UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF
INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

An Effective Cyberbullying Detection Model for the Malay Language Using Transformer Model in Social Media Platform X

Savinder Singh[1]* & Siti Hajar Othman[2]
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
Email: savindersingh@graduate.utm.my[1]; hajar@utm.my[2]

Submitted: 20/2/2025. Revised edition: 7/5/2025. Accepted: 15/5/2025. Published online: 27/5/2025
DOI: https://doi.org/10.11113/ijic.v15n1.520

***Abstract*—In the digital era, social media has transformed communication but has also facilitated cyberbullying, significantly affecting individuals' mental health, particularly within the Malay-speaking community. Despite the growing concerns, developing effective cyberbullying detection systems for low-resource languages like Malay has been limited. This research addresses this gap by introducing a transformer learning model specifically designed for detecting cyberbullying in Malay language tweets. The work begins with an extensive literature review to consolidate the current understanding of cyberbullying detection techniques. A substantial dataset will be curated from X and manually annotated, forming the basis for model training and evaluation. The research employs machine learning models and transformer models to improve overall detection accuracy and robustness. Advanced NLP techniques, including transformer models and transfer learning, will be utilized to navigate the complexities of the Malay language and ensure accurate detection. The proposed model's performance will be rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score with tests to ensure its robustness against different cyberbullying tactics. This research has produced a high-performing detection model that enhances the safety of Malay-speaking internet users and provides insights into cyberbullying dynamics within the community with advanced NLP for low-resource languages. The results demonstrate that deep learning Transformer models, particularly DeBERTa, have outperformed traditional machine learning models in accuracy for Malay language cyberbullying detection. Among all models tested the DeBERTa transformer model achieved a high F1 score (0.8277), Accuracy (0.8380), precision (0.8202) and recall (0.8454) in terms of cyberbullying detection and F1 score (0.6813)), Accuracy (0.7352), precision (0.6643) and recall (0.7335) for sentiment analysis which significantly outperforming traditional machine learning baselines. These results demonstrate the effectiveness of Transformer architecture in capturing linguistic nuances in Malay and confirm their superiority for cyberbullying detection. This research contributes to improving online safety for Malay-speaking users and advances the development of NLP tools for low-resource languages.**

***Keywords*—Cyberbullying, Malay Language, Transformer Model, Machine Learning**

I. INTRODUCTION

Cyberbullying refers to the practice in which individuals use electronic devices to threaten, tease, harass, or in any way cause discomfort to others, particularly within cyberspace. It simply means a criminal act that is often planned and directed towards an individual with the aim of causing them anguish or humiliation. Sharing provocative and slanderous texts, publishing personal data, impersonation, other threats, or publishing materials that are not suitable for open publication are examples of cyberbullying [1], and are often characterized as deliberate and repeated harm inflicted by, for instance, using computers, cell phones, and other electronic gadgets. However, not all abusive conduct among teenagers falls under the category of bullying. Firstly, bullying has to be driven deliberately rather than occurring unintentionally. Essentially, bullying entails harm caused to other people with the intention of doing it and in a continuous manner, while accidental harm is not considered bullying. For instance, omission can be defined as ignoring someone unintentionally, which does not qualify as bullying. If the threat is repeated, then that constitutes bullying.

Secondly, harassment involves repeated and intentional behavior that violates social or personal boundaries. This distinguishes it from isolated or accidental incidents. For example, a single disagreement or argument does not constitute bullying even if it happens more than once. However, when someone is consistently insulted, threatened, or physically harmed, it is considered bullying. The key distinction lies in the deliberate intent and the repeated nature of the harmful actions.

Malaysian Communications and Multimedia Commission (MCMC) has conducted a survey between 2018 and January 2020, in Malaysia and they documented a total of 3,762 instances of cyberbullying. Nevertheless, this figure does not reflect the full extent of cyberbullying incidents in the country, as many cases remain unreported. Victims might opt not to report these incidents due to feelings of distress, shame, or ignorance about where to seek help [2]. Consequently, experts believe that the actual number of cyberbullying cases is significantly higher than what's reported [2]. Addressing this concern requires media platforms and other online systems to incorporate efficient measures for identifying and preventing cyberbullying occurrences.

To address this issue, recent advancements in machine learning (ML) and deep learning with Transformer-based models like BERT, DistilBERT, and DeBERTa offer significant improvements in natural language processing (NLP). These models are capable of understanding contextual information in text, making them suitable for detecting harmful language patterns on social media.

Motivated by the lack of effective tools for cyberbullying detection in low-resource languages like Malay, this paper proposes a Transformer-based deep learning model trained on Malay-language tweets. It aims to evaluate and compare the performance of Transformer models against traditional machine learning methods.

II. RELATED WORKS

Cyberbullying is now recognized as a growing issue among frequent internet users. Its effects have gained increasing attention from researchers in recent years, including studies by Ahmad *et al.* [3] and Chan *et al.* [4]. These studies primarily examined the psychological impact of cyberbullying, its prevalence, and strategies for intervention. However, with advancements in artificial intelligence, current research is shifting toward the automatic detection and mitigation of cyberbullying. This evolution is largely driven by developments in Natural Language Processing (NLP) and machine learning methods, which enable automated detection and monitoring of online interactions [5].

