Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Master in Data Science

Master Thesis

Text Classification of Cyberbullying Comments: A Study on the Applicability of Various BERT Models

Author: <<SEONG-MIN GONG>>

Madrid. <<June, 2023>>

Master Thesis
Máser en ***Ciencia de Datos***

Title: **Text Classification of Cyberbullying Comments : A Study on the Applicability of Various BERT Model**
June, 2023

Author: <<**Seong Min Gong**>>

Supervisor:
Emilio Serrano Fernandez

ETSI Informáticos
<<Departmento de Inteligencia Artificial>>
UPM

Co-supervisor:
Zanardini Damiano

ETSI Informáticos
<< Departmento de Inteligencia Artificial >>
UPM

Abstract

Despite the significant advantages of the contemporary digital landscape, it also provokes considerable societal dilemmas, one of which is cyberbullying. Characterized by attacks or derogatory comments made under the cloak of online anonymity, this detrimental behavior induces psychological distress in victims and negatively impacts the overall wellness of digital communication environments. In light of this, there is a growing demand for solutions employing Natural Language Processing (NLP) technologies.

In response, this study explores the application of NLP technologies, sophisticated tools that empower computers to comprehend and process human language, in the context of cyberbullying. Three pre-trained models, BERT, RoBERTa and DeBERTa which are proficient in deciphering context and inferring connotations from textual data, were utilized in this investigation.

A dataset comprising cyberbullying comments extracted from digital platforms such as Twitter and Kaggle was compiled. An evaluation of the models was conducted to determine their efficacy in detecting malicious comments, thereby revealing areas for improvement during the fine-tuning process. The results showed that all three models achieved an accuracy of over 93%. Particularly, the BERT Multilingual model demonstrated an accuracy of 94.1%. However, during the transfer learning process, the RoBERTa and DeBERTa models, excluding the BERT models (Base: 84.9%, Large: 86.6%, Multilingual: 85.9%), exhibited lower accuracy in the range of the 63%.

Based on the findings, a need for additional research into the optimization of fine-tuning strategies and model development reflecting linguistic diversity was identified. By pursuing these research directions, the performance of text classification models intended for cyberbullying comment prevention can be greatly enhanced. This approach serves as a crucial step towards managing cyberbullying issues more effectively and minimizing the harm they cause, thereby fostering healthier communication within online communities.

Keywords: NLP, BERT, RoBERTa, DeBERTa, Cyberbullying Detection, Text Classification.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The subject of this paper pertains to text classification for cyberbullying detection. As digital communication technologies have advanced, and further catalyzed by the effects of the COVID-19 pandemic, online activities have significantly increased. In particular, platforms like Twitter, TikTok, and YouTube provide a level of anonymity that encourages more free expression of opinions. Unfortunately, this also leads to an exponential increase in cyberbullying, where the freedom of expression is abused to foster hate speech and biased remarks.

Cyberbullying has emerged as a global social problem, with effective countermeasures still an ongoing challenge. This problem is especially devastating for teenagers who extensively use social network services. According to a 2022 study, [1] 46% of American teenagers aged between 13 and 17 have experienced cyberbullying, leading to psychological stress and depression. Furthermore, research shows that children and teenagers under 25, as both victims and perpetrators, are 2.1 and 1.23 times more likely, respectively, to demonstrate suicidal behavior[2].

This issue isn't confined to teenagers alone. In South Korea, instances of celebrities ending their lives due to online harassment have become worryingly commonplace[3]. These severe consequences underscore the urgent need for society to actively respond to cyberbullying.

However, detecting and responding to cyberbullying poses significant challenges. Specifically, social media platforms must process massive amounts of data, rendering manual review of all posts virtually impossible. As a result, many cases of cyberbullying remain undetected [4]. South Korea attempted to address this issue by discontinuing comment services, but this came at the expense of the usefulness of internet comments, proving it is not a complete solution [5]. Furthermore, the use of text classification models for cyberbullying detection can lead to ethical issues related to privacy protection. In the freedom-assured online space, people may not want their posts to be monitored and judged by machines, raising concerns about cybersecurity and privacy infringement.

Moreover, the complexity and nuances of language make it difficult for text classification to perfectly interpret messages. Due to the subtlety of contextual information and the ambiguity of language, there is a high potential for misclassification, which can lead to false cyberbullying alerts and miss real incidents of bullying.

Nevertheless, the research and attempts to apply such technology using text classification in the actual online environment are crucial. Hence, this study aims to demonstrate the efficiency and practicality of using cutting-edge text classification models, such as BERT, RoBERTa, and DeBERTa, which have shown top-tier performance. Through this research, it is hoped that a significant contribution will be made to the advancement of cyberbullying detection technology. Ultimately, the goal is to provide an essential tool for reducing cyberbullying and protecting victims.

1.2 Objectives

The main objectives of this research are as follows :

- Comparison of the Performance of Various Text Classification Models:

This study, It is aimed to verify the effectiveness and efficiency of cyberbullying detection using high-performance text classification models such as BERT, RoBERTa, and DeBERTa. Specifically, conducting a performance comparison analysis focusing on the BERT Base, Large, and Multilingual models, the RoBERTa Base, Large, and XLM models, and the DeBERTa Base model.

- Optimization of Cyberbullying Detection Using Text Classification Models:

Based on an understanding of how each model detects cyberbullying, It is aimed to propose solutions for optimizing these models.

Through this study, this research seeks to identify the most effective text classification model for cyberbullying detection, with the ultimate goal of making a significant practical contribution to the response to cyberbullying.

1.3 Thesis Organization

This thesis is organized into six main sections.

In Chapter 1, 'Introduction', the background and motivation for this research are presented, along with the objectives of the study.

Chapter 2, 'Related Work and State of Arts', discusses the state-of-the-art language models such as BERT, RoBERTa, and DeBERTa, and reviews previous studies related to these models.

In Chapter 3, 'Methodology', the research methodology is detailed. This includes the implementation environment, dataset, as well as application of the BERT, RoBERTa, and DeBERTa models.

Chapter 4, 'Results', introduces the design of experiments and evaluation metrics. Then, it compares and analyzes the performance results of each model.

In Chapter 5, 'Conclusion and Future Work', the key conclusions of this research are derived and suggestions for future research directions are proposed.

Lastly, Chapter 6, 'References', lists all references consulted during the writing of this thesis."

2 Related work and State of Arts

2.1 Related work

With the advent of the big data era, the importance of analysis techniques to process big data has been elevated. Among these techniques, Natural Language Processing (NLP) is being utilized in various fields such as language translation, chatbots, and speech-to-text conversion. Notably, with the proliferation of social networks on the Internet, the detection of cyberbullying has emerged as a major research subject. Both past and present studies have employed machine learning and deep learning models for NLP to investigate cyberbullying.

For instance, Xiang Zhang et al.[6] observed that the pronunciation of spelling errors in informal conversations on social media remains constant, and used the pronunciation-based convolutional neural network (PCNN) to study the performance of cyberbullying datasets online, achieving a performance of 98% . Sweta and Amit[7] conducted research on the cyberbullying text classification problem by comparing the performance of the DNN-based models LSTM, CNN, and BLSTM with machine learning-based models like regression, SVM, random forest, and naïve Bayes using datasets extracted from various social media platforms like Wikipedia, Formspring, and Twitter. Cynthia Van Hee et al.[8] divided the subjects into bullies, victims, and bystanders using SVM and further divided the damage categories into threats, insults, and sexual conversations in their cyberbullying-related study. In South Korea, Naver[9], a major portal site, launched a chatbot with applied CNN+BiLSTM+LSTM model technology to detect malicious comments on the portal site.

Following these studies, the development of the BERT model, specialized for natural language processing by Google AI Language Team in 2018, brought a new phase to the research on detecting cyberbullying. This resulted in an innovative shift in the NLP field, spurring active research across multiple languages and tasks. There are diverse studies comparing the traditional deep learning, machine learning models with BERT.

Fatma Elsafloury et al.[10] conducted a study on detecting malicious comments related to cyberbullying using LSTM and BERT, but concluded that BERT did not assign importance scores to linguistic features related to cyberbullying. Bob Sanders[11] compared the performance using SVM, logistic regression, and PreTrainedCyberbullying BERT models, demonstrating BERT's superior performance.

Mohammed AL-Hashedi et al.[12] constructed an emotion detection model based on the premise that cyberbullying could provoke negative emotions, combined it with BERT, and researched the efficiency of cyberbullying comment detection. Chen and Rodey[13] conducted research comparing performances using RoBERTa and DeBERTa, BERT's derivative models, for sequential labeling of Dialogue Act, Sentiment/Emotion. Bayode and Babitha [14] used a total of eight algorithms including BERT, CLNet, RoBERTa, XLM-RoBERTa, and a combination of SBERT+SVM, a traditional machine learning model, to research effective algorithms for cyberbullying

However, these studies typically compared the performance of cyberbullying detection using either a single BERT model or a few derivative models. In contrast, the study aims to comprehensively compare various BERT derivative models, including BERT Base, Large, Multilingual, RoBERTa Base, Large, XLM, and DeBERTa. The goal is to understand the unique characteristics of each model and identify the most effective model for cyberbullying detection. This study aims to provide new criteria for the model selection in the text classification of cyberbullying comments and provide substantial help in addressing the cyberbullying issue. This distinctive approach and objective set our research apart from previous studies.

2.2 State of Arts

The Transformer is a natural language processing model presented by Google, building upon the encoder-decoder structure of the Seq2Seq model. [15] The use of self-attention techniques (bi-directional contextualized representation) allows the model to learn information regardless of the length of input data. During the training phase, the output layer following the decoder model can be modified to fit the desired task.

