

A DEEP LEARNING-BASED CYBER FRAMEWORK FOR BI-LINGUAL SENTIMENT ANALYSIS

Jiale Qi^{1*}, Siti Norul Huda Sheikh Abdullah², Fariza Fauzi³, Weixing Bai⁴

^{1,2,3} Center For Cyber Security, Faculty Of Information Science And Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

⁴School of Mechanical and Electrical Engineering, Zhengzhou University of Technology,
Yingcai Street, Huiji District, Zhengzhou City, Henan Province, P.R.China

Emails: p167251@siswa.ukm.edu.my^{1*}, snhsabdullah@ukm.edu.my^{2*},
fariza.fauzi@ukm.edu.my^{3*}, 20013007@zzut.edu.cn⁴

ABSTRACT

Automatically identifying abusive online content is more challenging when harmful expressions appear in different languages. This study examines the classification of cyber abuse in English and Chinese using two pre-trained Transformer models, Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), under a unified five-class labelling scheme covering discrimination, cyberbullying, hate speech, sexual harassment and normal content. The dataset was created by combining English and Chinese web-sourced data in three evaluation settings: English-only, Chinese-only and bilingual English-Chinese classification. The experimental findings reveal distinct language-specific behavior in the models. On the English dataset, RoBERTa performed better overall, achieving 91.16% accuracy and an F1-score of 91.11%, compared to BERT's 89.26% accuracy and an F1-score of 89.25%. In contrast, on the Chinese dataset, BERT performed better, achieving 87.28% accuracy and an F1-score of 84.01%, while RoBERTa achieved 83.55% accuracy and an F1-score of 82.40%. In the bilingual English-Chinese setting, both models produced similar results: BERT achieved 90.36% accuracy and 90.28% F1 score, while RoBERTa achieved 90.52% accuracy and 90.47% F1 score. These results suggest that the performance of transformer models is influenced by language-specific characteristics and data distribution. Therefore, the study provides empirical support for selecting cyber abuse detection models according to the target language environment, rather than relying on a single model across all multilingual settings.

Keywords: *Cyber Abuse Detection; English-Chinese Classification; Harmful Content Detection; BERT; RoBERTa; Multilingual Content Moderation*

1. INTRODUCTION

In the domain of natural language processing, researchers have demonstrated a notable interest in the manner in which individuals articulate emotional responses to disparate languages on social media platforms. With the increasing proliferation of social media platforms, review websites and global communication tools, user-generated text frequently incorporates multiple linguistic and cultural elements [1], [2], [3], [4]. This presents significant challenges to sentiment analysis, whilst concomitantly engendering new opportunities, such as the addressing of phenomena such as hate speech [5], [6], [7], cyberbullying [8] and sexual harassment [9] across age groups, national borders, and genders. In addition, conventional monolingual models are unable to process multilingual data in an effective manner [10]. The inability of models to accurately capture nuances in emotional expression across languages is a significant challenge. Consequently, there is a need to develop models capable of analyzing sentiment across multiple languages.

Despite the considerable success of sentiment analysis research in monolingual tasks, the majority of studies remain constrained to single-language scenarios [11]. Nonetheless, real-world applications (e.g. review evaluation) often have multilingual contexts (especially mixtures of English and Chinese), hence pose new challenges for using cross-lingual models and non-native representations [12], [13], [14]. Although pre-trained transformer models like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) are known to achieve high performance on English tasks, there are few systematic comparisons of their performance across other languages [15], [16]. As a result, we need to validate and compare them using both mixed Chinese-English texts and pure English texts. This methodology

is intended to identify the most appropriate models for cross-lingual lexical-semantic association, also evaluating their performance on their referent monolingual task.

In the 1990s and the early 21st century, traditional methods such as Hidden Markov Models (HMM) [17], Naive Bayes [18], and Support Vector Machines (SVM) [19] were extensively employed for a variety of tasks, including part-of-speech tagging, sentiment classification, and named entity recognition.[20]. Nevertheless, the aforementioned models were dependent on manually crafted features and expert dictionaries, which proved to be ill-suited for large-scale data. [21]]. Since 2020, the field of natural language processing has undergone a paradigm shift towards a model centered on pre-trained Transformers. The introduction of BERT and its variants, such as RoBERTa and DeBERTa, has led to significant advancements in text comprehension. In addition, sequence-to-sequence models, including T5 and BART, have shown a marked expansion in their generative capabilities. The GPT series models have exhibited remarkable few-shot learning performance, facilitated by parameter-efficient fine-tuning techniques, such as LoRA and adapters [22] A further reduction in the costs of customizing models has been achieved.

1. To construct a bilingual English-Chinese cyber abuse dataset under a unified five-class labelling scheme, covering discrimination, cyberbullying, hate speech, sexual harassment, and normal content.
2. To develop Transformer-based cyber abuse classification models using BERT and RoBERTa models for English-only, Chinese-only, and bilingual English-Chinese detection settings.
3. To evaluate and compare the performance of BERT and RoBERTa models across the three language settings using accuracy and F1-score, and to examine how language-specific characteristics influence model performance.

Organized of this paper is done based on the major stages of the study. The second section of the work based on multilingual cyber abuse detection and transformer-based text classification, investigate cross-lingual representations and language-dependent model behaviour. In Section 3, we explain the construction of the dataset, that is the collection, cleansing, alignment of labels and preparation of the English, Chinese and combined English-Chinese corpora. In Section 4, the experimental design is presented for the adaptation of BERT and RoBERTa for the five-class cyber abuse classification task. This section consists of tokenization, fine-tuning settings, classifier configuration and evaluation routines. The experimental findings regarding accuracy, precision, recall, and F1-score are reported in this section, and the differences in the two monolingual and a bilingual scenario are discussed. Section 5 concludes this paper. Furthermore, it summarizes the main findings. The section also outlines the implications of these findings for selecting language-sensitive models to moderate multilingual content.