In a research study, Emon *et al.* [6] investigated the detection of cyberbullying through the adoption of machine learning and deep learning models based on Transformer technology. They employed both Count Vectorizer and Term Frequency–Inverse Document Frequency (TF-IDF) vectorization methods, along with a mix of machine learning and deep learning algorithms to identify abusive language in Bengali content. The study gathered a dataset comprising 4,000 comments extracted from platforms like Facebook, YouTube, and the news website Prothom Alo. The dataset was divided into seven categories: slang terms, expressions of animosity, personal insults, and breaches of political decorum, anti-feminist rhetoric, positive statements, and neutral remarks. Among all the tested models, the Recurrent Neural Network (RNN) performed the best, achieving an accuracy of 0.8220. This result highlights the effectiveness of RNN in identifying abusive text.

Romim *et al.* [7] conducted a study on identifying hate speech in Bengali by employing networks like Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM). They used word embeddings trained using FastText, Word2Vec, and GloVe to improve textual representation. The study assessed how different deep learning models and word embedding methods performed. Interestingly, the Support Vector Machine (SVM) model produced the best result with an accuracy of 0.875, outperforming the deep learning models.

Islam *et al.* [8] conducted research to detect abusive comments using a dataset collected from Facebook and YouTube. To improve classification accuracy, they applied several machine learning methods such as Multinomial Naïve Bayes (MNB), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree, Random Forest, Stochastic Gradient Descent (SGD), Ridge Classifier, Perceptron, and k-Nearest Neighbors (KNN). To manage class imbalance, they used undersampling and a Bengali stemmer for preprocessing. Their findings showed that the SVM model achieved the highest accuracy, scoring 0.88 and outperforming other classifiers.

Aurpa *et al.* [9] explored Transformer-based models to detect offensive comments on Facebook. They applied models like BERT and ELECTRA to a dataset of 44,000 comments, aiming to improve classification accuracy. During evaluation, the BERT-based model achieved a testing accuracy of 85%, while ELECTRA reached 84.92%, demonstrating the effectiveness of Transformer models in detecting abusive content.

Van Hee *et al.* [10] studied cyberbullying detection in Dutch by analyzing posts from the Ask.fm platform. The dataset was categorized into content types such as profanity (curse), defamation, encouragement, insult, sexual content, threat, and non-cyberbullying content. To improve classification, they extracted features such as word n-grams, subjectivity lexicons, character n-grams, term lists, and applied topic modeling. These features were used to train a modified Support Vector Machine, and the integration of these characteristics significantly improved the F-score for both Dutch and English datasets—demonstrating the value of feature engineering across languages.

In the research by H. Margono *et al.* [11], Indonesian cyberbullying was analyzed using a dataset from platform X that contained offensive language. The data was divided into two categories: cyberbullying and non-cyberbullying. Classification algorithms including Naïve Bayes (NB), Decision Trees (DT), and Neural Networks (NN) were tested. Their findings showed that using Naïve Bayes achieved 100% accuracy in correctly distinguishing between cyberbullying and non-cyberbullying instances.

A. Research on Cyberbullying for the Malay Language

While several Malaysian researchers have explored the detection of cyberbullying on local social media platforms, very few have focused specifically on machine learning approaches for detecting cyberbullying in the Malay language.

In a study by Zainol *et al*. [12], the researchers examined cyberbullying terminology in Malay tweets using the Apriori algorithm for association analysis. Their work aimed to uncover common cyberbullying terms through association rules derived from frequently used words in tweets. They also applied clustering techniques to enhance the identification of cyberbullying instances in Malay-language content, highlighting the importance of pattern recognition methods in this domain.

According to research by T. K. Hua *et al.* [13], cyberbullying in Malay tweets can be categorized into three main themes: attacks based on intelligence, physical appearance, and sexual orientation.

R. Maskat *et al.* [14] proposed using a combination of sentiment, emotion, and toxicity polarity scores to automatically categorize a corpus of Malay-language cyberbullying tweets. Due to the impracticality of manual labeling for large datasets, they utilized the Bidirectional Encoder Representations from Transformers (BERT) model to develop an automatic labeling approach. The method was based on detecting emotional polarities such as anger, negative sensitivity, and toxicity.

In a study involving 45,580 imbalanced tweets, Nurina Farhanah Binti Johari *et al*. [15] evaluated the effectiveness of several machine learning models, including Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). These models were tested with three feature extraction techniques: Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Word2Vec embeddings. Their findings indicated that the combination of Logistic Regression and TF-IDF yielded the best performance. To improve model robustness, they applied Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and fine-tuned the hyperparameters. The optimized LR + TF-IDF model achieved an F1-score of 0.46, demonstrating its potential in detecting cyberbullying in Malay tweets.

III. METHODOLOGY

This paper utilizes learning methods, with Transformer based models like DeBERTa, BERT, DistilBERT and XLM R for identifying instances of cyberbullying in tweets written in Malay language. These models have been customized for this purpose by utilizing existing knowledge from diverse datasets which aid them in comprehending intricate sentence formations as well as linguistic nuances, in Malay language.

