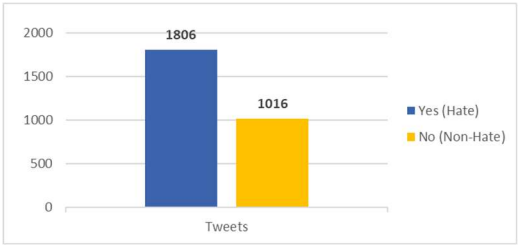

Figure 1: Total Tweets on each label in newly collected dataset

Upon examining the dataset, it is observed that certain keywords contribute significantly to the number of hate tweets. For instance, the keyword "anjing", which means "dog", appears frequently in hate speech tweets. However, it is important to note that even though some keywords have offensive or profane meanings, they are still being categorized as "Yes" which means hate speech. Any unrelated text to hateful, offensive, or profane is labeled as "No," which is non-hate speech. This observation underscores the deeper challenges involved in hate speech detection, where context plays a crucial role.

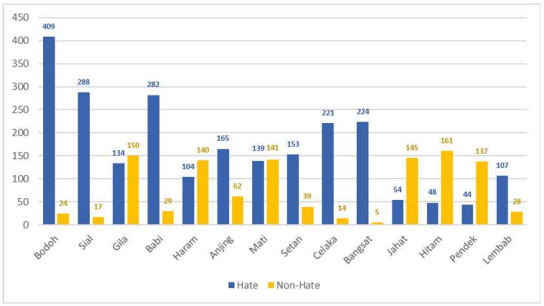

Figure 2: Total Count of each keyword in newly collected dataset

Table 1 below provides examples of tweets collected and labeled for the study. The label indicates whether the entry was classified as "Yes" or "No". For example, Tweet 1 contains an offensive statement and is labeled as "Yes". On the other hand, Tweet 2, which uses similar offensive language in a different context, is labeled as "No". This table gives you an idea of the type of data we used for this study and shows the range of expressions that were categorized and examined and where context and usage play crucial roles in determining whether an entry is considered hate speech or not.

Table 1: Example of tweets in newly collected Malay Hate Speech Dataset

Tweets	Label
T1: *Tak sangka, pompuan tu lagi bangang dari anjing yg dia dukung tu. Kasihan.* Translate: I didn't expect it, that woman is even more stupid than the dog she is carrying. Pity.	Yes
T2: *Allah kesiannye. Kenapa la nak langgar bkn anjing tu tgh belari ke apa.* Translation: Oh, poor thing. Why would they hit it? It's not like the dog was running or anything.	No

3.2 Model Development

The critical point of the model development in this experiment revolves around the Malay HSD model. This integration aims to enhance the model's capacity to detect expressions of hate speech. The step-by-step process is explained in the following sections.

3.2.1 Data Preparation

The most commonly used data pre-processing techniques in Malay text analysis are removing stop words and tokenization. However, for this proposed model, we will use each transformer model's tokenizer, which differs from using a tokenizer specifically made for the Malay language. Data pre-processing or cleaning involves removing stop words or common Malay words like 'dan', 'ialah', 'adalah', etc. Malay stop words can be obtained here [27], which is already compiled into one .txt raw file.

The pre-processing also involves changing all text to lowercase and removing any links, symbols or numbers as it is irrelevant to understanding the tweets. This whole pre-processing and data-cleaning step helps restructure the dataset by eliminating noise and reducing computational load during analysis, contributing to a more focused and meaningful text representation [26].

The dataset will be divided into training, validation, and testing sets in the subsequent steps according to the standard research procedure of 8:1:1. This means 80% of the data will be used for training the model, 10% for validating it, and the remaining 10% for testing [1]. The training set will teach the model what to do, and the validation set will help fine-tune its parameters and keep it from overfitting by showing how well it does on data it hasn't seen in training. Finally, the testing set will be used to evaluate how well the model generalizes to new, unseen data, ensuring it performs well in real-world scenarios.

As mentioned above, the tokenization for this proposed model will be using the existing tokenizer on each one of the transformer models, where the outputs of each tokenizer vary. The functions called from TensorFlow for both the tokenizer and model are the same. For example, the XLNet model will use 'xlnet-base-cased', the BERT model will use 'bert-base-cased', and the RoBERTa model will use 'roberta-base'.