2. LITERATURE REVIEW

With the increasing amount of multilingual and multimodal data on the internet, sentiment analysis has become an important aspect of natural language processing [23], [24]. Models like BERT and RoBERTa [16], [25] are good examples of models that can extract context-aware features. However, cross-lingual sentiment analysis faces three major challenges. Languages differ greatly in their vocabulary, syntax, expressions, etc. Thus, training on a model for a language cannot be transferred to a monolingual model. Second, a language might have limited resources and poorly annotated data. Another important consideration is that culture differences shape how people express emotion, suggesting that we need sturdier models. To address these challenges, the researchers have suggested several solutions including employing the pre-trained multilingual series, creating unified datasets with translation functionalities and enhancing cross-lingual consistency through contrastive learning.

Recent studies [26], [27], [28] have shown that it is particularly difficult to analyses the feelings behind different languages, especially when trying to apply this to languages that don't have much data. To overcome this challenge, an understanding of the emotional orientation of different linguistic contents is needed prior to pre-processing [29]. Multimodal sentiment analysis is the key to understanding sentiment more deeply. By fusing the auditory, visual and textual modalities, we can enhance the flexibility and generalization capability of multilingual sentiment analysis models, and keep the semantic integrity and increase the realistic application value.

2.1 BERT and RoBERTa models

BERT Model is a pretrained Transformer encoder that represents text by jointly modelling left and right contextual information. The pretraining process of the model combines masked language modelling with next-sentence prediction, thereby allowing the model to learn contextual representations that can be adapted to downstream

natural language processing tasks. In the present study, the BERT model is utilised as a foundational transformer model for the classification of cyber abuse into five distinct categories.

RoBERTa extends the BERT training framework by removing the objective of predicting the next sentence, applying dynamic masking, and increasing the scale of the data used for pretraining. It is evident that these changes enhance the system's capacity to identify contextual patterns in the context of text classification tasks. For this reason, RoBERTa is included as a comparative model to examine whether its optimized pretraining strategy provides advantages for English, Chinese, and bilingual English-Chinese cyber abuse detection [28], [30].

2.2 Differences Between BERT and RoBERTa Models

As illustrated in Table 1 below, BERT and RoBERTa exhibit significant discrepancies in their model architectures, which are primarily manifested in their training objectives, tokenization methods, masking mechanisms, sentence embedding approaches, and training data.

Table 1: The architecture design and capabilities of BERT and RoBERTa

Comparison aspect	BERT	RoBERTa	Relevance to this study
Model type	Pretrained Transformer encoder	Optimized pretrained Transformer encoder	Both models are suitable for text classification and are evaluated under the same cyber abuse detection task.
Pretraining objective	Masked language modelling with next-sentence prediction	Masked language modelling without next-sentence prediction	The difference in pretraining design may affect how each model captures contextual information.
Tokenization strategy	WordPiece tokenization	Byte-level Byte-Pair Encoding tokenization	Tokenization may influence how English, Chinese, and bilingual text are represented before classification.
Masking strategy	Static masking during pretraining	Dynamic masking during pretraining	Dynamic masking allows RoBERTa to learn from varied masked positions during training.
Sentence-pair modelling	Uses segment embeddings	Does not rely on segment embeddings in the same way as BERT	This distinction reflects different assumptions in input representation.
Pretraining scale	Trained on a smaller corpus than RoBERTa	Trained on a larger and more diverse corpus	Pretraining scale may contribute to RoBERTa's stronger performance on English and bilingual data.
Role in this experiment	Baseline transformer model	Comparative optimized transformer model	The comparison helps assess whether optimized pretraining improves English-Chinese cyber abuse classification.

As outlined in the preceding sections, the fundamental characteristics of the BERT and RoBERTa models have been described. The subsequent chapter will elaborate on the research methodology employed in this paper.

2.3 Experimental Evaluation and Validity in Cyber Abuse Detection

Detecting cyber abuse is key to experimental evaluation. Models should be tested under conditions that reflect the linguistic and categorical complexity of online harmful content. In multilingual contexts, the evaluation should not be based solely on the aggregate performance scores. The same model might behave differently in different languages, label categories and data distributions. A model that is good on English data is not necessarily trustworthy on Chinese data or a mixed bilingual corpus.

Detecting cyber abuse across languages is a challenging problem. The evaluation process is especially important because abusive expressions are affected by a number of factors, such as language structure, cultural usage and platform specific writing patterns. User-generated content is noisy, with spelling variations, informal wording and abbreviated expressions. These properties may affect tokenization, contextual representation, and final classification performance. The robustness of the models in separate monolingual and bilingual environments should be investigated and not only in a single dataset condition.

In this experiment we will look at validity from a number of different angles. First we will analyze internal validity which asks whether or not the observed differences between models are due to the model architecture and training configuration and not due to inconsistent pre-processing, data-splitting or parameter-setting. Next we will discuss external validity, i.e. how far we can generalize findings to other languages, datasets and types of online abuse. Third, we will look at the construct validity, which asks if the chosen measures of evaluation adequately reflect the goal of the classification. In the context of a five class cyber abuse task, accuracy is not enough as it could hide poor performance on some categories.

The study is more likely to be correct because the same evaluation process is used for BERT and RoBERTa when looking at English, Chinese and a mix of the two languages. The way the labels are set up, how the data is split, how the models are trained and how they are judged is the same for all languages. This makes it easier to compare the models. The study looks at which model performs better overall and how the model behaves in different language situations. The results show that models that are sensitive to language can be used to detect cyber abuse in many languages.

