BERT Based Models for Multiclass Cyberbullying Detection

Ali Jassim Mohammed and Wafaa Mohammed Saeed
Department of Software Engineering, College of Information Technology, University of Babylon, 51001 Hillah, Iraq
AliJassimM.sw@student.uobabylon.edu.iq, it.wafaa.mohammed@uobabylon.edu.iq

Keywords: BERT for Cyberbullying, Cyberbullying Detection, Deep learning. Machine Learning Models, Transformer Models.

Abstract: Cyberbullying is a complicated digital social media challenge that has intensified with the rapid spread of social media platforms. It takes many categories, including hate speech, verbal abuse. Due to the large volume of text generated daily, traditional methods based on rules or human review are no longer sufficient for accurate and effective detection of this phenomenon. The objective of this research is to evaluate the performance of a fine-tuned BERT model in multi-category cyberbullying detection and compare them with several traditional machine learning algorithms, including Support Vector Machine (SVM), logistic regression, and random forests. In this research we use a multi-category cyberbullying dataset, and the BERT model evaluated using performance metrics such as accuracy, precision, recall, and F1 score, The experimental showed that the BERT model achieved the best result overall performance metrics, where an accuracy is 0.8474, outperforming the conventional machine learning models such as SVM, RF and LR, the result of this research demonstrated the efficiency and ability of BERT in detection of multiclass cyberbullying, Furthermore this research illuminates the significant potential of transformer-based models in handling complex contextual and semantic information within social media cyberbullying texts.

1 INTRODUCTION

In the current time, the internet is widely used for communication, which may lead to the spread of harmful behaviors, most notably cyberbullying, which is an extension of traditional bullying but in digital ways via social media [1]. Cyberbullying is a type of online harassment that includes sending abusive messages, posting hurtful comments, spreading rumors, sharing inappropriate pictures or videos, and cyberstalking. [2], [3]. Cyberbullying is considered a complex phenomenon that can be multifaceted, which does not take one form or one type of aggressive behavior, rather, it takes different types that reflect multiple ways of abuse and harm through digital media. Cyberbullying can be categorized based on the nature of action used and the platform through which it occurs, to help to this phenomenon to understanding deeper and development a new effective strategy to overcome this phenomenon. There are several types below that can be briefly [4]-[6]. Flooding this type of cyberbullying represented by flooding the victim through a repetitive and unreasonable comments or post. That target is to disturb and hinder his ability to effectively participate in discussions or activation on social media platforms. this type of cyberbullying is often used in forums or chat rooms where the constant repetition distracts and isolates the victim from normal interaction. Masquerade this type of cyberbullying happens when the bully disguises the victim through creating a fake account or imitating by writing and communication style, the target of this cyberbullying to deceive others or damage the victim's reputation by posting offensive content attributed to the victim. Flaming or bashing This type of cyberbullying refers to the outbreak of verbal altercations online, where the bully sends messages filled with insults, hurtful language, and obscene content. This can occur privately between two individuals or publicly within groups and forums. Trolling The bully deliberately plays the role of a "provocateur," posting controversial or disagreeing comments with the intent of igniting conflict or provoking anger. Although these comments may not appear directly offensive or offensive, Harassment occurs when the bully repeatedly sends messages full of insults and hurtful phrases to the victim, these messages are mostly sent via an email. Denigration In this type of cyberbullying, the bully spreads rumors

or false claims intended to damage the victim's reputation or social relationships. These lies may be disseminated via social media platforms. Outing occurs when the bully reveals personal or embarrassing information about the victim and shares it on social media platforms without permission This information is often acquired by the bully as a result of a previous relationship with the victim. Exclusion this type of cyberbullying is embodied in exclusion the victim from sharing posts in groups or any activity across the internet, such as remove name from a chat group or ignoring the invitation to play a video game. Cyberbullying is not limited to these types only, but there are also aspects and forms of cyberbullying that are often such as Cyberstalking, Impersonation, and more [7].

the research objective is to evaluate performance of fine-tuned BERT model for multiclass cyberbullying detection compared to traditional machine learning algorithm such as Support Vector Machine and Naive Bayes, the significance of this research is to providing a comparative analysis with in a multiclass datasets, That reflect the complex nature of cyberbullying, while demonstrating the potential impact of proposed model in improving the accuracy of automated detection system and fostering safer digital environment.