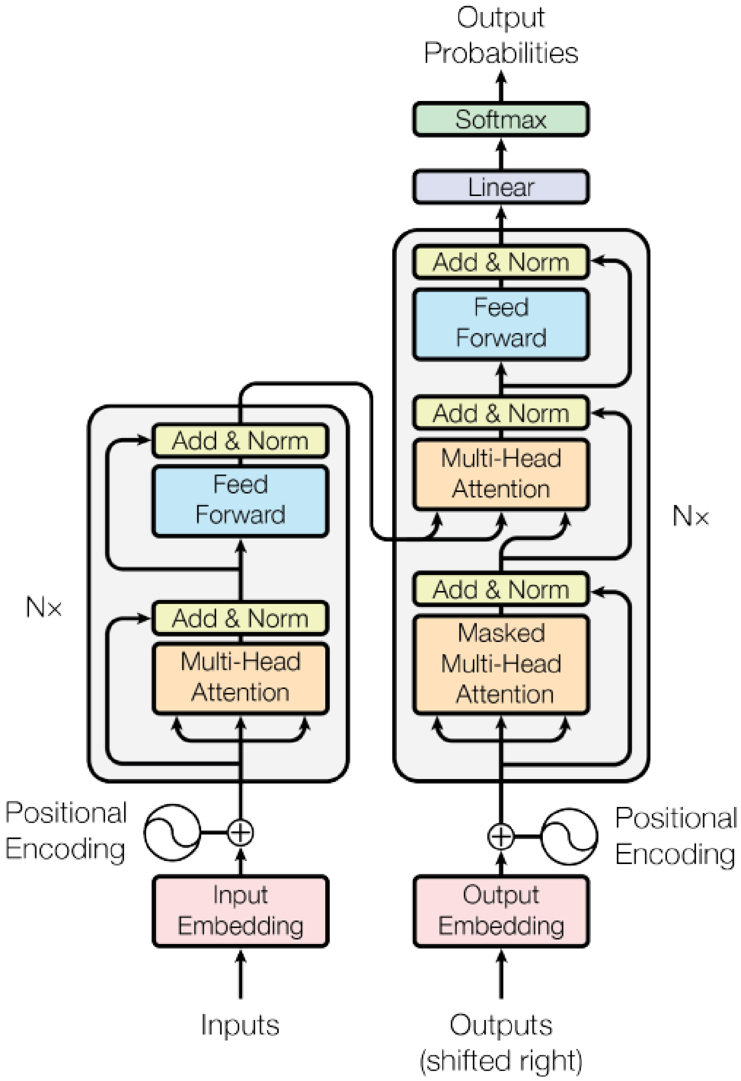


Figure 1. Transformer model architecture (Vaswani et al., 2017)

Subsequently, numerous pre-trained natural language processing models, such as BERT, RoBERTa, XLNet, ELECTRA, and DeBERTa, were developed based on the Transformer. In this study, the pre-trained natural language processing models BERT, RoBERTa, and DeBERTa, all based on the Transformer, were utilized to detect cyberbullying.

XLNet was excluded due to its permutation-based learning method, which, despite its ability to capture bidirectional context and its independence from data corruption, is computationally inefficient and complex [16].

ELECTRA, which employs a 'Replaced Token Detection' method where a portion of the words is replaced with different words and the model is trained to predict which words were replaced, was also not utilized [17]. While this approach could work effectively in diverse contexts, performance can be restricted when training data is limited.

Therefore, due to their computational efficiency and effectiveness in dealing with limited data, BERT, RoBERTa, and DeBERTa were selected for this research. This choice aims to ensure an optimal approach for cyberbullying detection.

2.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[18] is a revolutionary approach in pre-training language models, overcoming the shortcomings of previous methods. Unlike traditional left-to-right models, which often fall short in tasks that require an understanding of the bidirectional context like Question Answering (QA), BERT is designed to account for the context from both the left and the right in all layers. This capability enhances the fine-tuning-based pre-training approach originally introduced by OpenAI's GPT.

The BERT model undergoes pre-training using large English corpus datasets, including BookCorpus and English Wikipedia. A unique approach in BERT's pre-training is the application of the masked language model (MLM) method. This involves randomly masking a portion of the input tokens and predicting the original word based only on the context of the masked word. Moreover, it utilizes a next sentence prediction strategy that allows for joint pre-training on a pair of sentences.

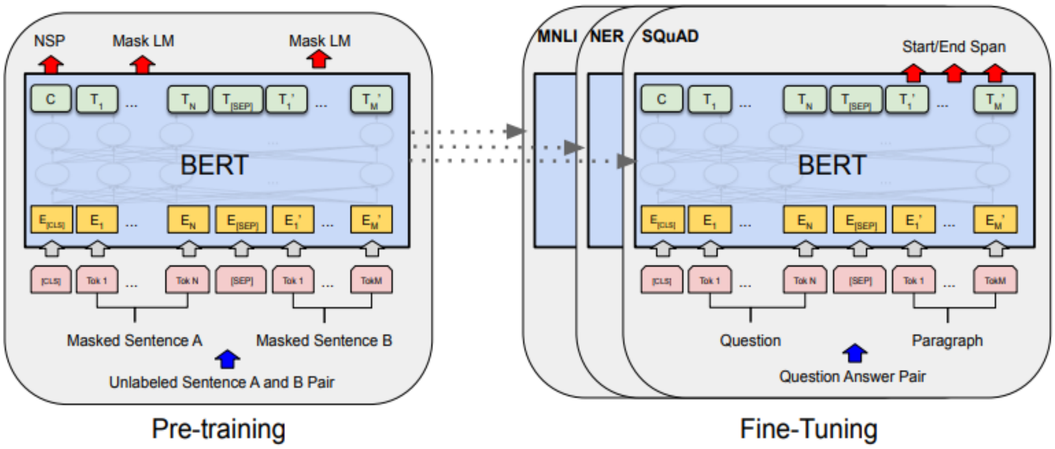


Figure 2. Pre-training and fine-tuning procedures for BERT (Devlin et al., 2018)

As seen in the figure, except for the output layer, the architecture of BERT remains consistent during both pre-training and fine-tuning. The same pre-training model parameters are initialized for various subsequent tasks, and all parameters are fine-tuned during the fine-tuning process.

BERT is available in two versions, distinguished by their model size:

- BERT BASE: Consisting of L=12, H=768, A=12, and Total Parameters=110M. It is designed with the same hyperparameters as OpenAI GPT for performance comparison.
- BERT LARGE: Consisting of L=24, H=1024, A=16, and Total Parameters=340M.

*(** L refers to the number of layers, H refers to the hidden size, and A refers to the number of self-attention heads)*

The input embedding of BERT is a summation of token embeddings, segment embeddings, and position embeddings.

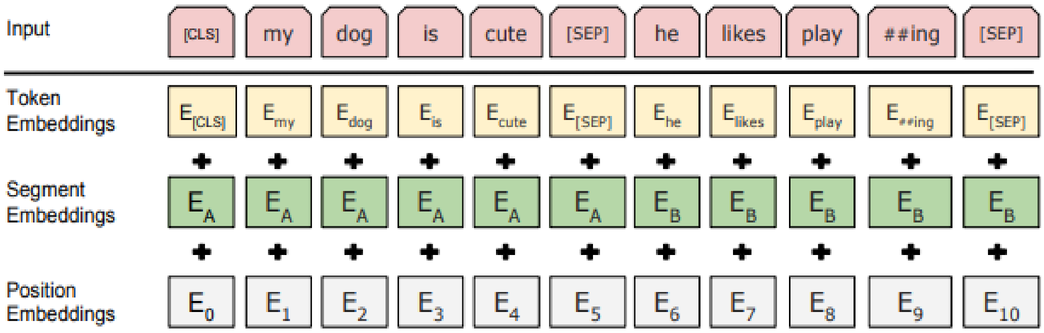


Figure 3. BERT input presentation (Devlin et al. 2018)

2.2.2 RoBERTa

2.2.2.1 RoBERTa

RoBERTa emerged from iterative research on BERT's pre-training, which revealed that BERT was significantly undertrained. This led to the proposal of better methods for training the BERT model [19].

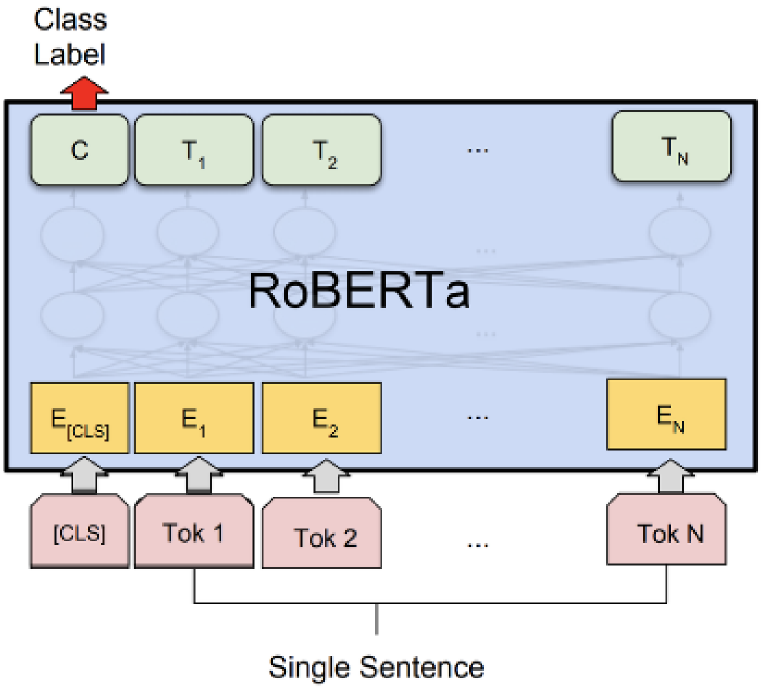


Figure 4. ***Pre-training and fine-tuning procedures for RoBERTa (Liu et al., 2019)***

This model was trained with 160GB of uncompressed text, which includes CC News, OpenWebText, and Stories, along with the data originally used for BERT. Unlike BERT, which only performs masking once during the pre-processing phase, RoBERTa employs dynamic masking, which generates a different masking pattern for each sequence inserted during the learning process. When compared with the original BERT model and static masking, dynamic masking displayed similar or slightly improved performance, leading to its use in the RoBERTa model.