All transformer models will use the 'base' configuration, which consists of a 12-layer encoder. This configuration is suitable for general tasks and moderately complex data. For more extensive and complex datasets, the 'large' configuration, which includes a 24-layer encoder, can be employed to capture more complex patterns and dependencies within the data [29].

3.2.2 Model Architecture

We implemented a combination of the transformer-based model, Convolutional Neural Networks (CNNs), and Capsule Networks to build a model framework for finding hate speech. Here's a detailed explanation:

- Input Layer:

The model starts with two and three input layers: 'input_ids', 'token_type_ids', and 'attention_mask', each with a shape of 128. These inputs are generated from the transformer's tokenizer.

- Transformer Model:

XLNet model ('xlnet-base-cased'), BERT model ('bert-base-model') and RoBERTa model ('roberta-base') process the tokenized inputs to produce a 'last_hidden_state' output, which provides contextual embeddings for the input text which also being called 'sequence_output' in this model.

- Convolutional Layer:

The 'sequence_output' is fed into a Convolutional Neural Network (CNN) layer. The CNN layer applies 128 filters with a kernel size of five and uses a ReLU activation function to capture local patterns in the text data. The output of the CNN layer is then processed by a GlobalMaxPooling1D layer to reduce its dimensionality.

- Capsule Network Layer:

Simultaneously, the 'sequence_output' is also passed to a Capsule Network layer. The Capsule layer consists of 10 capsules, each with 16 dimensions, and uses a routing mechanism with three iterations to capture spatial hierarchies and complex relationships between words. The capsule output is reshaped and then passed through a GlobalMaxPooling1D layer to further reduce its dimensionality.

- Combination Layer:

The outputs of the CNN and Capsule layers are concatenated to form a combined feature representation.

- Dense Layer with Dropout:

The combined features are fed into a Dense layer with 64 units and a ReLU activation function, followed by a Dropout layer with a rate of 0.2 to prevent overfitting.

- Output Layer:

The final output layer is a Dense layer with a single unit and a sigmoid activation function, which is suitable for binary classification (detecting hate speech or not).

- Training Configuration:

The model is compiled using the Adam optimizer with a learning rate of 1e-5, binary cross-entropy loss, and binary accuracy as the metric. Class weights are applied to handle class imbalance, and callbacks such as early stopping, learning rate reduction, and model checkpointing are used to improve training stability and performance. Figure 3 and 4 shows the architecture model that focus on CNN and Capsule Layer and the general architecture for the proposed model respectively.

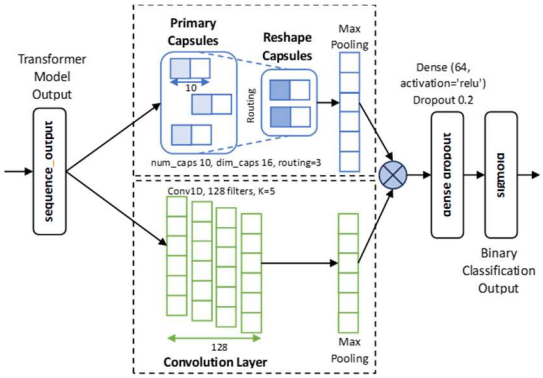

Figure 3: The model architecture on CNN and Capsule Network Layer

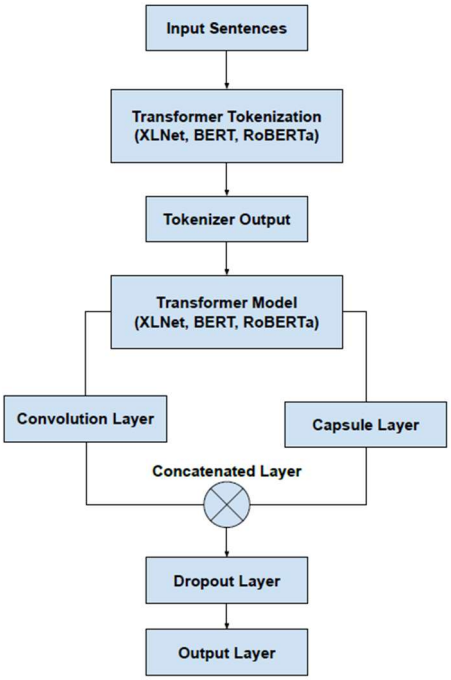

Figure 4: The proposed model architecture

The algorithm for proposed model is shown in Algorithm 1. This hybrid architecture is designed to effectively handle the nuances of Malay hate speech.