2 RELATED WORK

In recent years, cyberbullying has gained attention from research, especially with the growing reliance on machine learning and deep learning techniques. For instance. Lwendi et al. [8] studied the performance of serval deep learning model, such as RNN, LSTM, and GRU, using a binary dataset taken from the Kaggle platform, the result show the Bi-LSTM outperformed compared to other model, reflecting the ability to handle cyberbullying data.

In the other haned, Paul et al. [9] proposed BERT (Bidirectional Encoder Representations from Transformers) model to detect cyberbullying, the model was evaluated basned on their datasets Formspring, Twitter, and Wikipedia, each one containing about 12k, 16k, and 100k, and a binary labale. the study's results confirmed that BERT significantly outperformed traditional models such as SVM and logistic regression, and demonstrated better performance compared to other deep learning models like CNN and Bi-LSTM enhanced with attentional mechanisms.

The reasech, Muneer et al. [10] presented a framework based on ensemble stacking using datasets from Facebook and Twitter, integrating several models such as CNN and Bi-LSTM, while leveraging from word representations using CBOW and Skip-gram. The study demonstrated that this ensemble approach achieves better results than relying on a single traditional machine learning model.

The study by Dewani, et al.[11] aimed to develop an active framework to detecting cyberbullying in Roman Urdu Database based on deepening models, particularly RNN-LSTM, RNN-Bi-LSTM, and CNN, the overall data preprocessing include handling Unicode encodings and encoded text formats, cleaning noisy text, and removing stop words to enhance data quality and improve training efficiency, the study show that RNN-LSTM and RNN-Bi-LSTM outperformed on other models and archive accuracy 85.5%, 85% respectively with an F1-Score 0.70%, 0.67%. These findings indicate their strong ability to capture and model complex linguistic patterns in Roman Urdu texts effectively.

In the field of Arabic, Saadi et al. [12] focused on using an SVM model enhanced with the Cuckoo Search optimization algorithm, depending on TF-IDF text representations and Count Vectorization. The results showed that incorporating the optimization algorithm increased the classification accuracy from 85.8% to 87.1%, highlighting the effectiveness of this approach in improving the performance of cyberbullying detection in Arabic texts.

Nevertheless, a comprehensive evaluation of transformer-based models within multiclass cyberbullying detection remains limited, highlighting the need for further systematic investigation.

3 METHODOLOGIES

In this research, we use a BERT model to detect cyberbullying in multiclass datasets. The proposed approach seeks to classify online text into multiple cyberbullying categories rather than rely on binary classification. BERT is fine-tuned on labeled data for capture contextual and semantic information within text more effective. This capability allows the model to better understand subtle linguistic patterns and implicit abusive language. The performance of the model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the BERT-based model achieve strong performance across all classes. The results indicates that transformer-based models is well-suited for complex cyberbullying detection tasks. Compared to traditional machine learning approaches, BERT provide more accurate and reliable predictions. Overall, this research

highlights the effectiveness of BERT for multiclass cyberbullying detection.

3.1 Overview of the proposed Approach

The proposed fine-tuned BERT model consists of several steps to detect multiclass cyberbullying, starting by dataset preparation, where the datasets divided into training and test sets, followed by text preprocessing to clean and standardize the data before model training and then fed into to the BERT model to training dataset, finally evaluation model by using performance metrics such as accuracy, precision, recall, and F1-score, Figure 1 illustrates the general steps of proposed model.

3.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a deep linguistic model developed by Google AI in 2018. BERT is considered one of the revolutionary models in NLP, it introduces for the first time a fully Bidirectional representations of the text, enabling to understandings full context of a word based on each word from both fronts and back in the sentence. BERT was designed based on the Transformer architectures and depending on the Encoder part from Transformer. Additionally, BERT was pre-trained on a massive amount of data using an unsupervised task, then reuse or fine-tuned on a specific task such as classification, question and answer, and entities recognition [13]. BERT architectures are illustrated in the Figure 2 [14].

3.3 Input Representation

In BERT, each text should be transformed into a numerical representation consisting of [13]:

- Token Embedding. Each text is converted into a unit called Tokens by using a tool called WordPiece, then this token is converted to a numerical representation based on a model dictionary.
- Segment Embedding. Is used when two sentences are entered in the Next Sentence Prediction(NSP) task. Segment IDs are used to distinguish between each sentence (0 for the first sentence and 1 for the second sentence).
- Position Embedding. Due to the Transformer not understanding the order of the words, Position Embedding is added to determine the order of each word in the sentence.