Moreover, RoBERTa discarded the Next Sentence Prediction (NSP) task that was used in BERT. This decision was based on research findings indicating that NSP could potentially have a negative impact on model performance.

RoBERTa models can be classified into two types according to their size:

- RoBERTa BASE: L=12, H=768, A=12, Total Parameters=110M
- RoBERTa LARGE: L=24, H=1024, A=16, Total Parameters=340M

**L signifies the number of layers, H stands for hidden size, and A represents the number of self-attention heads.*

2.2.2.2 XLM-RoBERTa

XLM-RoBERTa is a specialized version of the RoBERTa model developed by Facebook AI. It's designed to handle multiple languages effectively, making it a versatile language processing tool [20]. The model is trained using a massive and well-balanced dataset called "CommonCrawl," which includes text from over 100 languages.

In evaluations comparing XLM-RoBERTa with other algorithms like BERT Large, mBERT, XLM-15, and XLM-Rbase, across seven language datasets, including English and Spanish, XLM-RoBERTa demonstrated outstanding performance. It outperformed the others, particularly showing a slight improvement in English compared to BERT Large, which is impressive.

With its multilingual capabilities and proficiency in understanding natural language, XLM-RoBERTa proves to be highly valuable for various language processing tasks. It excels in environments where multiple languages are involved and where accurate comprehension and question answering are essential. As a result, XLM-RoBERTa can be effectively applied in applications requiring multilingual natural language understanding and question answering.

2.2.3 DeBERTa

DeBERTa is a transformer-based neural language model that advances previous state-of-the-art pre-trained language models (SOTA PLMs) by leveraging two techniques: the disentangled attention mechanism and an enhanced mask decoder. [21]

In disentangled attention, while each word in BERT's input layer is represented by a single vector - the sum of word and position embeddings, in DeBERTa, each word is represented using two vectors that encode content and position separately. The attention weights between words are computed using disentangled matrices based on content and relative position. This approach is founded on the view that a word's attention weight depends not only on content but also on its relative position.

Similar to BERT, DeBERTa uses masked language modeling (MLM) for pre-training. MLM is a fill-in-the-blank task where the model is trained to use surrounding words to predict what the masked word is. DeBERTa utilizes the content and position information of the context word for MLM. The disentangled attention mechanism initially determines content and relative position for the context word, but it doesn't initially decide on the word's absolute position, which is a crucial factor in most predictions. The implication is that sentences with similar local contexts but different syntactic roles heavily rely on absolute position, underscoring the importance of considering absolute position in the model. DeBERTa incorporates absolute word position embedding just before the softmax layer, which decodes masked words based on aggregated contextual embeddings of word contents and positions.

DeBERTa's performance is compared with transformer-based PLMs of similar structure (composed of 24 layers with a hidden size of 1024), including BERT, RoBERTa, XLNet, and ELECTRA, across eight natural language understanding (NLU) tasks from GLUE. Despite being trained on 78GB of data compared to the 160GB used for pre-training RoBERTa, XLNet, and ELECTRA, DeBERTa consistently outperforms BERT and RoBERTa across all tasks. Furthermore, DeBERTa surpasses XLNet's performance in six out of eight tasks. Notably, significant differences are observed in MRPC (outperforming XLNet by 1.1%, RoBERTa by 1.0%), RTE (outperforming XLNet by 2.4%, RoBERTa by 1.7%), and CoLA (outperforming XLNet by 1.5%, RoBERTa by 2.5%). DeBERTa also outperforms other SOTA PLMs in terms of average GLUE score, including the ELECTRA large and XLNet large models.

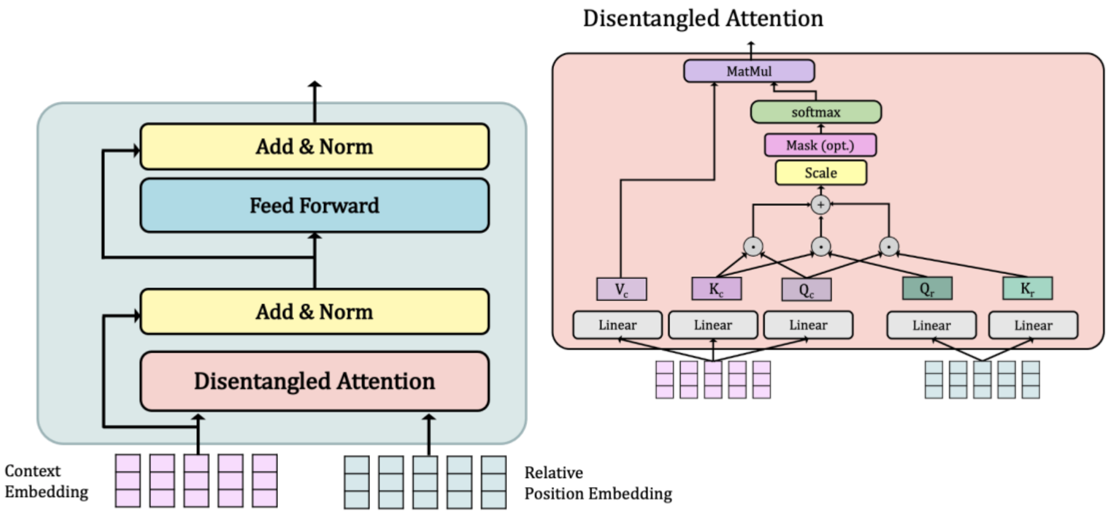


Figure 5. DeBERTa Encoder using Disentangled Attention(He et al., 2020)

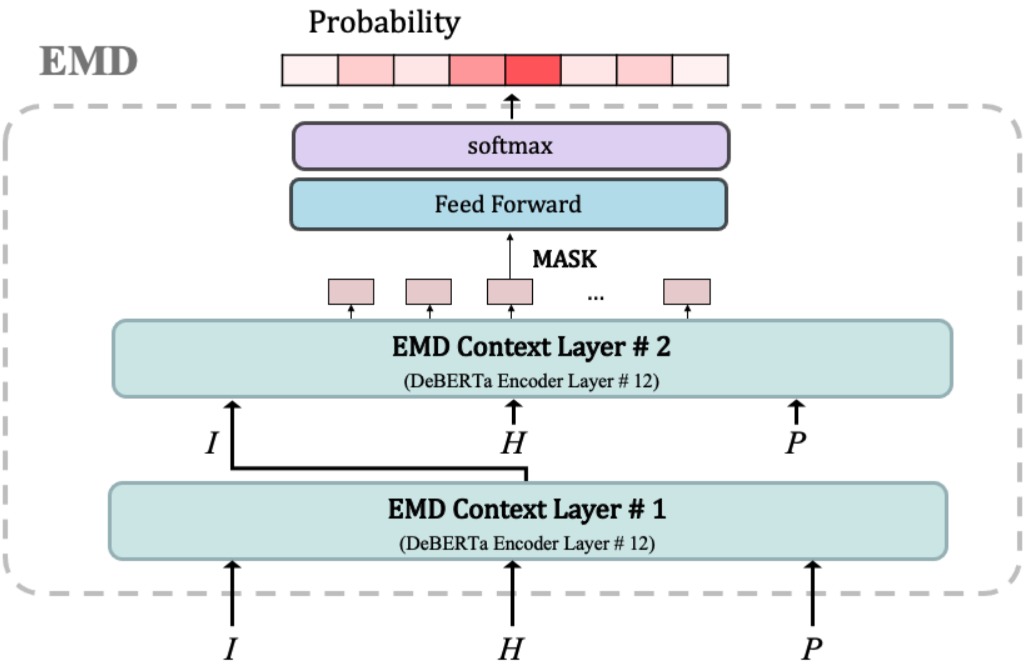


Figure 6. DeBERTa EMD (He et al., 2020)

Like BERT and RoBERTa, the DeBERTa model is available in Base and Large configurations, with model settings mirroring those of its counterparts.

3 Methodology

3.1 Implementation Environment

All experiments were conducted in the Google Colab Pro Plus environment. Google Colab is a cloud-based development environment where all necessary resources are pre-prepared, making installation and environment configuration easy. In addition, it provides GPUs and TPUs, allowing for the rapid processing of complex and computationally intensive tasks.

Natural Language Processing (NLP) requires abundant computational power for the analysis and processing of large volumes of data. For such complex and computation-intensive tasks, the use of high-performance GPUs is essential. Google Colab Pro offers various GPU models for this purpose.

GPU Model	Architecture	CUDA Core Count	Tensor Core Count	Memory	NVLink Support	PCIe Connection
A100	Ampere	6,912	826	40GB, 80GB, 160GB	Supported	Supported
V100	Volta	5,120	640	16GB, 32GB	Supported	Supported
T4	Turing	3,072	320	16GB	Not Supported	Supported

Table 1. Comparison of specifications by GPU models

Nowadays, Google Colab provides 40GB for the A100, and 16GB each for the V100 and T4.

During the process, some base models could be implemented sufficiently with the provided TPUs. However, as the size of the dataset increased and Large models were used, a GPU memory problema had occurred. This issue was particularly severe when conducting experiments using the BERT and RoBERTa Large models. Therefore, in these cases, it was necessary to utilize the A100 GPU. This enabled us to prevent situations where training had to be terminated due to GPU memory overload.

The flexibility, performance, and a variety of GPU options of Google Colab Pro greatly contributed to achieving the high-performance computation processing needed in this study.

3.2 Dataset

3.2.1 Dataset collection

One of the main objectives of this study is the identification and classification of various forms and patterns of cyberbullying. To significantly enhance the diversity of data, data was collected from multiple social media platforms, thereby not relying on a single online platform, via Kaggle[22],[23], which offers data aggregated from various sources.

Datasets classified with a range of labels such as hate speech, aggression, insult, gender discrimination, and racial discrimination were integrated. This approach was necessitated due to the multi-faceted nature of cyberbullying. A primary advantage of integrating diverse datasets is the enhancement of sample diversity. This was intended to facilitate more accurate recognition of cyberbullying by the model, which can occur in a wide array of situations and contexts.