A comparison between machine learning models will be executed too to ensure we can conclude the best approach to detect cyberbullying for the Malay Language.

A. Dataset Collection and Labelling Process

Fig. 1 illustrates the overall methodology employed for dataset collection and labelling, providing a summary of the steps taken to gather, preprocess, and annotate the Malay-language tweets for cyberbullying and sentiment analysis.'

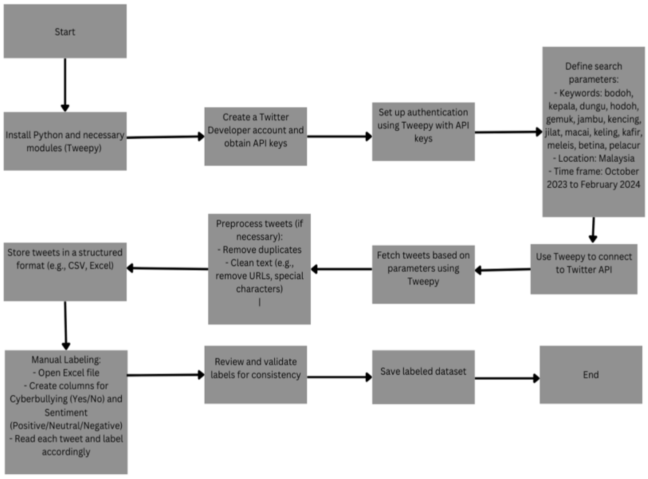
Start
Install Python and necessary modules (Tweepy)
Create a Twitter Developer account and obtain API keys
Set up authentication using Tweepy with API keys
Define search parameters:
- Keywords: bodoh, kepala, dungu, hodoh, gemuk, jambu, kencing, jilat, macai, keling, kafir, meleis, betina, pelacur
- Location: Malaysia
- Time frame: October 2023 to February 2024
Use Tweepy to connect to Twitter API
Fetch tweets based on parameters using Tweepy
Preprocess tweets (if necessary):
- Remove duplicates
- Clean text (e.g., remove URLs, special characters)
|
Store tweets in a structured format (e.g., CSV, Excel)
Manual Labeling:
- Open Excel file
- Create columns for Cyberbullying (Yes/No) and Sentiment (Positive/Neutral/Negative)
- Read each tweet and label accordingly
Review and validate labels for consistency
Save labeled dataset
End

Fig. 1. Dataset collection methodology

1) Installing Dependencies

The process begins with setting up the environment by: Installing Python and necessary modules, such as Tweepy, which is used to access the X API.

2) Creating a X Developer Account

To access the X API for collecting tweets an X Developer account is required in order to obtain the API keys.

3) Setting Up Authentication

With the utilization of Tweepy tool, the API keys for authentication purposes are employed to connect to Xs API and retrieve tweets.

4) Defining Search Parameters

This paper delves into the use of disrespectful language often linked to bullying or mistreatment, in the Malay language. These terms consist of words as bodoh (stupid), kepala (head), dungu (dumb), hodoh (ugly), gemuk (fat), jambu (pretty boy), kencing (lie), jilat (lick), macai (lackey), keli*g (a derogatory term for Indians), kafir (infidel), meles (a slur against Malays), betina (a derogatory term for females), and pelac*r (prostitute). In discussions related to bullying or harassment these terms are commonly used, playing a role in identifying cyberbullying within the Malay language setting.

The main emphasis of this study is directed towards Malaysia’s region where the Malay language is predominantly utilized in cultural settings as a vital element, for communication among individuals in society. It is crucial to grasp the aspects along, with the circumstances surrounding the application of these expressions to develop precise cyberbullying detection models that align with the culture.

The dataset collection for this study took place between October 2023 and February 2024 to gather, up to date data that reflects the trends and communication patterns, in Malaysia accurately.

5) Fetching Tweets Using Tweepy

Once the parameters are set, tweets matching the specified criteria are collected using Tweepy. This ensures that only relevant tweets within the specified time frame and region are gathered.

6) Preprocessing the Tweets

The tweets that was collected will go through a preparation stage to improve the quality and significance of the dataset used for analysis purposes. It all kicks off with getting rid of any repeated tweets to prevent any information from skewering the results or hindering how well the model works. By keeping entries the dataset becomes a better reflection and easier to handle. This ultimately boosts efficiency in handling data and accuracy, in identifying cyberbullying patterns.

After that comes the step of cleaning the text to get rid of any information that might affect the analysis and performance of the model. This means getting rid of any URLs since they usually don't add value and aren't related to the language patterns being studied. Furthermore all special characters and extra elements are taken out to make sure the text is standardized. These initial steps help fine tune the dataset by focusing linguistic features needed for detecting cyberbullying. This phase of preparation plays a role, in enhancing both the accuracy and efficiency of the machine learning models being used in this research.