Algorithm 1 Proposed Algorithm

Input: dataset $W = \{w_1, w_2 \ldots w_n\}$
Output: evaluation metrics *(Accuracy, Precision, F1 score, Recall)*

#Preprocessing
train_data, val_data, test_data ← split (dataset)
For each tweet in dataset
  LowerCase (tweet)
  Remove (URL, symbols, number, Malay stop words from tweet)
end for

#Tokenization and Transformer Model
Token ← tokenization using XLNet/BERT/RoBERTa tokenizer (dataset)
  Token: input_ids, token type_ids, attention_mask
Sequence Output ← XLNet/BERT/RoBERTa model (token)

#CNN Layer
CNN Layer (sequence output)
  Conv1D ← Conv1D (*filters=128, kernel_size=5, activation='relu'*)
  Pool1 ← GlobalMaxPooling (Conv1D)

#CapsuleNetwork Layer
Capsule Layer (sequence output)
  Capsule ← CapsuleLayer (*num_capsules=10, dim_capsules=16, routing=3*)

```
14:    Capsule ← Reshape (target_shape= (-1, 10 * 16)) (capsule)
15:    Pool2 ← GlobalMaxPooling (Capsule)

#Concatenated, Dense, Dropout and Output Layer
16:  Combined (Pool1, Pool2)
17:  Dense ← Dense (64, activation = 'relu') (combined)
18:  Dropout ← Dropout (0.2) (dense)
19:  Output ← Dense (1, activation = 'sigmoid') (dropout)

#Compile the model and Define callbacks
20:  Compile (optimizer = Adam, loss = BinaryCrossentropy, metrics = BinaryAccuracy)
21:  Callbacks ← EarlyStopping, learning rate reduction, model checkpointing

#Train and Evaluate
22:  For each epoch in range (0,30)
23:    Train (train_data, val_data)
24:    Evaluate Accuracy and Val_Accuracy
25:  end for
26:  Test (test_data)
27:  Evaluate (Accuracy, Precision, F1 score, Recall)
28:  Return (Accuracy, Precision, F1 score, Recall)
```

3.2.3 Handling Imbalanced Data

Imbalanced data is a common issue in hate speech detection, where the number of hate speech instances may be significantly lower than non-hate speech instances. To address this, the model in this study uses class weights, which assign higher weights to less frequent classes to ensure the model pays more attention to them during training [28]. To address the imbalanced dataset, class weights are assigned to the loss function. Class weights make the model pay more attention to the minority class, giving higher importance to correctly classifying instances of this class. Additionally, early stopping, learning rate reduction, and model checkpointing are used to prevent overfitting and improve model performance, ensuring the model generalizes well to both majority and minority classes.

3.3 Experimentation

3.3.1 Environment Setup

This experimental research was conducted with careful consideration of both hardware and software components. The hardware includes a laptop with an Intel Core i7 processor, 32 GB DDR4 RAM, a 1 TB solid-state drive (SSD), and a 4GB GPU. The software setup consists of Python as the programming language, Scikit-Learn and TensorFlow as the frameworks, and Jupyter Notebook as the application platform.

3.3.2 Model Setup

The training process utilizes both existing and newly collected datasets, employing several optimization techniques. The learning rate is set to 1e-5, with a batch size of 32. Early stopping is implemented with patience set to 2 epochs, and dropout regularization is applied at a rate of 0.2. The Adam optimizer is used to enhance the training efficiency.

3.4 Evaluation Metrics

The performance metrics used in this experiment are the confusion matrix, accuracy, precision, recall, and F1 measure, as most previous researchers have used these measurements to compare the results [26].