The final representation of each word = Token Embedding + Segment Embedding + Position Embedding, illustration in Figure 3 [15].

3.4 BERT Component

BERT is composed of several essential components, each of which plays a critical role in enabling the model to capture deep contextual representations of language. These components are detailed below[16].

1) Pre-training task in BERT. BERT model in pre-training depending on the specific tasks, targeted to help understanding language and context: these tasks split into two main categories.
2) Masked Language Model (MLM). In the MLM, some words are randomly hidden from input, and required from the model predict missing words based on the full context.
3) Next Sentence Predication (NSP). In the NSP, required from the model to determine whether the second sentence actually follow to the first sentence in the text, which help the model to understand the relationship between sentence.

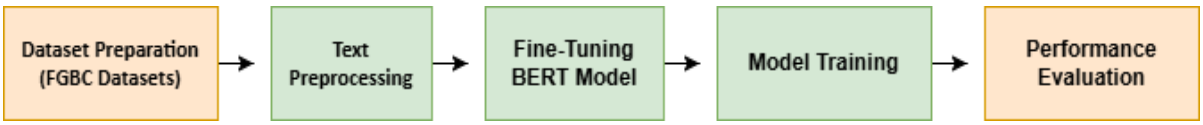

Figure 1: General steps of proposed BERT model.

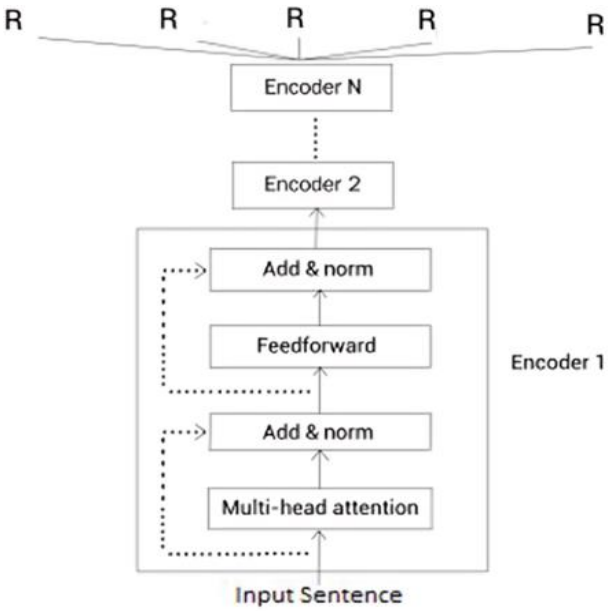

Figure 2: The BERT architecture.

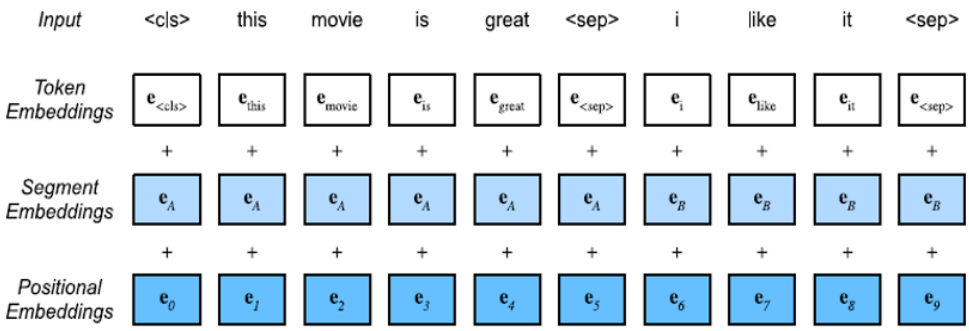

Figure 3: BERT Embedding.

3.5 Fine-Tuning for Multiclass Cyberbullying Detection

For the cyberbullying detection task, the pre-trained BERT model is fine-tuned by adding a task-specific classification layer on top of the [CLS] token representation. The [CLS] token serves as an aggregate representation of the entire input sequence and is commonly used for classification tasks.

During fine-tuning, all model parameters are updated jointly using labeled cyberbullying data. The model is optimized using a multiclass cross-entropy loss function, enabling it to learn decision boundaries among multiple cyberbullying categories as well as non-cyberbullying content. Fine-tuning allow BERT to adapt its generic language representation to domain-specific patterns such as offensive expressions, implicit harassments, and context dependent abuse.