3. METHODOLOGY

This section delineates the methodological framework employed in this study. The following aspects are addressed: the data processing workflow, the data preparation steps, and the sampling methods based on three-fold cross-validation [31], as well as parameter selection, label merging, label mapping relationships, and evaluation metrics.

Table 2: Classification structure

Original label	Original count	Original count	Merged category count	Class index
age discrimination	771	discrimination	5,962	0
gender discrimination	3,191			
race discrimination	2,000			
cyber bully	5,539	cyberbullying	5,539	1
hate speech	4,312	hate speech	4,312	2
sexual harassment	7,978	sexual harassment	7,978	3
normal	10,147	normal	10,147	4

Table 2 shows how the attributes were processed for the English cyber sentiment dataset. Some of the original labels, like age, gender, and race discrimination, were put together into one category called 'discrimination'. This was done to make the labels more complete and to make the classification more consistent. The other categories were kept the same.

Table 3: The structure of label count

Original label	Original count	Standardized category	Merged category count	Class index
歧视 (discrimination)	6,004	discrimination	6,004	0
网络霸凌 (cyberbullying)	9,303	cyberbullying	9,303	1
仇恨语言 (hate speech)	5,728	hate speech	5,728	2
性骚扰 (sexual harassment)	7,293	sexual harassment	7,293	3
非性骚扰 (non-sexual harassment)	6,501	normal	6,301	4

This study created a set of data about bad behavior online that includes both English and Chinese languages. This data was made by combining sets of data that were already available to the public and that came from places where people share information about software. The English data was collected from sets of data about hate speech, harassment, and sexual harassment. Age discrimination, gender discrimination and racial discrimination were grouped under a broader category termed 'discrimination', whilst other types—such as cyberbullying, hate speech, sexual harassment and general content—were retained as separate categories. The Chinese data was subsequently adjusted to conform to this five-category classification structure, thereby enabling standardized classification processing.

During the data cleaning and pre-processing stage, we removed redundant columns, duplicate records, missing text entries and inconsistent category names. In addition, we deleted or modified labels that were unclear or used infrequently. We then merged the English and Chinese subsets to form a bilingual dataset comprising five standardized categories: discrimination, cyberbullying, hate speech, sexual harassment and normal content. The resulting dataset has 64,313 samples and provides a shared way of classifying English and Chinese cyber misbehavior.

The cleaned bilingual corpus was used to train BERT and RoBERTa classifiers. Each input text was first converted into a token sequence and then encoded into contextual representations by the transformer model. The classification layer was used to assign each sample to one of the five cyber misbehavior categories. This experimental setting allows the models to be evaluated on their ability to learn from both English and Chinese samples within the same detection task.

3.1 The Preparation of the English-Chinese Cyber Sentiment

The final set of data is made smaller and written as the English-Chinese network opinion dataset [32]. By using set theory concepts such as redundancy and similarity, researchers can organize this process and improve large-scale opinion analysis more efficiently. In other words, all of the data is processed, merged, and categorized through pre-processing, and the final amount of data is 64,313 samples.

Table 4: Tabulation for the English-Chinese cyber sentiment dataset

Class index	Standardized category	Number of samples
0	discrimination	11,968
1	cyberbullying	14,842
2	hate speech	10,040
3	sexual harassment	11,015
4	normal	16,448
Total		64,313

3.2 BERT and RoBERTa models in the Cross-Lingual English-Chinese Dataset

We constructed the architectures for the BERT and RoBERTa models, as shown in Fig 1, which were used for pipelined operations on different datasets [33]. BERT and RoBERTa divide the knowledge representation processing (i.e. keyword filtering, normalization and annotation) into two parts: BERT and RoBERTa are used to process this input separately, and the results are then subjected to unified keyword classification.

Let us assume that each token, E , is connected to an embedding vector, the token embedding (w_i) is the vocabulary, the positional embedding (p_i) excludes segment embedding, while the segment embedding (s_i) is typically inactive in the RoBERTa method.

$$E_i = w_i + p_i + s_i, \text{ where } X = [E_1, E_2, \dots, E_n] \in R^{n \times d} \quad (1)$$

The transformer then receives the token formulation as Equation (1) with dimensions, X representing the number of rows n , and the number of dimensions d . In contrast to BERT, RoBERTa applies embedding using multi-head self-attention (Equation 2) based on a SoftMax model that learns dependencies from WQ , WK , and WV learned matrices.

In self-attention, the input representation X is projected into three different representations through learnable weight matrices. The query vector $Q = XW^Q$ indicates what information the current token is seeking. The key vector $K = XW^K$ indicates what information each token can offer. The value vector $V = XW^V$ contains the actual information to be passed forward. The attention mechanism compares Q and K to calculate relevance scores and then uses these scores to weight the corresponding values.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Once data pre-processing was complete, this paper divided the dataset into a training set, a validation set and a test set in a ratio of 80:10:10. The training set is primarily used for model parameter learning, the validation set is used for hyperparameter optimisation and monitoring the training process, whilst the test set is used for an independent performance evaluation of the final model.