7) Storing Tweets in Structured Format

The prepped tweets are saved in organized forms, like CSV or Excel files to make sure they're well organized and easy to access for analysis and tagging purposes. This structured storage method makes it simple to connect with AI workflows for data handling and processing, for identifying cyberbullying and categorizing sentiments

8) Manual Labeling

Table 1 shows the sample of the dataset labelling process. This process consists of a very important step of labeling which is a part of the preparation of the dataset for analysis. This process begins with the dataset being opened in its most basic form which is usually in the form of an Excel spreadsheet in order to help with the labelling process. Important information about each tweet is stored in columns, which are created.

One column is created to determine whether there is cyberbullying or not, called ‘Cyberbullying (Yes/No)’. Using this, the annotators are able to determine if a tweet is cyberbullying or not based on certain guidelines. Another column is set for the purpose of sentiment analysis with the label ‘Sentiment (Positive/Neutral/Negative)’ to determine the general emotional message of the tweet.

Using a dual label method not only aids in identifying cyberbullying instances it also offers deeper understanding of the emotional changes, in the data set to allow for a more thorough examination of cyberbullying patterns in the Malay language by reviewing individual tweets, for suitable categorization based on their content.

TABLE I. DATASET SAMPLE

Tweet Content	**Cyberbullying**	**Sentiment Analysis**
betina ni ajak tunggang airmata dan darah palestine demi nak lawan Anwar Ibrahim kononnya.. menyalak kat twitt..aku harap kau bukan lesbi....	Yes	Negative
But*h beti*a gem*k ni patut mampus.	Yes	Negative
Koya nyaaaaa muka beti*a ni. Macam natang	Yes	Negative

9) Reviewing and Validating Labels

The dataset contains an ID column, a text column, and a label column. The labels are then reviewed and validated to make sure that they are consistent and accurate throughout the dataset.

10) Dataset Size

The total dataset collected consists of 19,170 sentences. This dataset forms the core input for training and evaluating the deep learning models for cyberbullying detection.

11) Saving the Labeled Dataset

After validation, the labeled dataset is saved for further use in training deep learning models.

B. Deep Learning Transformer Model Process

Fig. 2 shows the pipeline of the deep learning transformer model which was utilized in this experiment.

1) Create Hugging Face Dataset

Both Sentiment and Cyber datasets with their splits are converted to Hugging Face DatasetDict containing Dataset instance for each split. This will be used in further processing and training using Hugging Face.

2) Tokenize Datasets and Encode Labels

The Datasets are tokenized, and labels are encoded to integers to use them effectively by HF Trainer to train models.

3) Finetune Transformers (Deberta, Bert, xlm_roberta, Distilbert)

The training process for the machine learning model utilized specific hyper parameters to optimize its performance and ensure accurate detection of cyberbullying. The number of training epochs was set to 4, meaning the entire dataset was passed through the model four times during the training phase. This choice balances sufficient learning with the avoidance of overfitting, where the model becomes overly tailored to the training data.

A learning rate of 2e-5 was employed, representing the step size for adjusting the model's weights during gradient descent. This relatively small value ensures gradual and stable convergence, reducing the risk of overshooting the optimal solution. Additionally, the batch size for both training and evaluation was set to 64 per device, which provides a balance between computational efficiency and model performance by processing a moderate number of examples at once.

In order to standardize the model and avoid overfitting issues a weight decay of 0.01 was utilized, this approach penalizes weights in the model encouraging more adaptable solutions. The assessment method was designated as "epoch" implying that the model was evaluated on the validation dataset following each training epoch. This allows for tracking performance enhancements and identifying possible problems during training.

Furthermore, mixed precision training was also employed with fp16 = True, which is 16 bit floating point precision. This paper employs this approach to increase the training speed and decrease memory consumption without sacrificing the model accuracy. In addition, these training arguments were chosen and set to the right values to optimize the process of learning the model for the cyberbullying detection task in the Malay language dataset.

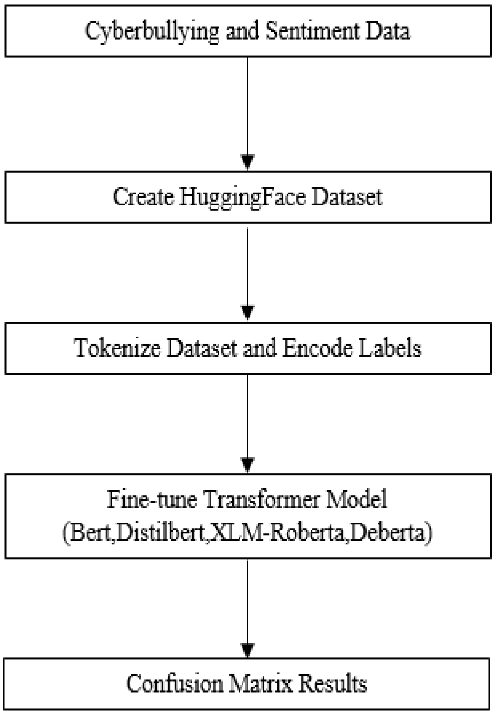
Cyberbullying and Sentiment Data
Create HuggingFace Dataset
Tokenize Dataset and Encode Labels
Fine-tune Transformer Model (Bert,Distilbert,XLM-Roberta,Deberta)
Confusion Matrix Results

Fig. 2. Pipeline for deep learning transformer model

C. Machine Learning Model Pipeline

Fig. 3, illustrates the process flow of the machine learning model employed in this study to carry out the comparative analysis with the transformer model.