3.6 Advantages of BERT for Cyberbullying Detection

BERT is particularly well-suited for cyberbullying detection for several reasons. First, its bidirectional attention mechanism enables the model to capture contextual nuance and implicit meanings that are common in abusive language, including sarcasm and indirect insult. Second, the use of subword tokenization allow BERT to handle misspelling, slangs, and informal language frequently observed in social media text. Finally, the fine-tuning process allows the model to generalize across different cyberbullying categories, making it effective for multiclass classification scenarios.

3.7 Dataset and Preprocessing

In this research, we use a IEEE FGBC (Fine-Grained Balanced Cyberbullying) dataset [17]. that has more than 47000 tweets taken from a twitter, its categories into six major cyberbullying category, age, ethnicity, gender, religion and other type of cyberbullying, in addition into not cyberbullying category, each of this category contain approximately 8000 of sentence, cyberbullying and distributing data consistently and systematically. This strategy enhances the accuracy and reliability of predictions in multi-category classification and ensures that the models can handle complex real-world texts prevalent on social media platforms, minimizing bias towards underrepresented categories and thus achieving higher performance and more credible results. to ensure that all categories were equally represented and to minimize model bias towards the most common categories, Balancing categories is crucial in multiclass classification tasks, as it helps improve the models' ability to fairly and accurately distinguish between all types of bullying and reduces the likelihood of overlooking underrepresented categories that could be critical in real-world applications. The prep-processing applied to the dataset before passing to the model that including data cleaning non-text symbols, special characters, and links such as https or https. This stage focused on remove noise form datasets with preserving the core meanings, which contributes to improving the model's ability to learn and extract important features from texts in datasets. Figure 3 presents the distribution across this type. Figure 4 shows the distribution of word length in IEEE FGBC datasets. Figure 5 shows the most frequent words of each category in the Cyberbullying datasets.

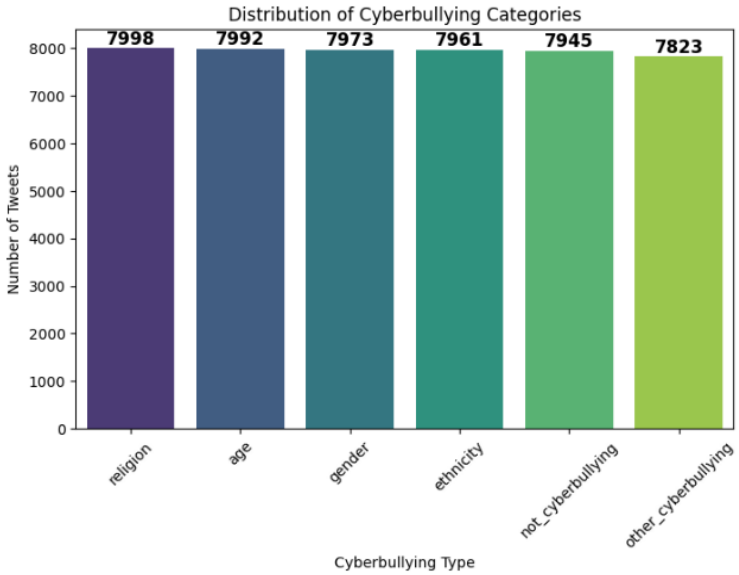

Figure 4: Distribution of cyberbullying types in IEEE FGBC (Fine-Grained Balanced Cyberbullying) datasets.

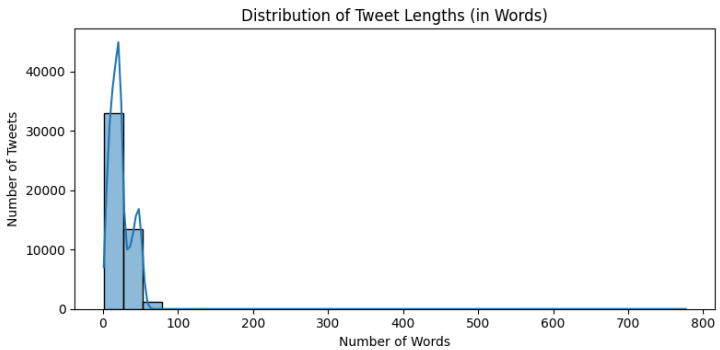

Figure 5: Distribution of word length in IEEE FGBC (Fine-Grained Balanced Cyberbullying) datasets.