Accuracy – The proportion of cases that are correctly classified. The best accuracy is 100%, indicating that all the predictions are correct.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision - Calculated proportion of cases predicted as good that is correct.

$$Precision = \frac{TP}{TP + FP}$$

Recall- Calculated by the ratio of the true positive to the addition of the true positive and false negative.

$$Recall = \frac{TP}{TP + FP}$$

F1-Measure - The total accuracy of the model or the classifier

$$F1\ measure = 2\frac{P \times R}{(P + R)}$$

4. RESULTS

This section presents all the results from the experimentation on each model for both datasets.

4.1 Model Performance

The proposed models were evaluated on both the existing HateM dataset and the newly collected dataset. The results demonstrate that the Transformer models, particularly when integrated with CNN and Capsule networks, perform significantly better than

existing approaches. Given the unbalanced nature of the datasets, the F1 score was chosen as the primary metric for model performance evaluation, as it provides a balanced measure that accounts for both precision and recall. Tables 2 and 3 show the results for whole evaluation metrics in both the HateM dataset and the new dataset, respectively. Figure 5 shows the comparison chart of the F1 score and accuracy for both datasets in all models.

Table 2: Results for HateM dataset

Model	Pre	Rec	F1	Acc
XLNet	78.56	75.89	76.64	75.78
BERT	78.98	76.35	77.12	76.23
RoBERTa	79.67	77.06	77.67	77.45

Table 3: Results for the new dataset

Model	Pre	Rec	F1	Acc
XLNet	82.51	82.33	82.40	82.33
BERT	83.58	83.75	83.59	83.74
RoBERTa	84.71	84.45	84.54	84.45

Although the BERT+CNN+Caps model shows good results compared to the other model XLNet+CNN+Caps, with an F1-score of 77.12% and accuracy of 76.23% for the HateM dataset and an F1-score of 83.59% and accuracy of 83.74% for the new dataset, the RoBERTa+CNN+Caps model still outperforms BERT+CNN+Caps in all evaluation metrics. The difference in F1-score and accuracy is 0.55% and 0.64%, respectively, in the HateM dataset and 0.95% and 0.71%, respectively, in the new dataset. When comparing the performance across these two datasets, it is clear that the new dataset's performance improvements are more distinct compared to the existing HateM dataset. This may indicate that the new dataset has characteristics that better influence the model's efficiency.

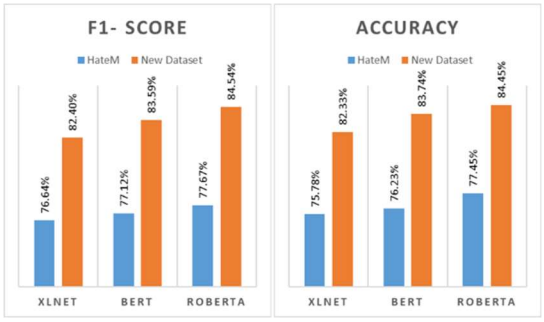

Figure 5: The comparison results on the F1 Score and Accuracy for each model on both datasets

4.2 Generalizability Analysis on Datasets

The evaluation of the newly collected dataset demonstrates that the Transformer-based models, particularly the RoBERTa+CNN+Caps model, generalize well to new data. The performance metrics on the new dataset are significantly higher compared to the existing HateM dataset. Specifically, the RoBERTa+CNN+Caps model in the new dataset achieves an F1-score of 84.54% and an accuracy of 84.45%, which are improvements of 6.87% and 7.0%, respectively, over the RoBERTa+CNN+Caps model in HateM dataset. This indicates that the models have effectively learned to detect hate speech in varying contexts and scenarios, making them robust and reliable for practical applications.

The robustness of the models is evident from their consistent performance across different datasets. The RoBERTa+CNN+Caps model outperforms not only other models on the HateM dataset but also shows significant improvements on the new dataset. This suggests that the model does not overfit the specific characteristics of the HateM dataset but is capable of generalizing to new, unseen data. This ability to generalize is crucial for real-world applications where the nature of hate speech can vary over time and across different social and cultural contexts. The consistent performance improvements, with a higher F1-score and accuracy on the new dataset, underscore the model's robustness and its potential effectiveness in diverse and dynamic environments.