In addition, the integration of data from various sources enables a balance to be achieved in the model, preventing it from excessively favoring a single label. It was predicted that this could play a crucial role in enhancing the model's generalization performance. Through such an approach, it is aimed to provide a deeper understanding of the myriad aspects of cyberbullying, thereby contributing to the development of a more reliable cyberbullying detection model.

As noted above, the aim of this study is to detect cyberbullying, not confined to specific topics or forms, hence, all labels were consolidated into a binary classification of 0 and 1.

Given that the datasets each had different formats and labeling methods, uniformity was ensured by restructuring all columns into [Text, oh_label].

To ensure that the training dataset is not biased towards a single label, we randomly sampled an equivalent number of instances from general comments and a cyberbullying dataset for the task. This approach aimed to maintain a balanced representation of both classes during the training process.

The integrated dataset was organized as shown in the following [*Table 2*].

Dataset	**Non-CyberBullying**	**CyberBullying**	**Total**
Kaggle	5,993	2,806	8,799
Twitter	19,446	43,124	62,570
Aggression_parsed	101,082	39,747	115,846
Toxicity_parsed	144,324	14,782	156,094
Intergrated_data	270,845	72,482	343,327

Table 2. Dataset information

3.2.2 Data Preprocessing

In this study, the *html* package of Python was utilized in the process of text data preprocessing. Initially, a transformation of HTML entities into regular text was carried out. During this process, any patterns within the text matching ***#[xX]?\w+;*** were located and an ***'&'*** was appended to the beginning, thereby transforming text that could be identified as HTML tags into regular text. Subsequently, the unescape function from the html package was employed to convert HTML characters into Unicode characters.

Next, URLs and email addresses were removed, a step conducted to refine the data by discarding unnecessary information such as URLs and email addresses.

Subsequent operations involved the removal of or conversion into regular text of HTML tags and HTML entities. Throughout this step, HTML entities such as ***&lt;/b&gt;, &quot;, quot;, &amp;, amp;, &lt;, lt;, &gt;, gt;*** were transformed into regular text.

Following this, brackets and text enclosed within brackets in the text were discarded. This was a measure taken to remove any superfluous information embedded within the text.

Lastly, hashtags, user handles, and multiple spaces were eliminated or converted into single spaces. This step was designed to exclude non-essential information and prevent errors that could arise from the number of spaces during the text analysis process.

Once the text data had been preprocessed in this manner, stopword removal was performed on each row. Stopwords are frequently used in sentences but bear no particular meaning, such as articles, 'be' verbs, and prepositions that form sentence structures. It was deemed beneficial to remove these words as they do not significantly contribute to discerning the topic of the context, thus allowing the dimensionality of the data to be reduced and the burden on the algorithm to be lessened. The function nfx.remove_stopwords from the nfx package was used for stopword removal.

3.3 BERT

In this study, three variants of BERT, namely Base, Large, and Multilingual, were employed for the task of cyberbullying text classification. Although the collected datasets were in English, the Multilingual model was included to assess its generalization ability across multiple languages.

The BERT model was developed based on TensorFlow and numerous related resources have been implemented using TensorFlow. Therefore, in this study, the three different models of the BERT were trained using TensorFlow. This is for facilitating the utilization of resources such as pre-trained weights, model architectures, and training data associated with BERT models.

The datasets were partitioned into training and testing sets in a 70:30 ratio for all three models. Subsequently, 50% of the test datasets were utilized as validation sets

3.3.1 BERT-base

In the case of the BERT Base model, an experiment was conducted using a randomly extracted subset of 46,000 out of the 140,000 collected datasets, in order to observe performance differences based on the size of the dataset.

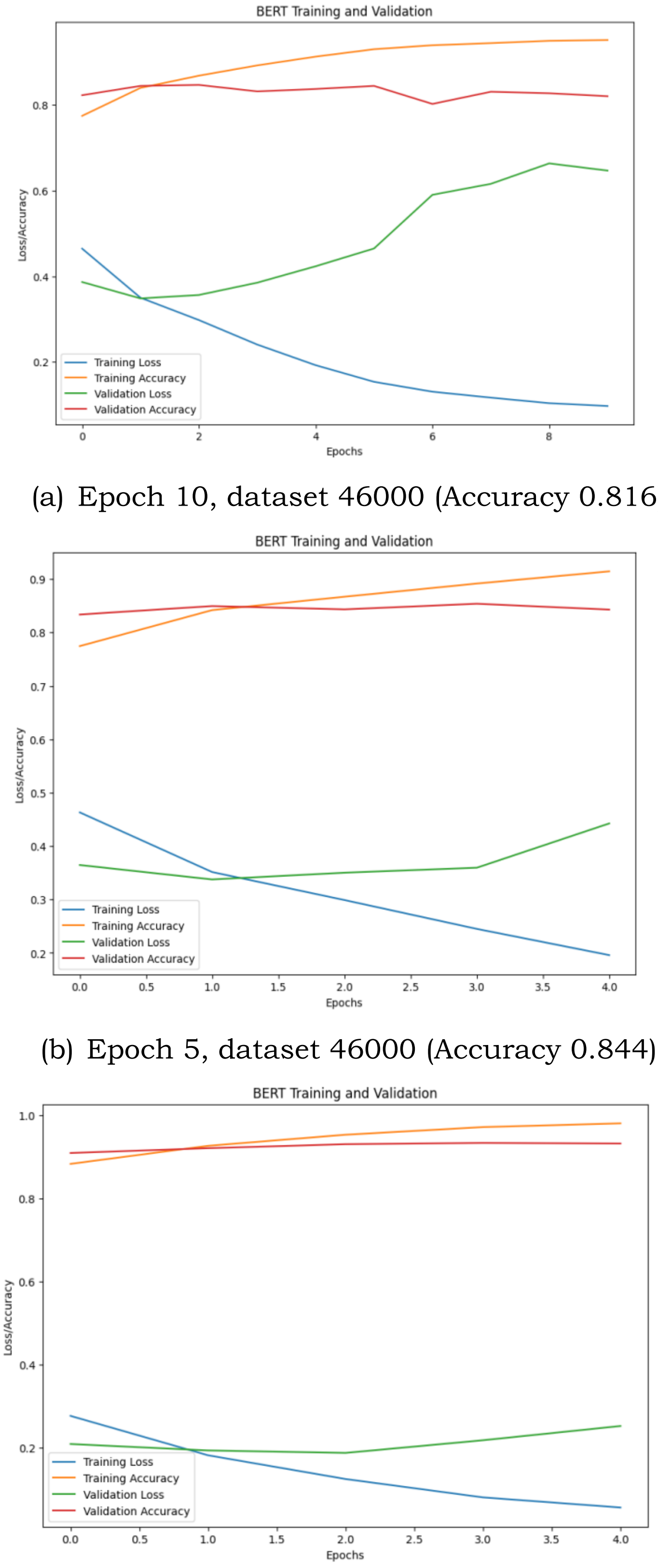
The *'bert-base-cased'* tokenizer, provided by Hugging Face's Transformers library, was utilized to transform the texts in the datasets into tokens during the tokenization phase. Following this, mask and segment information was generated for each token, which then served as input to the model. All text data was padded to maintain a length of 128.

For the binary classification of comment categorization, a *sigmoid* activation function was employed in the Dense layer, and the Rectified *Adam* optimizer was utilized for model optimization. To prevent overfitting, the *droprate* was set at 0.1.

The BERT Base model was implemented to operate effectively in a TPU environment. The *f.distribute.cluster_resolver.TPUClusterResolver* provided by TensorFlow was used to connect to a TPU cluster, and the model was executed within that cluster.

The batch size used for training was set at 32, and the initial epoch was set at 10. However, due to signs of overfitting, the number of epochs was reduced to 5. This adjustment resulted in improved performance in the learning rate.

Also, Significant differences were observed in overfitting and accuracy depending on the size of the dataset, prompting the use of larger datasets in subsequent model experiments. [*Figure 7*]



(a) Epoch 10, dataset 46000 (Accuracy 0.816)

(b) Epoch 5, dataset 46000 (Accuracy 0.844)

(c) Epoch 5, dataset 140000 (Accuracy 0.934)

Figure 7. Comparison graph of accuracy by Epoch and dataset (a),(b),(c)

Finally, the trained model was used to predict cyberbullying comments in the test dataset, and the model performance was evaluated by comparing these predictions with actual labels. The results were evaluated using F1 score and confusion matrix.

3.3.2 BERT-large and BERT-multilingual

BERT Large and Multilingual models were implemented using Hugging Face's Transformers library. Text data was tokenized using 'bert-large-cased' and 'bert-multilingual-cased' tokenizers respectively. In the process of conversion, mask and segment information for each token were generated and used as inputs to the models. The preprocessing of the input data was conducted in the same way as the base model.

However, due to the structural complexity and the capability of handling larger datasets of the BERT Large and Multilingual models, learning in a TPU environment was not feasible. Consequently, training was performed using an A100 GPU. The output of the models was extracted from the last hidden layer of the model and fed into a Dense layer.

These models were adapted AdamW as the optimization algorithm. This decision was made based on the judgment that AdamW, which includes weight decay in Adam for controlling model complexity and effectively preventing overfitting, is more suitable for the large models that deal with a more complicated structure and larger datasets. The learning rate was set to 1.0e-5 and weight decay, a type of L2 regularization, was set to 0.0025 to prevent overfitting. The activation function of the output layer was set to sigmoid, which is suitable for binary classification tasks.

During the model training process, the batch size was set to 32 and the number of epochs was set to 5. This was due to comparing the results of 10 epochs, higher accuracy and F1 scores were observed at 5 epochs. A separate validation dataset, apart from the training dataset, was utilized to objectively evaluate the performance of the models. This was used to measure the generalization ability of the models.[*Table 3*]

Models		Epoch	Accuracy
BERT	Large	5	93.1
		10	85
	Multilingual	5	94.1
		10	87.3

Table 3. Accuracy differences based on the number of epochs

3.4 RoBERTa

The data was randomly shuffled and subsequently divided into training, validation, and testing datasets according to a specific ratio. 70% of the data was assigned to training, with the remaining 15% each allocated to validation and testing.