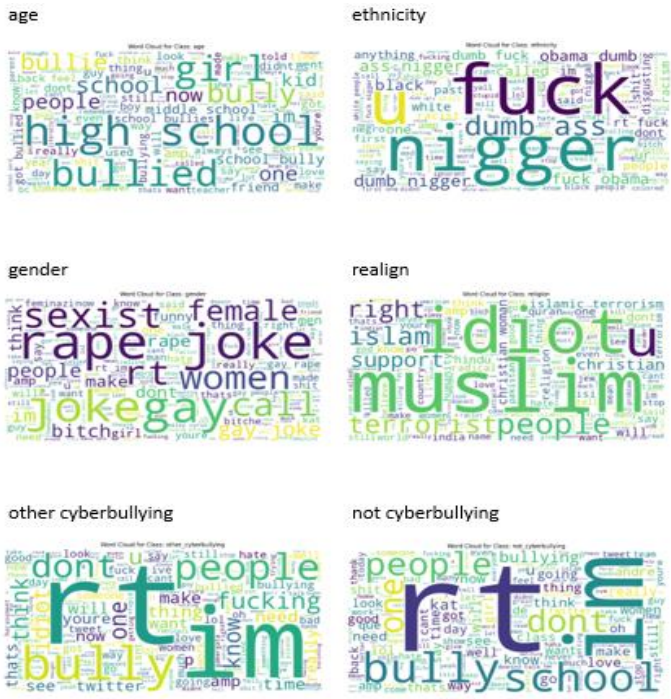

Figure 6: The most frequent words of each category in the IEEE FGBC (Fine-Grained Balanced Cyberbullying) datasets.

Table 1: Hyperparameters for our BERT model.

Parameter	Value	Description
BERT Checkpoint	bert-base-uncased	Pre-trained BERT model used as the base for fine-tuning.
Tokenizer	BertTokenizer	Tokenizer used to convert raw text into tokens compatible with BERT.
Train/Test Split	80/20	Percentage split of the dataset for training and testing.
Optimizer	AdamW	Optimizer used for fine-tuning BERT with weight decay.
Learning Rate	2e-5	Learning rate for training.
Batch Size	16	Number of samples per batch during training.
Epochs	5	Number of complete passes through the training dataset.

Table 2: Hyperparameters for baseline models.

Model	Text Representation	Key Hyperparameters
SVM	TF-IDF (max_features = 5000, n-grams = 1–2)	Kernel = Linear, C = 1.0, Class Weight = Balanced
Random Forest (RF)	TF-IDF (max_features = 5000, n-grams = 1–2)	n_estimators = 200, max_features = √, max_depth = None, class_weight = Balanced
Logistic Regression (LR)	TF-IDF (max_features = 5000, n-grams = 1–2)	Solver = lbfgs, Penalty = L2, C = 1.0, max_iter = 1000, Class Weight = Balanced

3.8 Performance Metrics Explained

Performance evaluation is an essential component of machine learning research, providing objective measures for assessing the effectiveness of classification models. In cyberbullying detection studies, evaluation metrics are used to determine how accurately a model can distinguish harmful content from non-harmful content. This study employs four widely used metrics: accuracy, precision, recall, and F1-score [18].

Accuracy measures the overall proportion of correctly classified instances and provides a general indication of the model's predictive performance. Precision evaluates the reliability of positive predictions by measuring how many instances

identified as cyberbullying are actually cyberbullying cases. Recall assesses the model's ability to detect relevant positive instances and reflects how effectively cyberbullying content is identified within the dataset. Finally, the F1-score combines precision and recall into a single measure, providing a balanced assessment of classification performance, particularly when class distributions are uneven. Together, these metrics offer a comprehensive evaluation of the proposed model by considering overall correctness, prediction reliability, detection capability, and the balance between precision and recall [18].

A confusion matrix is a performance tool used in a classification task, that summarizes prediction results compared to actual values (true values). This matrix helps to understand how well the model performs the classification task, especially to distinguish between different classes. it is particularly useful in binary and multi-class classification problems [19]. Figure 7 illustrates the confusion matrix.