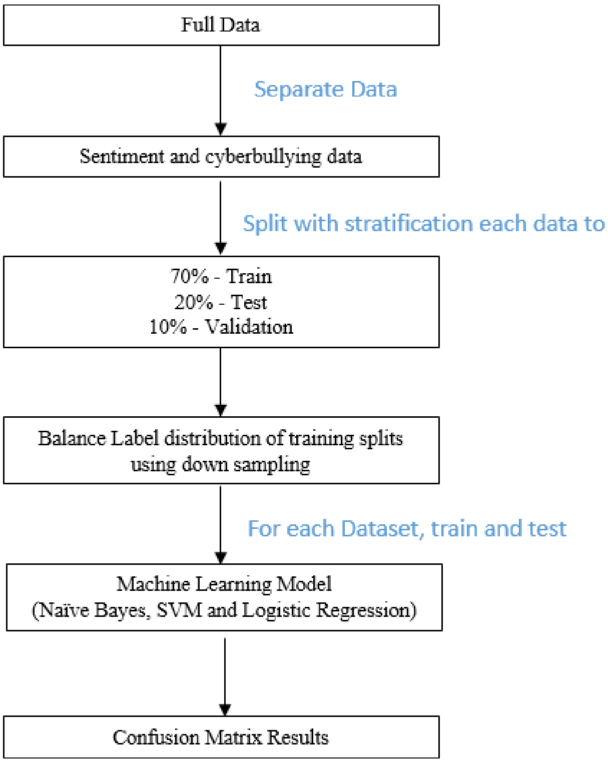
Full Data
Separate Data
Sentiment and cyberbullying data
Split with stratification each data to
70% - Train
20% - Test
10% - Validation
Balance Label distribution of training splits using down sampling
For each Dataset, train and test
Machine Learning Model (Naïve Bayes, SVM and Logistic Regression)
Confusion Matrix Results

Fig. 3. Pipeline for machine learning model

The procedure starts with a collection of data that includes information on both cyberbullying and sentiment analysis content. This data is separated into two categories. Sentiment Data and cyber Data. For processing purposes. The data within each category is subsequently divided into training (70%) testing (20%) and validation (10%) subsets using a method called stratification to maintain a distribution of classes. To tackle any disparities in class representation, within the training dataset a technique known as down sampling is employed to ensure that all classes are equally represented.

In both datasets three classic machine learning models such, as Support Vector Machine SVM) Naïve Bayes and Logistic Regression are set up and tested. These models are selected for their speed and accuracy in text categorization. The performance of the models is checked using metrics like Accuracy, Precision, Recall and F1 score, with confusion matrices created to offer a view of how the classification is done. The outcomes are stored for examination and comparison.

D. Performance Evaluation

The evaluation of machine and transformer models would also involve monitoring their performance by means of accuracy, precision, recall and the F1 score. Accuracy is defined by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [17]$$

Where TP represents true positive, TN denotes true negative, FP signifies false positive, and FN represents false negative.

Precision is a measure of performance that calculates the ratio of positives to the total number of positive predictions. It is expressed as follows:

$$Precision = \frac{TP}{TP + FP} \quad [18]$$

The recall metric is also used for measuring the model performance and is defined by the formula:

$$Recall = \frac{TP}{TP + FN} \quad [19]$$

In cases where there is an imbalance, in classes, within the data set the F score serves as a metric since it takes into account both precision and recall to offer an evaluation of the models performance. The computation of the F score involves using a formula.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [20]$$

IV. Results and Discussion

A. Deep Learning Transformer Model Implementation (Libraries and Tools)

For this paper, the deep learning implementation was performed using Python, which has broad support for machine learning and natural language processing (NLP) tasks. This, together with the flexibility and the seamless integration of Python with deep learning frameworks, made it the right choice for using state-of-the-art models.

The Hugging Face Transformers library was used to access and adjust trained Transformer models, like BERT, DistilBERT, XLM RoBERTa (XLM-R) and DeBERTa to identify cyberbullying and analyze the sentiment of tweets in Malay. The library also provides a structure for customizing Transformer models by adding classification heads and making necessary modifications for this work

The Hugging Face Datasets library was used for loading, processing and organizing the datasets. This library is compatible with the Hugging Face Transformers library which makes the pipeline smooth for performing tasks like tokenization and label encoding for NLP. It will help in effectively handling the data related tasks that occurred during training and evaluation, especially for large datasets.

Real-time tweets were collected using Tweepy, a Python wrapper for the X API. This tools will help to collect tweets based on particular keywords related to cyberbullying and sentiment, so that the dataset would be relevant and up-to-date. The data was important in helping to build a reliable training corpus for fine-tuning the model.

PyTorch serves as the foundation, for enabling smooth and effective model refinement within the Hugging Face library framework it allows for exploration of models through adaptable modifications, to training settings and structure adjustments to achieve the best possible results.