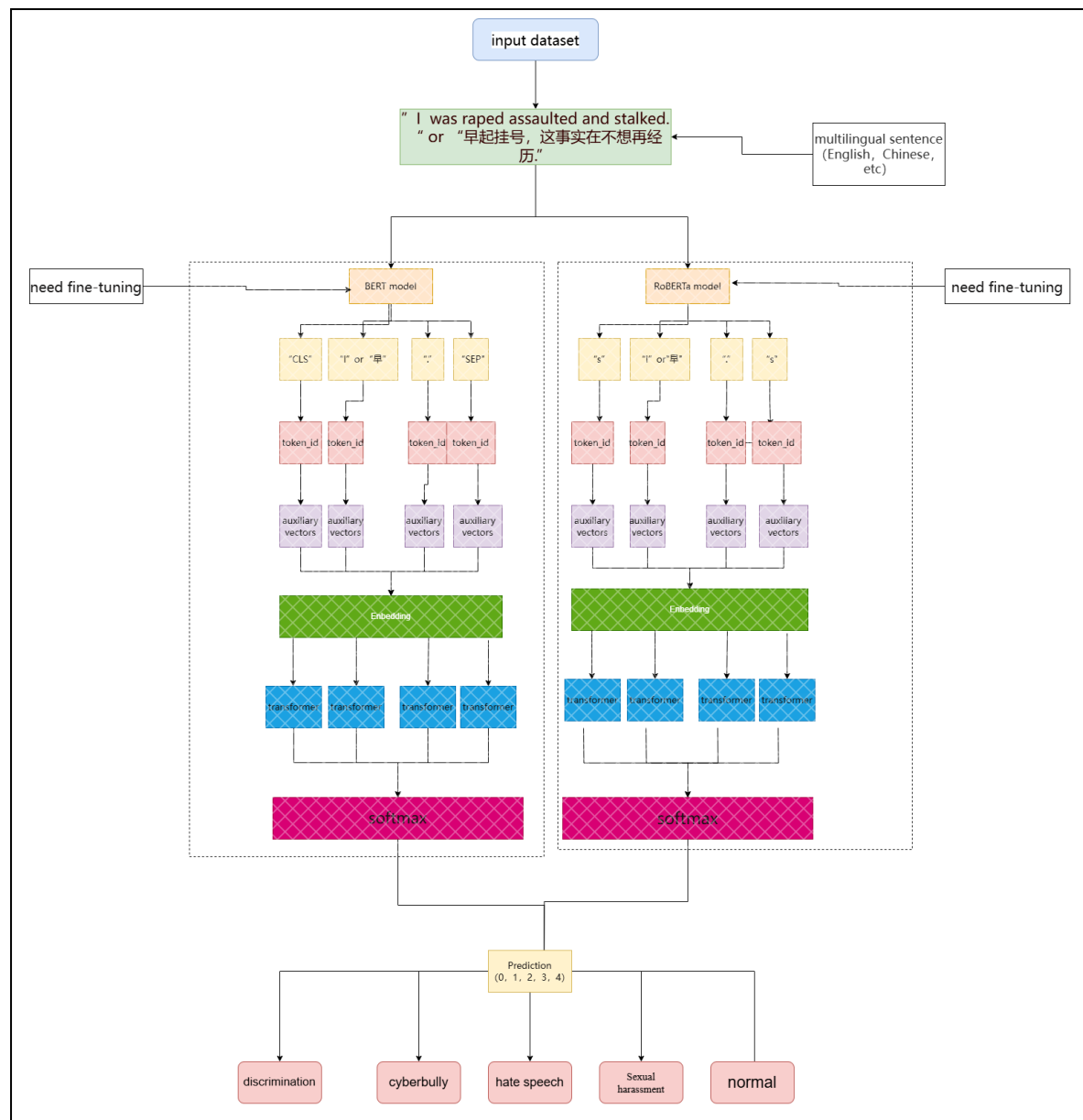


Fig.1: The flowchart of the BERT model and the Roberta model for English-Chinese Cyber Sentiment.

3.3 Evaluation Metrics

Both models mentioned in this article utilize the following evaluation metrics; I will also explain the formulas for calculating these metrics below [34]. These were used to evaluate model performance.

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

$$TPR = \frac{True\ Positive}{True\ Positives + False\ Negative} \quad (5)$$

$$FPR = \frac{False\ Positive}{False\ Positives + True\ Negative} \quad (6)$$

4. MODEL COMPARISON

The experiments showed that RoBERTa performed better than BERT across all the tasks tested, especially in the understanding of neutral sentiments. RoBERTa is trained on a larger dataset with better training, so it understands English and Chinese better in this case. Below is the histogram of the distribution of the length of the content text.

4.1 Parameter settings for data processing

During processing, it is essential to select the most appropriate parameters based on the distribution of text lengths within the dataset. Our experimental results corroborate previous findings: in multilingual sentiment analysis tasks, RoBERTa outperforms BERT in most cases, particularly in resource-constrained environments.

So, in this experiment, the model parameters were tuned a few times, including per-device training and evaluation batch sizes of 32, 64, and 128 [35], [36]. The same parameter settings were used for ten training runs to make a fair comparison. The two graphs below are a comparison of the training process of two models.

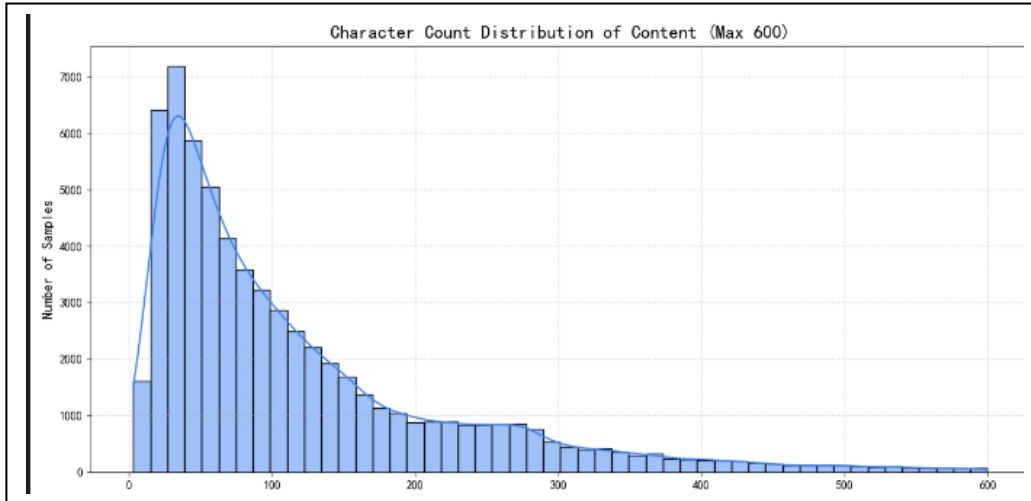


Fig. 2: The histogram illustrating content distribution.