The experimental environment was set up with the device configured as a GPU. The learning rate was set at 1e-5, the maximum input sequence length (max_len) was 256, and the batch size was set to 32. In the preprocessing phase, the dataset's elements were tokenized, padded to match the set maximum length, and the labels were then converted into tensors. Following this, the model was loaded onto the preconfigured device and training was initiated.

To avoid overfitting, the dropout rate was set at 0.1. The output of the model was processed by RobertaForSequenceClassification, which accepts text sequences as input and outputs logits values for each class. Since the default setting for the output layer was a softmax activation function, no additional activation function was set.

The training utilized the AdamW optimizer, with the number of epochs set to 5. Upon completion of training, performance was quantified using the f1-score.

3.4.1 RoBERTa-base

The base version of RoBERTa utilizes a Byte Level BPE tokenizer, which tokenizes original data at the byte level and applies the BPE algorithm. For training this model, the *RobertaTokenizer* and *RobertaForSequenceClassification* from the Hugging Face's *transformers* library were imported and '*roberta-base*' was called for both the tokenizer and model.

3.4.2 RoBERTa-large

The expanded version of the base, known as RoBERTa-large, also uses the BPE tokenizer. The difference is that the large model possesses more parameters, thereby significantly increasing the computational process and resource usage compared to the base version. For training this model, the *RobertaTokenize*r and *RobertaForSequenceClassification* from the Hugging Face's *transformers* library were imported, and '*roberta-large*' was called for both the tokenizer and model.

3.4.3 XLM-RoBERTa

XLM-RoBERTa is a model developed for multilingual classification based on RoBERTa. The fundamental algorithm remains the same as the original RoBERTa. For training this model, the *XLMRobertaTokenizer* and *XLMRobertaForSequenceClassification* from the Hugging Face's transformers library were imported, and *'xlm-roberta-base'* was called for both the tokenizer and model.

3.5 DeBERTa

The 'base' model of DeBERTa was employed for the project due to the limitations of the computational resources preventing the utilization of the larger version. The training process was conducted under similar conditions to those implemented for previous RoBERTa models. The dataset was divided into training, validation, and testing subsets at ratios of 70%, 15%, and 15%, respectively.

GPU was selected as the training device, with specific parameters set to a learning rate of 1e-5, a maximum length of 256, and a batch size of 32. In the preprocessing stage, elements in the dataset were tokenized, padded to the pre-established maximum length, and the labels were transformed into tensors. The model was then loaded onto the designated device.

Training was carried out using the *DebertaTokenizer* and DebertaForSequenceClassification from the transformers library. Both the tokenizer and the model called upon '*microsoft/deberta-base'*. When initializing the model, a dropout rate of 0.1 was set to prevent overfitting.

AdamW optimizer was used for the model's training with an epoch setting of 5. Upon completion of the training, the model's performance was evaluated based on its F1 score.

3.6 Transfer Learning

3.6.1 Validation Dataset

To prevent duplication with the training dataset, a separate dataset obtained from Hugging Face[24] was used for validation. This dataset, named 'hate_toxic_speech', consists of binary labels indicating whether the text is classified as cyberbullying or not, with values 0 and 1.

- 0 = General comment
- 1 = Cyberbullying comment

The dataset comprises of 42,045 general comments and 42,045 cyberbullying comments. Text preprocessing was performed following the same methodology as the training dataset, ensuring consistency in the input representation for the model.

3.6.2 Model Saving Methods

Following the completion of each model's training, the trained models were saved to evaluate their performance using a new dataset for transfer learning. Two approaches to saving a model are to either save only the weights of the trained model or to save the entire model. Saving only the weights can reduce file size and save/load time, increasing efficiency and usefulness for weight application in a new environment. However, it necessitates the separate definition of the same model structure.

On the other hand, saving the entire model can include the model structure, weights, and set optimizer, making it advantageous when training needs to be resumed or when sharing or reproducing a particular model state. Yet, it requires larger storage space. In this study, the weights were saved for the BERT model, while the entire model was saved for the RoBERTa and DeBERTa models due to time and storage constraints. The weights of BERT's Base, Large, and Multilingual models were saved using the *.save_weights( )* function. RoBERTa's Base, Large, and XLM models and DeBERTa's Base model were saved in their entirety using *model.save_pretrained()* and *tokenizer.save_pretrained()* methods.

- BERT (Base, Large, Multilingual) – save.weights()
- RoBERTa (Base, Large, XLM-RoBERTa) - *model.save_pretrained(), tokenizer.save_pretrained()*
- DeBERTa (Base) - *model.save_pretrained(), tokenizer.save_pretrained()*

3.6.3 BERT

The BERT models, for which only the weights were saved, had their model structure defined following the original training method. The saved weights were applied using the *.load_weight()* method when creating the models. Attempts were made to improve the model's performance by fine-tuning learning rate and epoch numbers. However, the highest accuracy was observed when the original training method was reproduced.

- Learning Rate: 1e-5
- Epochs: 5
- Batch Size: 32

3.6.4 RoBERTa and DeBERTa

The RoBERTa and DeBERTa's Base models were trained using the same dataset used for BERT's transfer learning. Initially, SequenceClassification and Tokenizer, suitable for each model, were invoked from transformers. The saved models were then applied.

- Model Invoke : ForSequenceClassification.from_pretrained(MODEL_PATH)
- Tokenizer Invoke : tokenizer = Tokenizer.from_pretrained(MODEL_TYPE)

Afterward, the learning rate, Max_Len, and batch size were set to 1e-5, 256, and 32, respectively, for transfer learning. The input data for training was preprocessed following the same method as in the training phase, and the saved models were then invoked for training. After the training was completed, the trained data was initialized, and the DataLoader was reset. The model was then switched to the evaluation mode using *model.eval().* The performance of the model was evaluated using the classification_report from sklearn.metrics.

- Learning Rate: 1e-5
- Epochs: 5
- Batch Size: 32

4 Result

In a results, for Precision, Recall, and F-1 Score of the three BERT models, it is observed that the BERT Base model presents a precision of 95.7% and a recall of 90.9% for label 0, resulting in an F1 score of 93.3%. For label 1, a precision of 91.2% and a recall of 95.9% is shown, also leading to an F1 score of 93.5%. These results indicate that the BERT Base model demonstrates superior performance for both labels.

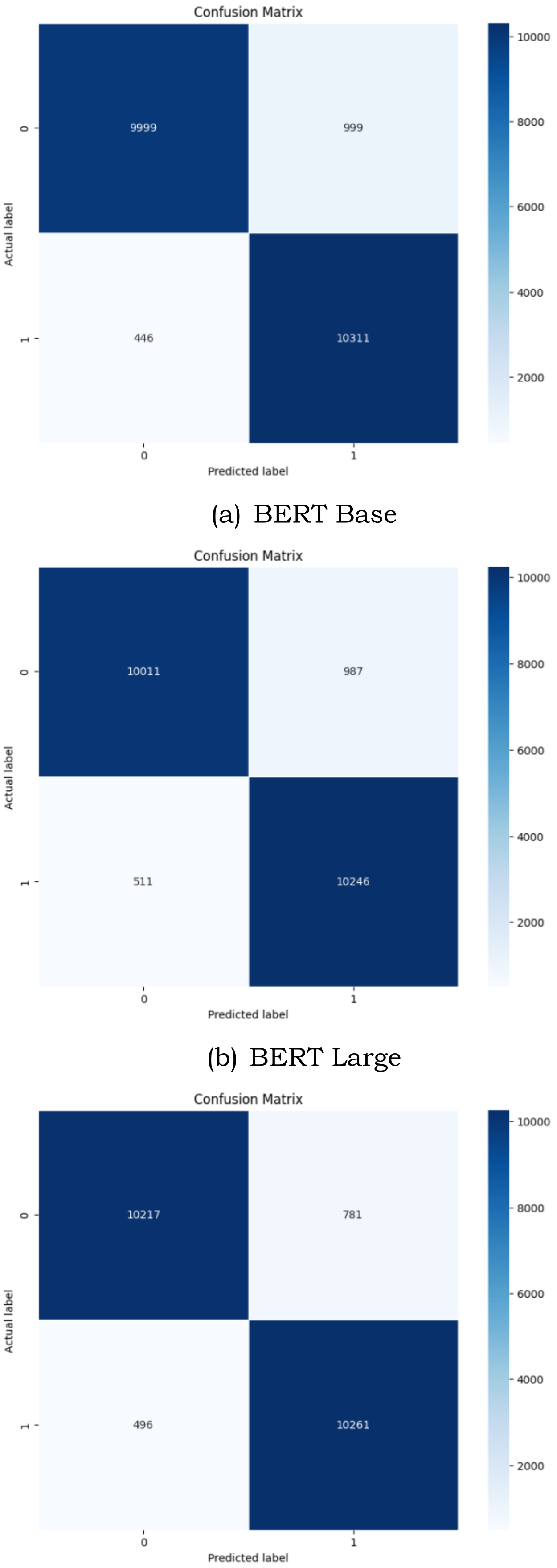
Next, the BERT Large model shows a precision of 95.1% and a recall of 91.0% for label 0, leading to an F1 score of 93.0%. For label 1, a precision of 91.2% and a recall of 95.2% is noted, also resulting in an F1 score of 93.2%. These results suggest that the BERT Large model exhibits very high performance for both labels.

Lastly, the BERT Multilingual model demonstrates a precision of 95.4% and a recall of 92.9% for label 0, leading to an F1 score of 94.1%. For label 1, a precision of 92.2% and a recall of 95.4% is noted, also resulting in an F1 score of 94.1%. These results show that the BERT Multilingual model exhibits high performance for both label 0 and 1. [*Table 4*]

All three BERT models show considerably high precision, recall, and F1 scores for both label 0 and 1. This suggests that they are highly effective for the classification of cyberbullying comments. Interestingly, when the same English dataset was used for the experiment, the Multilingual model was observed to produce superior results in all aspects, albeit marginally. This can be assumed to be due to the BERT Multilingual model, by learning various languages, is able to grasp diverse linguistic features, thereby enhancing its performance.