5. FINDINGS AND DISCUSSION

The experimental results reveal several important insights. Among the models evaluated, RoBERTa+CNN+Caps consistently outperformed XLNet+CNN+Caps and BERT+CNN+Caps in all evaluation metrics, including precision, recall, F1-score, and accuracy. The highest F1-score and accuracy were observed with RoBERTa+CNN+Caps on the newly collected dataset, indicating its exceptional ability to detect hate speech in the Malay language.

The newly collected dataset, which included a broader range of keywords and modern socio-political contexts, yielded higher performance metrics compared to the existing HateM dataset. This underscores the importance of using diverse and representative datasets for training robust hate speech detection models.

The consistent improvement in performance metrics across different datasets indicates that the models, particularly RoBERTa+CNN+Caps, generalize well to new, unseen data. This suggests that the models are not merely memorizing the training data but are effectively learning the underlying patterns of hate speech in Malay.

While the RoBERTa+CNN+Caps model shows significant improvements in Malay hate speech detection (HSD), there are limitations that need to be considered. One key challenge is the dataset imbalance, with non-hate speech tweets greatly outnumbering hate speech tweets. This imbalance could limit the model's effectiveness in real-world scenarios where hate speech is rarer. To address this, class weights were applied to ensure the model focuses more on the minority class during training.

Another limitation is the diversity of the datasets. Although the new dataset covers a range of socio-political contexts, more diverse datasets are needed to test the model's generalizability across different cultural and linguistic environments. The evaluation metrics, especially the F1-score, were chosen to balance precision and recall, providing a more accurate measure of performance in the face of dataset imbalance, which is critical for identifying hate speech in low-resource languages.

6. CONCLUSION AND FUTURE RESEARCH

The study successfully developed and evaluated advanced HSD models for the Malay language, with the RoBERTa+CNN+Caps model showing the best performance across all metrics. The new dataset, which was more diverse and representative of contemporary socio-political contexts, resulted in higher performance metrics, underscoring the importance of dataset diversity. The models also demonstrated strong generalizability, maintaining high performance on unseen data, which is crucial for real-world applications. While the results highlight the strengths of the proposed model, such as its ability to generalize well across different datasets and its strong performance on unseen data, certain limitations exist. The primary challenge lies in the dataset's imbalance, which could affect the model's ability to handle varied real-world scenarios. Additionally, the computational complexity of the model may limit its scalability for real-time applications.

Despite these challenges, the model represents a significant step forward in hate speech detection for low-resource languages. Its ability to adapt to the complexities of Malay hate speech demonstrates its potential for broader applications, serving as a foundation for future advancements in HSD models for other underrepresented languages. Future research should focus on expanding the dataset to include more diverse and balance samples to improve the model's generalizability. Additionally, optimizing the model to reduce its computational complexity could make it more suitable for real-time applications.

REFERENCES:

[1] Maity, K., Bhattacharya, S., Saha, S., & Seera, M. (2023). A Deep Learning Framework for the Detection of Malay Hate Speech. IEEE Access, 11, 79542–79552. https://doi.org/10.1109/access.2023.3298808

[2] Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal, 80, 110–121. https://doi.org/10.1016/j.aej.2023.08.038

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

[4] Schulze, E. (2019, February 4). EU says Facebook, Google and Twitter are getting faster at removing hate speech online. CNBC. Retrieved January 19, 2024, from https://www.cnbc.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds--.html

[5] Maskat, R., Zainal, M. F., Ismail, N., Ardi, N., Ahmad, A., & Daud, N. (2020). Automatic Labelling of Malay Cyberbullying Twitter Corpus using Combinations of Sentiment, Emotion and Toxicity Polarities. https://doi.org/10.1145/3446132.3446412

[6] Mehmood, I., Anwar, S., AneezaDilawar, N., IsmaZulfiqar, N., & Abbas, R. M. (2020). Managing Data Diversity on the Internet of Medical Things (IoMT). International Journal of Information Technology and Computer Science, 12(6), 49–56. https://doi.org/10.5815/ijitcs.2020.06.05

[7] Guterres, A. (2019). UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH. Retrieved June 21, 2024, from https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

[8] Nortajuddin, A. (2020, July 22). Does Malaysia Have A Cyberbullying Problem? The ASEAN Post. Retrieved January 19, 2024, from https://theaseanpost.com/article/does-malaysia-have-cyberbullying-problem