Figure 2 shows how long the text is, with the longest part set at 600 characters. The overall distribution shows a clear pattern to the right, with most of the samples concentrated in the shorter text range, particularly between about 20 and 100 characters. The number of samples peaks at around 30 to 40 characters, then gradually decreases as the number of characters increases. The number of samples decreases a lot once the character count is over 150. After 300 characters, there are only a few long text samples left. This shows that short texts are more common in the dataset, while long texts are less common, but there is still a long-tail effect. So, when training models or

preparing text for processing, we need to think about how short texts have more information and how cutting out long parts of text can affect how well the model performs.

4.2 English dataset Results

Table 5: BERT and RoBERTa model 10-Epoch result in the English cyber sentiment dataset

Epoch	BERT Model				RoBERTa Model			
	Acc	Prec	Recall	F1 Score	Acc	Prec	Recall	F1 Score
1	85.4%	86.0%	85.4%	85.3%	86.1%	87.4%	86.1%	85.9%
2	88.0%	88.0%	88.0%	88.0%	89.3%	89.6%	89.3%	89.3%
3	89.2%	89.4%	89.2%	89.2%	90.9%	90.9%	90.9%	90.9%
4	89.9%	90.1%	89.9%	89.9%	91.6%	91.7%	91.6%	91.5%
5	89.7%	90.0%	89.7%	89.7%	91.9%	92.2%	91.9%	91.9%
6	90.1%	90.1%	90.1%	90.1%	92.4%	92.5%	92.4%	92.3%
7	90.4%	90.4%	90.4%	90.4%	92.3%	92.4%	92.3%	92.3%
8	90.1%	90.1%	90.1%	90.1%	92.2%	92.2%	92.2%	92.1%
9	89.9%	89.9%	89.9%	89.9%	92.5%	92.6%	92.5%	92.5%
10	89.9%	89.9%	89.9%	89.9%	92.4%	92.4%	92.4%	92.3%
Average	89.3%	89.4%	89.3%	89.3%	91.2%	91.4%	91.2%	91.1%

Note. Acc = Accuracy; Prec = Precision. All values are presented as percentages, and the average row reports the mean result across ten epochs.

Table 5 shows the same metrics for the two models on an English dataset. It is clear to see that, across all four metrics, the RoBERTa model outperforms the BERT model.

4.2 Chinese dataset Results

Table 6: BERT model 10-epoch result in the Chinese cyber sentiment dataset

Epoch	BERT Model				RoBERTa Model			
	Acc	Prec	Recall	F1 Score	Acc	Prec	Recall	F1 Score
1	88.3%	86.3%	88.3%	84.6%	78.4%	80.5%	78.4%	76.4%
2	88.5%	91.1%	88.5%	84.2%	79.9%	82.9%	79.9%	77.6%
3	88.1%	80.7%	88.1%	83.8%	81.1%	83.0%	81.1%	80.0%
4	88.2%	80.8%	88.2%	83.9%	85.2%	86.2%	85.2%	84.3%
5	88.4%	83.7%	88.4%	84.4%	84.3%	84.9%	84.3%	83.1%
6	88.2%	84.6%	88.2%	84.4%	84.6%	85.1%	84.6%	84.1%
7	87.5%	82.3%	87.5%	83.8%	85.6%	86.9%	85.6%	84.5%
8	85.7%	82.9%	85.7%	83.9%	85.3%	86.4%	85.3%	84.3%
9	84.6%	82.9%	84.6%	83.6%	85.7%	86.4%	85.7%	85.0%
10	85.4%	82.4%	85.4%	83.5%	85.5%	86.4%	85.5%	84.6%
Average	87.3%	83.8%	87.3%	84.0%	83.6%	84.9%	83.6%	82.4%

Note. Acc = Accuracy; Prec = Precision. All values are presented as percentages, and the average row reports the mean result across ten epochs.

However, in the Chinese dataset shown in Table 6, the BERT model outperforms the RoBERTa model on all four metrics, which differs from the results observed in the English dataset.

4.3 The results of the bilingual cyber sentiment dataset

The results indicate that different models perform differently according to the language used. The results support the hypothesis that RoBERTa is better at modeling contextual patterns in English and bilingual data, which can be attributed to its optimized pretraining strategy and dynamic masking mechanism. In contrast, BERT performs better on the Chinese dataset, especially in terms of recall and F1-score. This result shows in Table VII remarks that BERT has improved the stability of Chinese cyber abuse classification in this experiment, and this may be because the process of BERT's tokenization and representation is more consistent with the characteristics of Chinese textual data [37]. Finally, the research results highlight the significance of selecting a model for

multilingual cyber abuse detection. To summarize, the research conclusions highlight the importance of selecting a model for multilingual cyber abuse detection that takes the target language distribution into account. This contrasts with relying on a single model across all settings.