Models		Label	Precision	Recall	F-1Score
BERT	Base	0	95.7	90.9	93.3
		1	91.2	95.9	93.5
	Large	0	95.1	91.0	93.0
		1	91.2	95.2	93.2
	Multilingual	0	95.4	92.9	94.1
		1	92.2	95.4	94.1

Table 4. Comparison of Performances by BERT Models

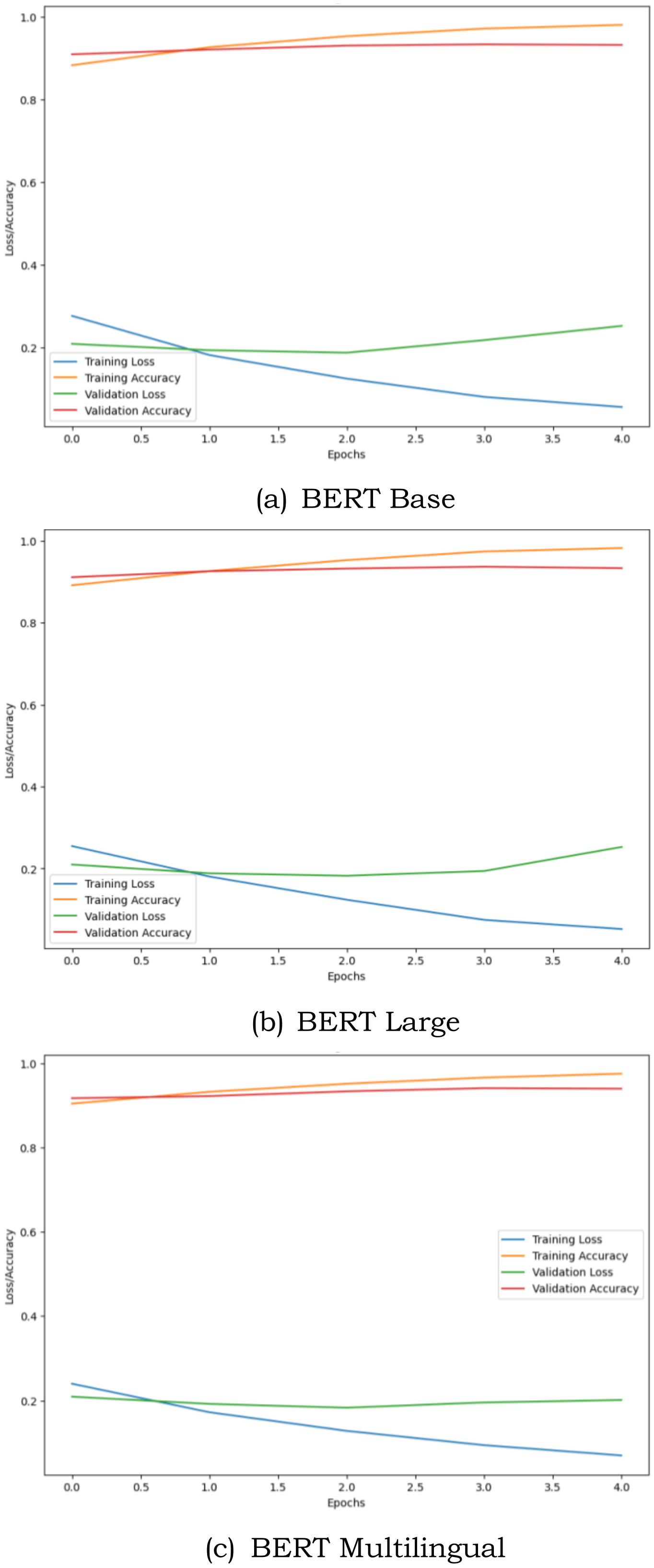


(a) BERT Base

(b) BERT Large

(c) BERT Multilingual

Figure 8. Confusion Matrix of BERT models by labels



(a) BERT Base

(b) BERT Large

(c) BERT Multilingual

Figure 9. Accuracy and Loss Graph of BERT models

Next, the performances of the three models of RoBERTa: Base, Large, and XLM, were assessed based on the Precision, Recall, and F-1 Score for each label. [*Table 5*]

Initially, the RoBERTa Base model presents a Precision of 95.2% and a Recall of 91.4% for label 0, resulting in an F-1 Score of 93.3%. For label 1, a Precision of 91.6% and a Recall of 95.3% are observed, culminating in an F-1 Score of 93.4%. These results highlight the accurate and balanced performance of the RoBERTa Base model for both labels.

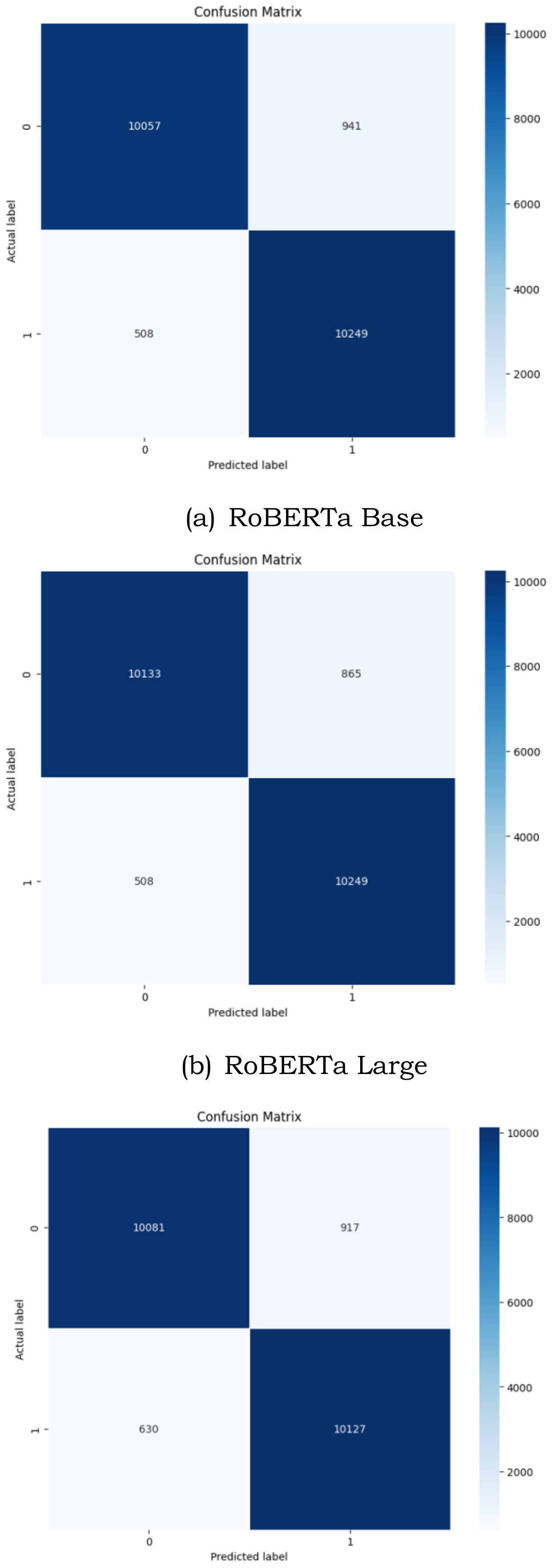
Subsequently, the RoBERTa Large model exhibits a Precision of 95.2% and a Recall of 92.1% for label 0, leading to an F-1 Score of 93.7%. For label 1, the Precision is 92.2%, and the Recall is 95.3%, which amounts to an F-1 Score of 93.7%. These outcomes denote the high performance of the RoBERTa Large model for both labels.

Lastly, the RoBERTa XLM model displays a Precision of 94.1% and a Recall of 91.7% for label 0, which leads to an F-1 Score of 92.9%. For label 1, the Precision is 91.7%, and the Recall is 94.1%, resulting in an F-1 Score of 92.9%. These results signify the RoBERTa XLM model's excellent performance for both label 0 and 1.

All three models of RoBERTa also exhibit notably high Precision, Recall, and F1 scores for both labels 0 and 1. Among these models, the Large model shows slightly higher performance for both cyberbullying and regular comments. Contrary to the BERT models where multilingual showed the highest performance, the RoBERTa model that classifies multiple languages, XLM, has lower performance than the RoBERTa Base model.