[9] Azman, N. F., & Zamri, N. A. K. (2022). Conscious or Unconscious: The Intention of Hate Speech in Cyberworld—A Conceptual Paper. International Academic Symposium of Social Science 2022. https://doi.org/10.3390/proceedings2022082029

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

[11] Firmino, A. A., de Souza Baptista, C., & de Paiva, A. C. (2024). Improving hate speech detection using Cross-Lingual Learning. Expert Systems With Applications, 235, 121115. https://doi.org/10.1016/j.eswa.2023.121115

[12] Liu, L., Xu, D., Zhao, P., Zeng, D. D., Hu, P. J. H., Zhang, Q., Luo, Y., & Cao, Z. (2023). A cross-lingual transfer learning method for online COVID-19-related hate speech detection. Expert Systems With Applications, 234, 121031. https://doi.org/10.1016/j.eswa.2023.121031

[13] Pérez, J. M., Luque, F. M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P. S., . . . Cotik, V. (2023). Assessing the Impact of Contextual Information in Hate Speech Detection. IEEE Access, 11, 30575–30590. https://doi.org/10.1109/access.2023.3258973

[14] Agarwal, S., Sonawane, A., & Chowdary, C. R. (2023). Accelerating automatic hate speech detection using parallelized ensemble learning models. Expert Systems With Applications, 230, 120564. https://doi.org/10.1016/j.eswa.2023.120564

[15] Roy, P. K., Bhawal, S., & Subalalitha, C. N. (2022). Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Computer Speech & Language, 75, 101386. https://doi.org/10.1016/j.csl.2022.101386

[16] Keya, A. J., Kabir, M. M., Shammey, N. J., Mridha, M. F., Islam, M. R., & Watanobe, Y. (2023). G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media. IEEE Access, 11, 79697–79709. https://doi.org/10.1109/access.2023.3299021

[17] Al-Hashedi, M., Soon, L. K., Goh, H. N., Lim, A. H. L., & Siew, E. G. (2023). Cyberbullying Detection Based on Emotion. IEEE Access, 11, 53907–53918. https://doi.org/10.1109/access.2023.3280556

[18] Putra, C. D., & Wang, H. C. (2024). Advanced BERT-CNN for Hate Speech Detection. Procedia Computer Science, 234, 239–246. https://doi.org/10.1016/j.procs.2024.02.170

[19] Yigezu, M., Kolesnikova, O., Sidorov, G., & Gelbukh, A. (2023). Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification. https://api.semanticscholar.org/CorpusID:265309481

[20] Mutanga, R. T., Naicker, N., & O, O. (2020). Hate Speech Detection in Twitter using Transformer Methods. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 11(9). https://doi.org/10.14569/ijacsa.2020.0110972

[21] Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: a multi-label hate speech detection dataset. Complex & Intelligent Systems, 8(6), 4663–4678. https://doi.org/10.1007/s40747-021-00608-2

[22] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1. https://doi.org/10.48550/arXiv.1810.04805

[23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.11692

[24] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. https://doi.org/10.48550/arXiv.1906.08237

[25] Jiang, A., & Zubiaga, A. (2021). Cross-lingual Capsule Network for Hate Speech Detection in

Social Media. https://doi.org/10.1145/3465336.3475102

[27] of Art, Challenges and Future Work. IEEE Access, 8, 24687–24696. https://doi.org/10.1109/access.2020.2968955

[28] Stopwords-Iso. (n.d.). stopwords-ms/stopwords-ms.txt at master · stopwords-iso/stopwords-ms. GitHub. https://github.com/stopwords-iso/stopwords-ms/blob/master/stopwords-ms.txt

[29] Singh, K. (2023, July 6). How to Improve Class Imbalance using Class Weights in Machine Learning? Analytics Vidhya. Retrieved May 29, 2024, from https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

[30] Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits. IEEE Access, 9, 154704–154716. https://doi.org/10.1109/access.2021.3128742

[26] Bakar, M. F. R. A., Idris, N., Shuib, L., & Khamis, N. (2020). Sentiment Analysis of Noisy Malay Text: State