Table 7: BERT and RoBERTa model results in the English-Chinese cyber sentiment dataset

Epoch	BERT Model				RoBERTa Model			
	Acc	Prec	Recall	F1 Score	Acc	Prec	Recall	F1 Score
1	90.1%	91.3%	90.1%	90.0%	90.0%	91.1%	90.0%	90.0%
2	90.6%	91.6%	90.6%	90.5%	90.8%	91.9%	90.8%	90.7%
3	91.0%	92.0%	91.0%	90.9%	91.1%	92.2%	91.1%	91.0%
4	90.8%	92.0%	90.8%	90.7%	90.4%	91.5%	90.4%	90.4%
5	90.6%	91.6%	90.6%	90.6%	91.0%	92.2%	91.0%	90.9%
6	90.4%	91.4%	90.4%	90.4%	90.8%	91.7%	90.8%	90.8%
7	90.3%	91.2%	90.3%	90.3%	90.4%	90.9%	90.4%	90.4%
8	90.6%	91.5%	90.6%	90.5%	90.4%	91.0%	90.4%	90.3%
9	89.4%	89.7%	89.4%	89.4%	90.2%	90.6%	90.2%	90.1%
10	89.7%	90.0%	89.7%	89.6%	90.0%	90.3%	90.0%	90.0%
Average	90.4%	91.2%	90.4%	90.3%	90.5%	91.3%	90.5%	90.5%

Note. Acc = Accuracy; Prec = Precision. All values are presented as percentages, and the average row reports the mean result across ten epochs.

Table 8: Parameters Comparison

Dataset	Model	Acc	Prec	Recall	F1 Score
English dataset	BERT Model	89.3%	89.4%	89.3%	89.3%
	RoBERTa Model	91.2%	91.4%	91.2%	91.1%
Chinese dataset	BERT Model	87.3%	83.8%	87.3%	84.0%
	RoBERTa Model	83.6%	84.9%	83.6%	82.4%
English-Chinese dataset	BERT Model	90.4%	91.2%	90.4%	90.3%
	RoBERTa Model	90.5%	91.3%	90.5%	90.5%

Note: This table deals with average parameters, Acc = Accuracy; Prec = Precision. All values are presented as percentages, and the average row reports the mean result across ten epochs.

Table 8 shows a comparison of BERT and RoBERTa in terms of accuracy, precision, recall and F1-score on the English, Chinese and combined English-Chinese datasets. The results reveal definite patterns in language-dependent performance. RoBERTa gets better performance in English dataset and has a small advantage in bilingual setting, while BERT has a better performance in Chinese dataset. This indicates that the language distribution of the data influences the performance of pretrained transformer models.

4.4 Confusion Matrix and ROC Curve

Performance varies across different languages. In terms of overall performance in English sentiment classification, RoBERTa has been shown to outperform BERT. For example, RoBERTa achieved 5,698 correct classifications, whilst BERT achieved 5,447; in hate speech classification, BERT's 3,832 correct classifications were surpassed by RoBERTa's 3,729. However, RoBERTa reduced the number of misclassifications for both the 'discriminatory' and 'normal' sentiment categories. Its overall performance is more balanced, with RoBERTa producing fewer misclassifications than BERT for both 'discriminatory' and 'normal' sentiments.

On Chinese datasets, however, the BERT model achieves higher accuracy in predicting cyberbullying, discrimination, hate speech and normal sentiments. RoBERTa outperforms BERT in bilingual datasets, achieving higher correct classifications in key categories such as normal (626) and discrimination (589) in Chinese, and hate speech (899) and cyberbullying (1386) in bilingual data. Both models are good at English sentiment classification, with AUC values of 0.99–1.00 across categories. The more balanced and stable AUC performance of RoBERTa, notably in the case of cyberbullying (AUC 0.98), indicates a better cross-lingual generalization.

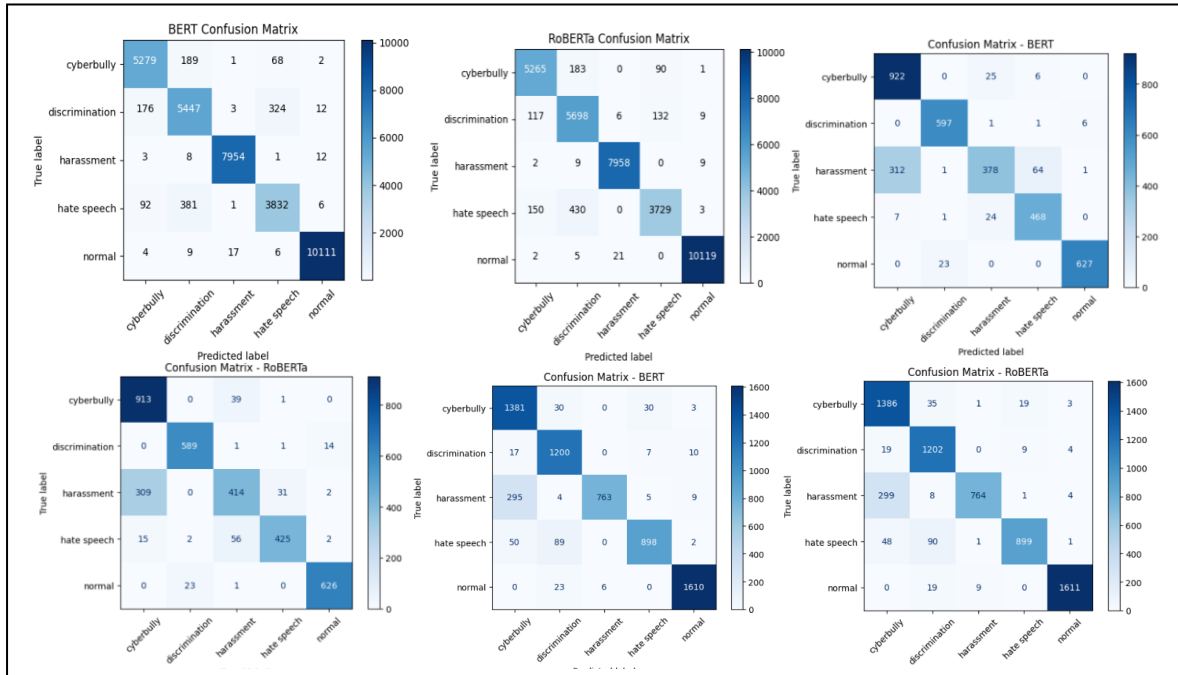


Fig. 3: Confusion matrix on BERT model and the Roberta model tested on bilingual dataset.