Models		**Label**	**Precision**	**Recall**	**F-1Score**
RoBERTa	Base	0	95.2	91.4	93.3
		1	91.6	95.3	93.4
	Large	0	95.2	92.1	93.7
		1	92.2	95.3	93.7
	XLM	0	94.1	91.7	92.9
		1	91.7	94.1	92.9

Table 5. Comparison of Performances by RoBERTa Models



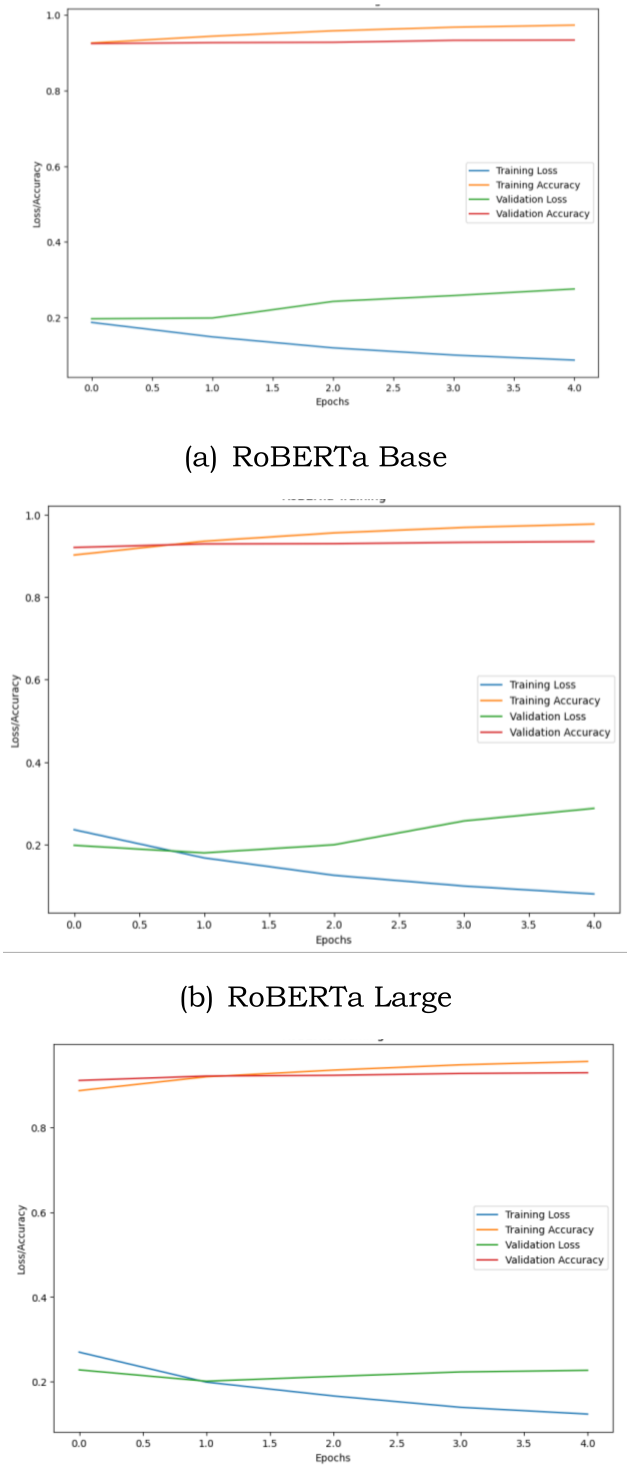
(a) RoBERTa Base

(b) RoBERTa Large

(c) RoBERTa XLM

Figure 10. Confusion Matrix of RoBERTa models by labels

However, a slight increase in validation loss was observed in the RoBERTa model. This phenomenon could suggest a possibility of the model overfitting to the training data, yet it does not serve as conclusive evidence of overfitting. To firmly ascertain overfitting, additional experiments or other overfitting indicators need to be referenced. Nonetheless, it cannot be denied that this phenomenon may have contributed, in part, to the performance improvement of the RoBERTa model. [*Figure 11*]



(a) RoBERTa Base

(b) RoBERTa Large

(c) RoBERTa XLM

Figure 11. Accuracy and Loss Graph of RoBERTa models

Finally, examining the performance of the DeBERTa model reveals significant results. For label 0, the DeBERTa Base model demonstrates a precision of 95.3%. In addition, the model yields a recall of 91.6%, and considering these two indices collectively, an F1 score of 93.4% is achieved.

Next, the DeBERTa Base model, when dealing with label 1, exhibits a precision of 91.7% and a recall of 95.4%. These high indices translate to an F1 score of 93.5%, indicating high precision and recall.[*Table 6*]

When compared with the base models of BERT and RoBERTa, DeBERTa shows a very slight performance improvement. The high performance of DeBERTa is further substantiated by the confusion matrix evaluated using the test data, where it exhibits performance metrics comparable to other models.

Models		Label	Precision	Recall	F-1Score
DeBERTa	Base	0	95.3	91.6	93.4
		1	91.7	95.4	93.5

Table 6. Comparison of Performances of DeBERTa Base

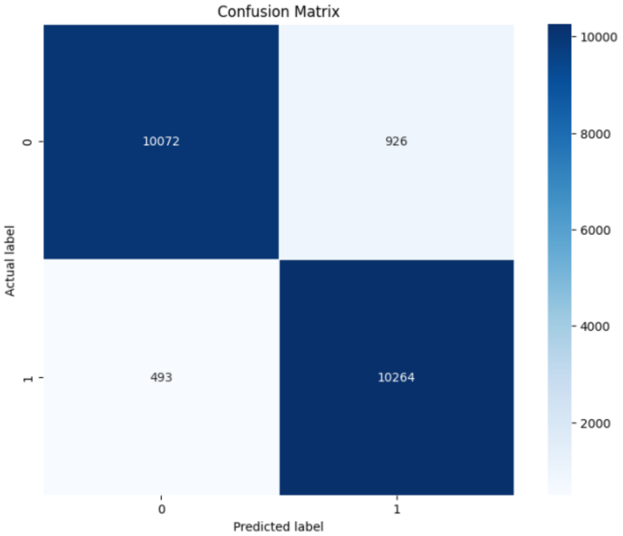


Figure 12. Confusion Matrix of DeBERTa Base by labels

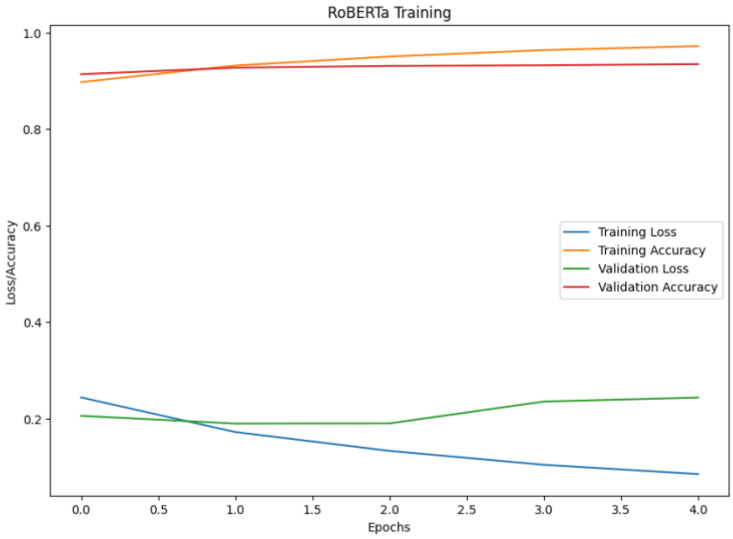


Figure 13. Accuracy and Loss Graph of DeBERTa Base

The comprehensive accuracy comparison among all models is presented as follows. Starting with the BERT model, the Base version demonstrated an accuracy rate of 93.4%, while the Large version delivered an accuracy of 93.1%. Significantly, the BERT Multilingual variant exhibited an impressive accuracy of 94.1%.

Moving onto the RoBERTa model, the Base version achieved an accuracy rate of 93.3% and the Large variant delivered an accuracy of 93.7%. Despite its performance being marginally lower compared to other models, the XLM version of RoBERTa maintained a commendable accuracy of 92.9%.

Lastly, the Base model of DeBERTa reported an accuracy of 93.5%, a performance on par with that of BERT and RoBERTa models.[*Table 7*]

In summary, current widely-utilized text classification models generally showcase high-level performance. While minute differences exist, the BERT Multilingual model emerged as the most suitable option for the cyberbullying dataset when tested on the same set of data.

These findings suggest that all models reliably performed at high standards. Consequently, to determine the most appropriate model for a given problem, careful consideration should be given to the unique attributes of each model and the relevant performance indicators.

Models		Accuracy
BERT	Base	93.4
	Large	93.1
	Multilingual	94.1
RoBERTa	Base	93.3
	Large	93.7
	XLM	92.9
DeBERTa	Base	93.5

Table 7. Comparison of Accuracy by Models

Further, trials were conducted on the dataset referenced in the Test section of an alternative Methodology using the preserved pre-trained model, with the results laid out below.[*Table 8*]

Models		Label	Precision	Recall	F-1Score	Accuracy
BERT	Base	0	85.8	83.8	84.8	84.9
		1	84.1	86.1	85.1	
	Large	0	84.7	89.1	86.8	86.6
		1	88.6	84.0	86.3	
	Multilingual	0	85.7	86.1	85.9	85.9
		1	86.2	85.8	86.0	
RoBERTa	Base	0	82.1	35.3	49.4	63.7
		1	58.7	92.2	71.7	
	Large	0	79.8	36.8	50.4	63.8
		1	59.0	90.7	71.5	
	XLM	0	81.7	46.7	59.4	68.3
		1	62.9	89.6	73.9	
DeBERTa	Base	0	80.9	33.3	47.1	62.9
		1	58.3	92.2	71.4	

Table 8. Comparison of Performances of Validation by models

BERT models, in all their variants (Base, Large, and Multilingual), demonstrate a steady show. In the Base version, both class 0 and class 1 had F1 scores nearly neck and neck at 84.8 and 85.1 respectively. The Large version held its ground too, with F1 scores of 86.8 for class 0 and 86.3 for class 1. The Multilingual version followed suit, with F1 scores of 85.9 for class 0 and 86.0 for class 1. In a nutshell, BERT models performed admirably across the board, with the Large variant standing out with the highest F1 scores.[*Figure 14*]

The RoBERTa models showed a different picture across their Base, Large, and XLM versions. The Base and Large versions struggled with class 0, scraping by with F1 scores of 49.4 and 50.4 respectively. However, when it came to class 1, they shone with impressive F1 scores of 71.7 and 71.5 respectively. The XLM version brought some balance to the table with F1 scores of 59.4 for class 0 and a robust 73.9 for class 1.[*Figure 15*]

In the DeBERTa model, the F1 score for class 0 was notably low at 47.1, compared to 71.4 for class 1, indicating that the model struggled to predict class 0 accurately.

In summary, while the BERT models demonstrated balanced performance across both classes, the RoBERTa and DeBERTa models underperformed in predicting class 0. This pattern suggests that the models were better tuned for predicting class 1.