During the data preprocessing stage for handling missing data and cleaning text inputs, with structured datasets transformation from raw tweet data, for model training and evaluation—a combination of Pandas and NumPy was utilized extensively.

Matplotlib and Seaborn were used for visualizing the results and plots like confusion matrices were created. Using these visualizations provided key insights into the models' performance and allowed for an in-depth comparison of their effectiveness.

Sklearn was employed for computing the evaluation metrics of the models, including accuracy, precision, recall, and F1-score, to evaluate the effectiveness of the models in identifying cyberbullying. Furthermore, confusion matrices were created to provide a more detailed analysis of model performance and help visualize misclassification patterns.

B. Experimental Set Up

This section describes the dataset, parameters, and processes that were used to fine tune and evaluate the transformer models for cyberbullying detection and sentiment analysis of Malay language tweets.

In this work, the dataset consists of 19,170 tweets, all in Malay Language. Using the Tweepy library, these tweets were collected from X. The time of collection spans from October 2023 to February 2024, and the keywords used are specific to cyberbullying and abusive behavior. Two tasks were performed for each tweet: cyberbullying detection and sentiment analysis. In the first task, tweets were labeled as Cyberbullying (Yes) or Non-Cyberbullying (No). In the second task, the tweets were assigned a class label corresponding to the sentiment of the tweet as Positive, Neutral or Negative.

The datasets were divided into three subsets for reliable evaluation of the models. The data was split into 70-20-10 for training, testing and validation, respectively. This split maintained the class distribution across the datasets, so that there was no class imbalance and the models were trained and evaluated on fairly representative data. Several hyperparameters were used during the fine tuning of the hyperparameters to fine tune the performance of the models. The learning rate was set to 2e-5 and the batch size was set to 16 with the number of epochs being 5. The weight decay was employed by using the AdamW optimizer to update the model weights during the training process. In addition, 500 warm-up steps were applied to gradually increase the learning rate, and the maximum sequence length for each input was set to 128 tokens to ensure consistency. Binary Cross-Entropy Loss was used for the cyberbullying detection task, while Categorical Cross-Entropy Loss was employed for the sentiment analysis task. To prevent overfitting, early stopping was implemented, halting training if the validation performance did not improve after several epochs.

The Hugging Face trainer API with PyTorch as the backend was used to fine-tune the Transformer models. The tweets were first tokenized using AutoTokenizer which split the text into subword tokens and added special tokens [CLS] and [SEP]). In this work, the following pre-trained Transformer models were fine-tuned: BERT, DistilBERT, XLM-RoBERTa (XLM-R) and DeBERTa. Each model was trained separately for the two tasks: a binary cyberbullying detection task and a multi-class sentiment analysis task. The training was conducted on NVIDIA GPU hardware to increase efficiency and decrease computation time.

Following adjustments and refinements were made to the models before assessing them on a subset of 20%. The evaluation criteria encompassed accuracy along, with precision and recall metrics to gauge performance comprehensively through the lens of strengths and limitations, in each models design nuances. Moreover the analysis delved into confusion matrices which shed light on the models predictive capabilities by distinguishing between positives and false positives.

The outcomes of each model such, as BERT, DistilBERT, XLM RoBERTa and DeBERTa were examined then compared against Machine Learning models. The assessment standards along with the confusion matrices were archived for future examination the models that performed the best were identified by looking at their F1 scores along with metrics offering understanding into how well they function in detecting cyberbullying and analyzing sentiments, in Malay language tweets.

This unique configuration guaranteed a foundation, for assessing performance consistently while showcasing the capabilities of Transformer models in detecting content and assessing sentiments on Malay language social platforms. The data slpit was carefully selected to offer a rounded method for training models that can adapt effectively to data sets.

TABLE II. CYBERBULLYING DETECTION RESULTS (TRANSFORMER MODEL)

Model	F1	Accuracy	Precision	Recall
Deberta	0.827693	0.838028	0.820193	0.845390
Bert	0.812238	0.822379	0.805629	0.832405
XLM roberta	0.822854	0.832812	0.815695	0.842349
Distillbert	0.796127	0.807512	0.789865	0.814496

Based on the results in Table II, among the individual models, DeBERTa and XLM-RoBERTa perform comparably well. DeBERTa achieves an F1-score of 0.8277 with high recall (0.8454), indicating it captures most true cyberbullying instances. XLM-RoBERTa performs slightly lower, but it's F1-score of 0.8229 and multilingual capabilities make it highly effective for Malay-language tweets.

BERT performs consistently with an F1-score of 0.8122 and accuracy of 0.8224, although it lags behind DeBERTa and XLM-RoBERTa in recall, suggesting it might miss some true cyberbullying cases. DistilBERT, while offering faster inference shows the lowest performance among the individual models, with an F1-score of 0.7961 and accuracy of 0.8075.

TABLE III. SENTIMENT ANALYSIS RESULTS (TRANSFORMER MODEL)

Model	F1	Accuracy	Precision	Recall
Deberta	0.681324	0.735263	0.664356	0.733502
Bert	0.645969	0.694575	0.627922	0.709089
XLM roberta	0.659984	0.701356	0.646180	0.730585
Distillbert	0.623004	0.670318	0.610320	0.685859

In Table III sentiment analysis, the performance of the models is generally lower than in cyberbullying detection, which could indicate that classifying sentiment in Malay-language tweets is more challenging.