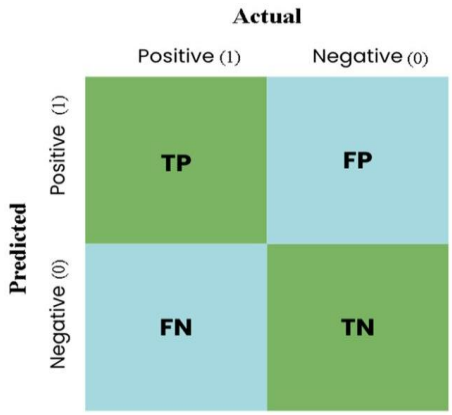

Figure 7: The confusion matrix.

4 HYPERPARAMETERS FOR OUR BERT MODEL

The BERT model was fine-tuned using selected hyperparameters to optimize performance for multi-class cyberbullying detection. Key settings, including the pre-trained checkpoint, tokenizer, optimizer, learning rate, batch size, and number of epochs, were chosen based on best practices and empirical testing to ensure effective learning and stable convergence.

The selected hyperparameters are summarized in Table 1 below.

5 HYPERPARAMETERS FOR OUR BASSLINE MODEL

For comparison, traditional machine learning models were implemented as baselines using TF-IDF features to represent text. These models include Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). Key hyperparameters, such as kernel type, regularization strength, number of estimators, and maximum tree depth, were set based on standard practices to ensure fair evaluation. The baseline models and their parameters are summarized in Table 2 below.

6 RESULTS AND DISCUSSION

The four models were trained and tested on the IEEE FGBC dataset used, which represents tweets categorized into six cyberbullying categories. Table 3 presents the main evaluation metrics for each model, including accuracy, precision, recall, and F1-score. And Figure 6 shows the confusion matrix for each class in each model.

Table 3: Evaluation matrix for each model.

Model	Accuracy	Precision	Recall	F1-score
SVM	0.8382	0.8429	0.8382	0.8380
Logistic regression	0.8306	0.8362	0.8306	0.8320
Random forest	0.8231	0.8259	0.8231	0.8232
ourBERT	0.8441	0.8452	0.8447	0.8432

6.1 Comparative Evaluation Framework

To evaluate the performance of a fine-tuned BERT model in multiclass cyberbullying detection against traditional machine learning algorithms, a unified experimental framework was adopted using the FGBC dataset, containing six cyberbullying classes. Traditional models (SVM, Logistic Regression, Random Forest) were trained on preprocessed, numerically represented text, while BERT was fine-tuned by adapting its classification layer to the dataset. All models were evaluated using accuracy, precision, recall, and F1-score, with confusion matrices generated for class-level analysis, ensuring a fair and systematic comparison.

6.2 Performance Analysis

The Table 3 shows that the BERT model, designed for multi-category classification, outperformed the three traditional machine learning models, achieving the highest accuracy of 84.41%, along with the best Precision, Recall, and F1-score values. This is due to the ability of BERT to handle context and capture deep semantics of data.

In the traditional machine learning method, we noted that SVM outperformed other traditional machine learning methods, such as LR and RF. This is attributed to the efficiency of SVM in handling high-dimensional space, where text is converted into a digital representation before passing into the training process.

The Logistic Regression model shows average performance compared to the other models used, while RF showed the lowest performance accuracy among traditional machine learning models. This is due to its reliance on single-attribute tree aggregation, which reduces the ability to capture contextual meaning between words.

Based on the above results these findings highlight the importance of using pre-trained transformer models in cyberbullying applications such as BERT, as their ability to understand the context of texts gives them a clear advantage over traditional models. In addition, these models can be applied in practical applications such as content moderation systems on social media platforms. Figure 8 illustrates the confusion matrices of the SVM, LR, RF, and BERT models used in the experiments [20].