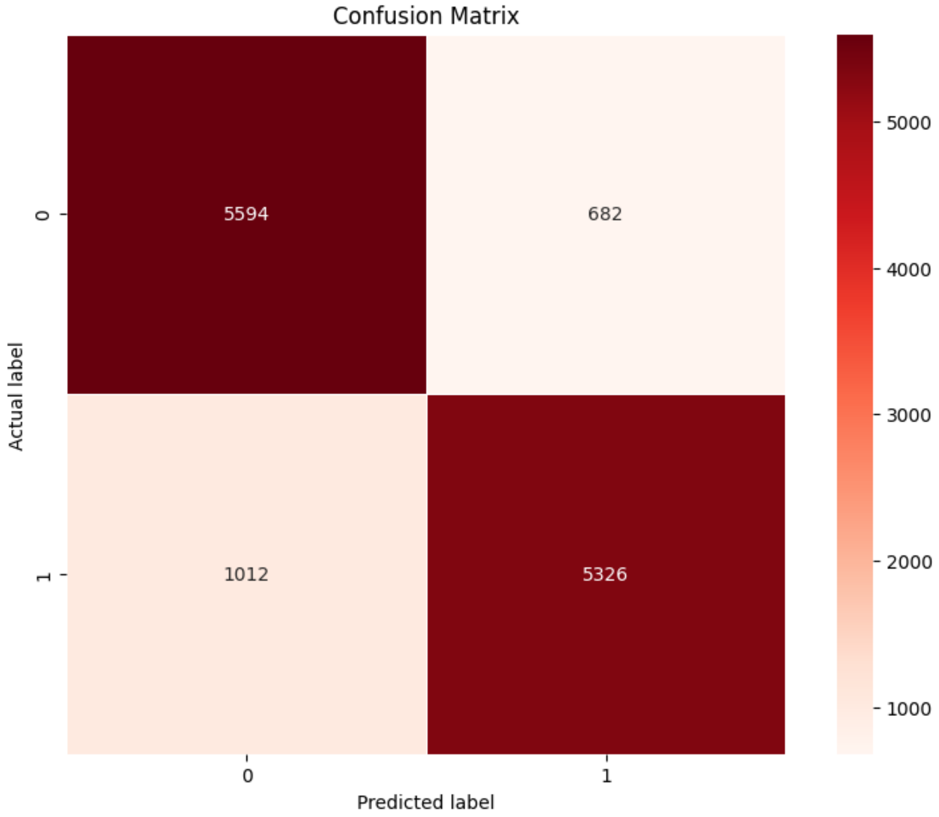


Figure 14. Confusion Matrix of BERT Large (Transfer Learning) by labels

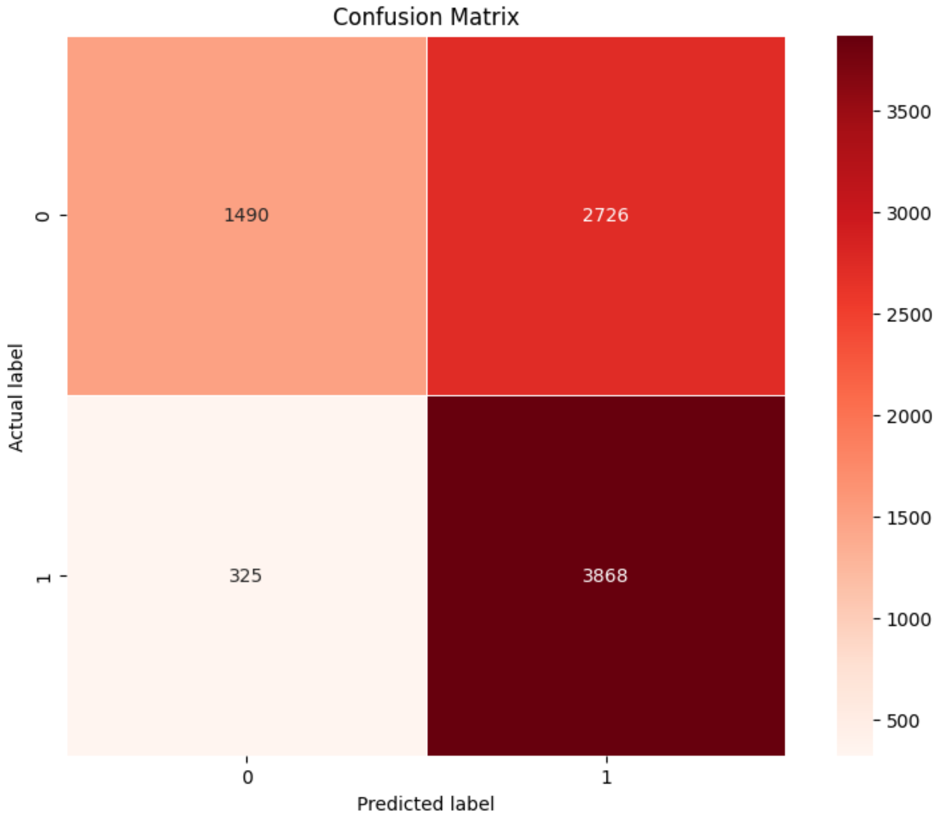


Figure 15. Confusion Matrix of RoBERTa Base (Transfer Learning) by labels

Interestingly, the models where only the weights were saved demonstrated relatively impressive results. On the contrary, the RoBERTa and DeBERTa models, where the entire model was saved, showed a noticeable decline in performance compared to their original training. Particularly, most models had a challenging time predicting general comments. Despite attempts at adjusting the learning rate and fine-tuning other details during the training process, no significant improvement was observed. This leads us to speculate that there may be inherent limits to transfer learning in text classification. Although fine-tuning is performed using pre-trained models, it appears that they can't be perfectly tailored to certain specific tasks. In the case of BERT, the models where only the weights were saved produced better results in the transfer learning process than the others. This suggests that the initial weights of each model may have a considerable influence on performance.

5 Conclusion and Future work

In this study, pre-training was conducted on various datasets using BERT's Base, Large, Multilingual, RoBERTa's Base, Large, XLM, and DeBERTa's Base models. During the pre-training phase, all models exhibited exceptional performance in cyberbullying comment classification, but opportunities for improvement were revealed during the fine-tuning process. Particularly, difficulties were encountered in the classification of generic comments during the fine-tuning stage. This has indicated the need for additional research into the optimization of fine-tuning strategies, considering their real-world application.

Additionally, the improved performance of the BERT Multilingual model, capable of handling multiple languages, underscored the importance of detection strategies for cyberbullying comments written in a variety of languages. It has highlighted the necessity for efforts to accurately detect cyberbullying comments in diverse linguistic environments, and called for research into model development that reflects such linguistic diversity.

Pursuing the research directions mentioned above can greatly enhance the performance of text classification models intended for the prevention of cyberbullying comments. This will serve as a crucial step towards managing cyberbullying issues more effectively and minimizing the harm caused by it.

6 References

1. Teens and Cyberbullying (Pew Research Center, 2022) https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/
2. Ann John, et al. "Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review." (2018).
3. Lee Sang Woo, 2019, (Korean National Human Right Commision) https://www.humanrights.go.kr/webzine/
4. Teens, Social Media and Technology (Pew Research Center, 2018) https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/
5. O Dae-suk, 2019, (Article) https://www.mk.co.kr/economy/view/2019/874232/
6. Xiang Zhang, et al. *"Cyberbullying detection with a pronunciation based convolutional neural network"*.Web.
7. Sweta Agrawal and Amit Awekar. *"Deep learning for detecting cyberbullying across multiple social media platforms"*.Web.
8. Cynthia Van Hee, et al. "Automatic detection of cyberbullying in social media text." *PloS one* 13.10 (2018): e0203794. Web.
9. Naver D2, https://d2.naver.com/helloworld/7753273
10. Fatma Elsafoury, et al. "When the timeline meets the pipeline: A survey on automated cyberbullying detection." *IEEE access* 9 (2021): 103541-63. Web.
11. Bob Sanders. *ANALYSIS OF A PRE-TRAINED BERT MODEL USED FOR CYBERBULLYING DETECTION IN TWEETS.* TILBURG UNIVERSITYWeb.
12. Mohammed Al-Hashedi, et al. "Cyberbullying Detection Based on Emotion." *IEEE Access* (2023)Web.
13. Yunhao CHEN and Hélene RONDEY. "Intent classification using contextual embeddings." Web.
14. Bayode Ogunleye and Babitha Dharmaraj. "Use of Large Language Model for Cyberbullying Detection." (2023).
15. Ashish Vaswani, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017)Web.
16. Zhilin Yang, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* 32 (2019)Web.
17. Kevin Clark, et al. "Electra: Pre-training text encoders as discriminators rather than generators." (2020).
18. Jacob Devlin, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." (2018).
19. Yinhan Liu, et al. "Roberta: A robustly optimized bert pretraining approach." (2019).
20. Alexis Conneau, et al. "Unsupervised cross-lingual representation learning at scale." (2019).
21. Pengcheng He, et al. "Deberta: Decoding-enhanced bert with disentangled attention." (2020).
22. Datasets - https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset Accessed 28 April 2023.
23. Datasets-https://www.kaggle.com/code/yemi99/fine-tuning-bert-with-tf-text-classification/input Accessed 28 April 2023

24. Datasets - https://huggingface.co/datasets/SotirisLegkas/binary_off_hate_toxic_new Accessed 10 June 2023
25. Chi Sun, et al. *"How to fine-tune bert for text classification?"*.Web.
26. Hee Jung Choi, et al. "Stanford MLab at SemEval-2023 Task 10: Exploring GloVe-and Transformer-Based Methods for the Explainable Detection of Online Sexism." (2023).
27. Yoo, J., et al. "A Comparative Study of Transformer-based Language Models: BERT and GPT-2." Proceedings of the Korean Society for Information Processing Conference 29.1.(2022)
28. Park, H., & Kim, K. "Recommendation System Using Bert-based Sentiment Analysis." Journal of Intelligent Information Systems 27.2 (2021)
29. Jamil M. Saquer "Application of Artificial Intelligence and Graphy Theory to Cyberbullying" (2020).
30. Elaheh Raisi. "*Weakly Supervised Machine Learning for Cyberbullying Detection*" (2019).