Among the individual models, DeBERTa achieves the best performance, with an F1-score of 0.6813 and accuracy of 0.7353. Its ability to capture nuanced language patterns helps it perform well across different sentiment categories. XLM-RoBERTa follows closely with an F1-score of 0.6599, benefiting from its multilingual capabilities.

BERT shows a decline in performance compared to its results for cyberbullying detection, achieving an F1-score of 0.6460 and accuracy of 0.6946. Similarly, DistilBERT shows the lowest performance, with an F1-score of 0.6230 and accuracy of 0.6703, reflecting the challenges in handling sentiment analysis effectively.

C. Comparison Between Machine Learning and Deep Learning

In this section we did a comparison between the results of machine learning and deep learning to ensure that we can determine the best technique and model to detect cyberbullying for the Malay Language as shown in the table below.

TABLE IV. CYBERBULLYING DETECTION RESULTS COMPARISON BETWEEN MACHINE LEARNING AND DEEP LEARNING)

Model	F1	Accuracy	Precision	Recall
Naïve Bayes	0.743371	0.748826	0.753045	0.781162
Logistic Regression	0.805353	0.817684	0.798504	0.820379
SVM	0.794285	0.806208	0.787924	0.811482
Deberta	0.827693	0.838028	0.820193	0.845390
Bert	0.812238	0.822379	0.805629	0.832405
XLM roberta	0.822854	0.832812	0.815695	0.842349
Distillbert	0.796127	0.807512	0.789865	0.814496

Based on Table IV the deep learning models (DeBERTa, BERT, XLM-RoBERTa, and DistilBERT) outperform traditional machine learning models (Naïve Bayes, Logistic Regression, and SVM) in F1-score, accuracy, precision, and recall. This indicates that Transformer-based models are more effective in capturing the complex patterns in text data, especially for identifying cyberbullying behavior.

DeBERTa achieves the highest individual model performance, with an F1-score of 0.8277 and recall of 0.8454, suggesting it is highly effective in identifying true cyberbullying cases.

Among the machine learning models, Logistic Regression performs best, achieving an F1-score of 0.8054 and accuracy of 0.8177, but it still falls short compared to the deep learning models. Naïve Bayes shows the weakest performance, with an F1-score of 0.7434, indicating it struggles.

TABLE V. SENTIMENT ANALYSIS RESULTS COMPARISON BETWEEN MACHINE LEARNING AND DEEP LEARNING

Model	F1	Accuracy	Precision	Recall
Naïve Bayes	0.600471	0.639541	0.599823	0.675938
Logistic Regression	0.617099	0.655451	0.610290	0.691745
SVM	0.61131	0.642671	0.609324	0.689007
Deberta	0.681324	0.735263	0.664356	0.733502
Bert	0.645969	0.694575	0.627922	0.709089
XLM roberta	0.659984	0.701356	0.646180	0.730585
Distillbert	0.623004	0.670318	0.610320	0.685859

The results in Table V show that sentiment analysis is a more challenging task than cyberbullying detection, as evidenced by generally lower F1-scores and accuracies across all models. The multi-class nature of sentiment analysis introduces more complexity, leading to a performance drop compared to the binary classification task of cyberbullying detection.

DeBERTa leads the individual models with an F1-score of 0.6813 and accuracy of 0.7353, followed closely by XLM-RoBERTa with an F1-score of 0.6599. Their superior performance highlights the importance of advanced deep learning models for handling nuanced text data.

BERT shows a decline in performance compared to its cyberbullying results, achieving an F1-score of 0.6460 and accuracy of 0.6946, indicating that it struggles more with multi-class sentiment classification.

Among the machine learning models, Logistic Regression again performs best, achieving an F1-score of 0.6171 and accuracy of 0.6555. However, it still lags behind the deep learning models, highlighting the limitations of traditional models for complex NLP tasks.

D. Discussion and Limitations

The findings highlight the ability of Transformer-based deep learning models to understand the complexities and subtleties of Malay-language social media content. Their strong performance in both cyberbullying detection and sentiment analysis is due to their capacity to model context, semantics, and linguistic variation more effectively than traditional machine learning models. Models like DeBERTa and XLM-RoBERTa benefit from deeper attention mechanisms and multilingual pretraining, which help them capture the nuances and informal language often found in Malay tweets.

Despite these advantages there are several limitations. Class imbalance particularly in sentiment categories affected overall precision and recall. Social media language also evolves rapidly, with new slang and cultural expressions emerging frequently. As a result, models trained on static datasets may struggle to maintain performance over time. Additionally, while Transformer models are accurate, they lack transparency, making them harder to interpret and trust in sensitive applications like automated content moderation.

Future work should focus on updating training data regularly to keep up with language trends. The integration of explainable AI (XAI) methods could also improve model transparency, making them more suitable for real-world deployment where interpretability is crucial.