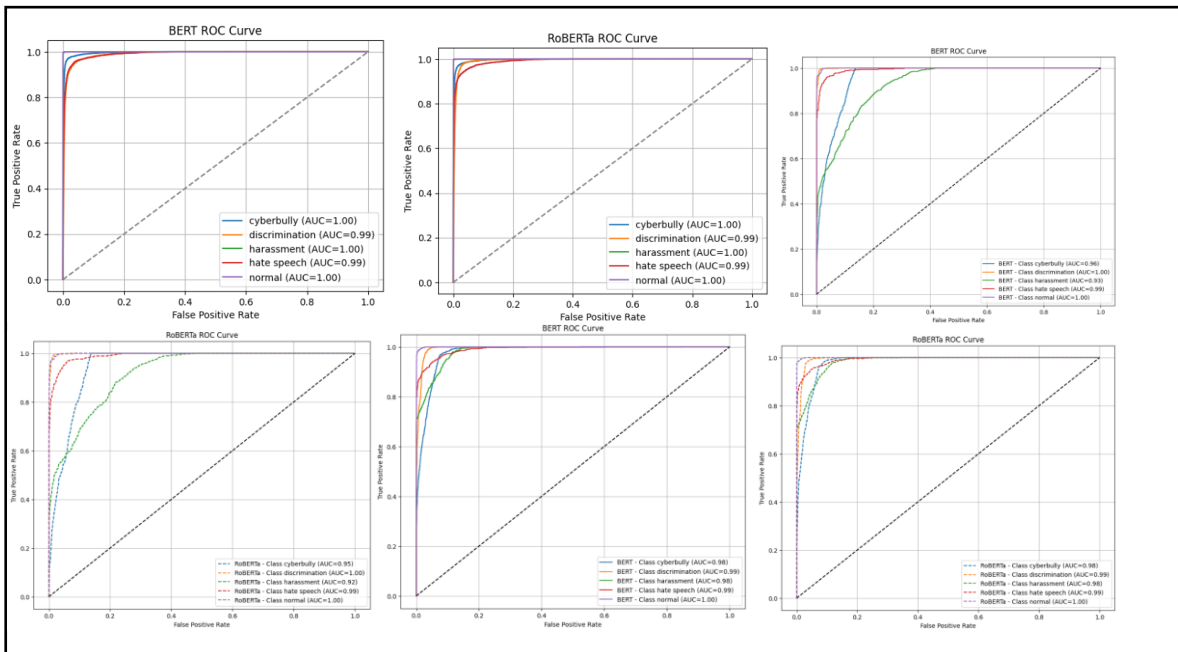


Fig 4: ROC curve on the BERT model and the RoBERTa model tested on the bilingual dataset.

5. CONCLUSION

The results suggest that the composition of the dataset by language affects the performance of the transformer-based cyber abuse detection. RoBERTa achieves better performance on the English dataset and a slight edge on the combined English-Chinese setting, while BERT performs better on the Chinese dataset. This pattern shows that the performance of the model is not consistent across languages, and the performance should be interpreted in the context of the specific linguistic context of the task. In the bilingual setting, the difference between BERT and RoBERTa is small, which indicates that both models can learn useful representations for the classification of English-Chinese cyber abuse. But their relative strengths seem to emerge in different data settings. RoBERTa has better performance when the data set has more English or mixed languages. BERT has a more stable performance on Chinese-only classification in this experiment. The experimental results show that the pretrained transformer models are effective for five-class cyber abuse detection on both English and Chinese data. Our findings suggest that models should be selected based on factors such as language distribution, label structure and empirical performance, rather than simply opting for the conventional Transformer model. We recommend that future research should encompass ensemble strategies for Chinese and bilingual user-generated text, as well as language-specific fine-tuning studies. These approaches are expected to enhance the reliability of multilingual content moderation systems.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Higher Education Malaysia under grant numbers FRGS/1/2021/ICT02/UKM/02/1 and IMAP/3/2024/ICT07/UKM/1.