V. CONCLUSION AND FUTURE WORK

It can be concluded that when it comes to detecting cyberbullying behavior online the transformers model demonstrated accuracy and overall performance when compared to machine learning models. This suggests that this approach is quite effective, for purposes. Transformer models like DeBERTa and XLM Roberta excelled in metrics highlighted their knack for capturing nuances in sentiment. Moreover transformer based models have shown better results than the machine learning models for both cyberbullying detection and sentiment analysis tasks. Moving forward there is a need to address the challenge of class imbalance which continues to pose a difficulty, in cyberbullying detection initiatives. For future work researchers could investigate methods such as data augmentation or creating synthetic data to even out the dataset better for minority categories such as positive emotions or cases not related to cyberbullying.

ACKNOWLEDGMENT

We would like to thank Universiti Teknologi Malaysia for their helpful feedback and support.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Zhu C, Huang S, Evans R, Zhang W. (2021). Cyberbullying among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures. *Front Public Health, 9*, 634909. Doi: 10.3389/fpubh.2021.634909.

[2] C. S. Lai, M. M. Mohamad, M. F. Lee, K. Mohd Salleh, N. L. Sulaiman and W. V. S. Chang. (2016). Cyberbullying: Understanding the Current Situations. *Prosiding Seminar Penyelidikan IYRES 2016: Menerajui Penyelidikan, Menempa Perubahan, Kuala Lumpur.*

[3] H. Ahmad Ghazali, A. Abu Samah, S. Z. Omar, H. Abdullah, A. Ahmad and H. A. Mohamed Shaffril. (2020). Predictors of Cyberbullying among Malaysian Youth. *Journal of Cognitive Sciences and Human Development*, *6*(1), 67–80.

[4] T. K. Chan, C. M. Cheung and Z. W. Lee. (2021). Cyberbullying on Social Networking Sites: A Literature Review and Future Research Directions. *Information & Management, 58*(2).

[5] Haidar, M. Chamoun and A. Serhrouchni. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. *Advances in*

Science, Technology and Engineering Systems Journal (ASTESJ), 2(6), 275−284.

[6] Emon, E.A., Rahman, S., Banarjee, J., Das, A.K., Mittra, T., (2019). A Deep Learning Approach to Detect Abusive Bengali Text. *2019 7th International Conference on Smart Computing & Communications (ICSCC),* IEEE. 1–5.

[7] Romim, N., Ahmed, M., Talukder, H., Islam, S., *et al.* (2021). Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. *Proceedings of International Joint Conference on Advances in Computational Intelligence,* 457–468.

[8] Islam, T., Ahmed, N., Latif, S. (2021). An Evolutionary Approach to Comparative Analysis of Detecting Bangla Abusive Text. *Bulletin of Electrical Engineering and Informatics*, *10*, 2163–2169.

[9] Aurpa, T. T., Sadik, R., Ahmed, M. S. (2022). Abusive Bangla Comments Detection on Facebook using Transformer-based Deep Learning Models. *Social Network Analysis and Mining, 12*, 1–14.

[10] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans and V. Hoste. (2018). Automatic Detection of Cyberbullying in Social Media Text. *PLoS ONE*, *13*(10).

[11] H. Margono. (2019). Analysis of the Indonesian Cyberbullying through Data Mining: The Effective Identification of Cyberbullying through Characteristics of Messages. Dissertation.

[12] Z. Zainol, S. Wani, P. N. Nohuddin, W. M. Noormanshah and S. Marzukhi. (2018). Association Analysis of Cyberbullying on Social Media using Apriori Algorithm. *International Journal of Engineering and Technology (IJET), 7*(4.29), 72−75,

[13] T. K. Hua, B. Abdul Hamid and S. M. Mat So'od, (2019). Linguistic Features of Cyberbullying. *International Conference on Advanced Research in Social Sciences*, London.

[14] R. Maskat, M. F. Zainal, N. Ismail, N. Ardi, A. Ahmad and N. Daud. (2020). Automatic Labelling of Malay Cyberbullying X Corpus using Combinations of Sentiment, Emotion and Toxicity Polarities. *ACAI 2020: 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya.*

[15] Nurina Farhanah Binti Johari and Juliana Jaafar. (2022). A Malay Language Cyberbullying Detection Model on X using Supervised Machine Learning, International Visualization, *Informatics and Technology Conference (IVIT).* Doi: 10.1109/IVIT55443.2022.10033395.

[16] N. F. B. Johari and J. Jaafar. (2022). A Malay Language Cyberbullying Detection Model on X using Supervised Machine Learning. *2022 International Visualization, Informatics and Technology Conference (IVIT)*, Kuala Lumpur, Malaysia. 325–332. Doi: 10.1109/IVIT55443.2022.10033395.

[17] 2020. Stopbullying.gov. what is cyberbullying.

[18] S. Hinduja and J. W. Patchin. (2018). Cyberbullying: Identification, Prevention,and Response. Cyberbullying.org.

[19] Cyberbullying Research Center. (2020). 2019 Cyberbullying Data.

[20] Cyberbullying Research Center. 2020. Summary of Our Cyberbullying Research (2007–2019).