REFERENCES

- [1] V. Srivastava and M. Singh, "HinGE: A Dataset for Generation and Evaluation of Code-Mixed Hinglish Text," in *Eval4NLP 2021 - Evaluation and Comparison of NLP Systems, Proceedings of the 2nd Workshop*, Association for Computational Linguistics (ACL), 2021, pp. 200–208. doi: 10.26615/978-954-452-056-4_020.
- [2] I. Sterner, "Multilingual Identification of English Code-Switching," 2024. [Online]. Available: <https://github.com/igorsterner/AnE>
- [3] M. D. Bui, K. von der Wense, and A. Lauscher, "Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision-Language Models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers*, 2025.
- [4] F. Alshargi, A. Abulohoom, S. Yagi, F. Jabr, L. Lulu, and A. Elnagar, "Aratox: a multi-dialect, multi-label arabic dataset and model benchmark for toxicity detection," *Language Resources and Evaluation*, 60:39, 2026, doi: 10.1007/s10579-026-09917-9.
- [5] "Multilingual-hate-speech-detection".
- [6] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," 2021. [Online]. Available: <https://github.com/punyajoy/HateXplain>
- [7] J. Schäfer and E. Kistner, "HS-EMO: Analyzing Emotions in Hate Speech." [Online]. Available: <https://github.com/Johannes-Schaefer/>
- [8] M. Dadvar, F. De, J. Roeland, and O. Dolf Trieschnigg, "Improved Cyberbullying Detection Using Gender Information," 2012. [Online]. Available: <http://www.noswearing.com/dictionary>
- [9] K. Maity, S. Bhattacharya, S. Saha, and M. Seera, "A Deep Learning Framework for the Detection of Malay Hate Speech," *IEEE Access*, vol. 11, pp. 79542–79552, 2023, doi: 10.1109/ACCESS.2023.3298808.
- [10] U. Ihsan, H. Ashraf, and N. Jhanjhi, "Multilingual Sentiment Analysis Using Deep Learning: Survey," Dec. 26, 2023. doi: 10.20944/preprints202312.1990.v1.
- [11] Ł. Augustyniak, M. Morzy, and T. Kajdanowicz, "Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark." [Online]. Available: <https://huggingface.co/datasets/Brand24/mms>
- [12] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale." [Online]. Available: <https://github.com/facebookresearch/cc>
- [13] Y. Xu, H. Cao, W. Du, and W. Wang, "A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations," Sep. 01, 2022, *Springer*. doi: 10.1007/s41019-022-00187-3.
- [14] L. Wu *et al.*, "Enhancing LLM Language Adaption through Cross-lingual In-Context Pre-training," pp. 27152–27166, Nov. 2025, doi: 10.18653/V1/2025.EMNLP-MAIN.1380.

- [15] A. Chakraborty, S. Joardar, D. K. Prasad, and A. A. Sekh, "BGPCN: A BERT and GPT-2-based Relational Graph Convolutional Network for hostile Hindi information detection," *Array*, Volume 28, 2025, Article 100601, 2025, doi: 10.1016/j.array.2025.100601.
- [16] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv, 2019. doi: 10.48550/arXiv.1907.11692.
- [17] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989, doi: 10.1109/5.18626.
- [18] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997, doi: 10.1023/A:1007413511361.
- [19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [20] A. Yadav and D. Kumar Vishwakarma, "A Deep Multi-Level Attentive network for Multimodal Sentiment Analysis."
- [21] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.
- [22] A. Campagner, F. Milella, D. Ciucci, and F. Cabitza, "Three-way decision in machine learning tasks: a systematic review," *Artif. Intell. Rev.*, vol. 57, no. 9, Sep. 2024, doi: 10.1007/s10462-024-10845-9.
- [23] N. Mohamed Ali, M. M. A. El Hamid, and A. Youssif, "SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS," *International Journal of Data Mining & Knowledge Management Process*, vol. 09, no. 03, pp. 19–27, May 2019, doi: 10.5121/ijdkp.2019.9302.
- [24] A. Rajesh and T. Hiwarkar, "Sentiment analysis from textual data using multiple channels deep learning models," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Nov. 2023, doi: 10.1186/s43067-023-00125-x.
- [25] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?" [Online]. Available: <https://github.com/google-research/bert>
- [26] A. A. Alsuwaylimi and Z. S. Alenezi, "Leveraging Transformers for Detection of Arabic Cyberbullying on Social Media: Hybrid Arabic Transformers," *CMC-Computers Materials & Continua*, 2025, doi: 10.32604/cmc.2025.061674.
- [27] S. Rajak and U. Baruah, "Hybrid transformer-convolution neural network architecture for emoji-aware hostile post classification with custom attention mechanisms," *Eng. Appl. Artif. Intell.*, 2025, doi: 10.1016/j.engappai.2025.111037.
- [28] K. Ghosh and A. Senapati, "Hate speech detection in low-resourced Indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments," *Natural Language Processing*, 2025, doi: 10.1017/nlp.2024.28.
- [29] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: <https://github.com/tensorflow/tensor2tensor>
- [30] C. C. F. Chu, C. C. H. Tong, C. H. Chiu, and S. C. Lam, "Early discovery of cyberbullying incidents in multiparty Chinese-English code-mixed colloquial dialogue: A generative AI approach," *Intelligent Systems with Applications*, 2026, doi: 10.1016/j.iswa.2026.200652.
- [31] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: what does it estimate and how well does it do it?," Apr. 2021, doi: 10.1080/01621459.2023.2197686.
- [32] Qi Jiale, Siti Norul Huda Sheikh Abdullah, Masnizah Mohd, Azianura Hani Shaari, and Ruan Haoda, *Cyber Cross-Lingual Sentiment Analysis Model Based On Deep Learning Methods*. Dubai, United Arab Emirates: IEEE, 2025. Accessed: Mar. 07, 2026. [Online]. Available: 10.1109/ICCR67387.2025.11292379
- [33] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [34] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice", doi: 10.1162/tacl.
- [35] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting Pre-trained Models for Chinese Natural Language Processing." [Online]. Available: <https://github.com/ymcui/MacBERT>
- [36] X. Chen, D. Wen, and D. Zuo, "Towards Detecting Chinese Harmful Memes with Fine-Grained Explanatory Augmentation," *Electronics (Basel)*, 2025, doi: 10.3390/electronics14173504.
- [37] H. Peng, E. Cambria, and A. Hussain, "A Review of Sentiment Analysis Research in Chinese Language," Aug. 01, 2017, *Springer New York LLC*. doi: 10.1007/s12559-017-9470-8.