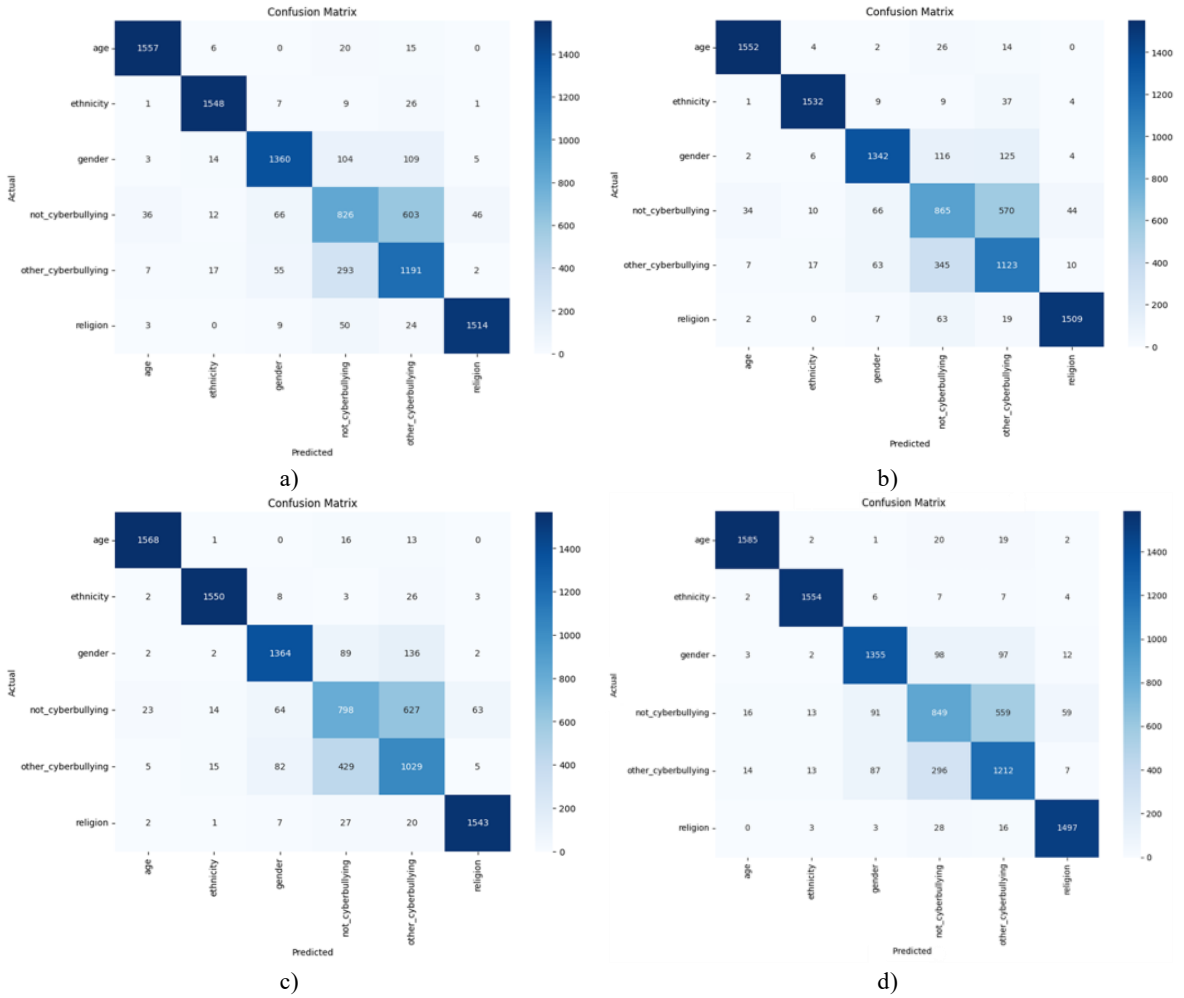

Figure 8: The confusion matrix of: a) SVM, b) LR, c) RF, d) BERT.

7 LIMITATIONS OF THE STUDY

Although the BERT model achieved promising results, this study has several limitations. First, the experiments were conducted on a single dataset, which may restrict the generalizability of the findings to other datasets and real-world scenarios. Second, the dataset was limited to the English language, which prevents evaluating the model's performance in multilingual or cross-cultural contexts. Additionally, the comparison was limited to traditional machine learning models, without including other advanced transformer-based architectures. Therefore, future studies should incorporate more diverse datasets, multilingual resources, and state-of-the-art language models to further validate the robustness and generalizability of cyberbullying detection approaches.

8 CONCLUSIONS

This research presented a comparative analysis of conventional machine learning algorithms and fine-tuned BERT for multiclass cyberbullying detection. The experimental results demonstrated that traditional machine learning models, including SVM, Logistic Regression, and Random Forest, achieved lower performance across evaluation metrics compared with the BERT-based approach. This difference can be attributed to the limited ability of conventional algorithms to capture complex contextual and semantic patterns in cyberbullying-related text.

In contrast, the fine-tuned BERT model showed superior capability in distinguishing between different categories of cyberbullying, achieving higher accuracy and improved performance across all evaluation metrics. These findings highlight the effectiveness of transformer-based architectures for complex text classification tasks and confirm their potential for reliable cyberbullying detection systems.

Future research should focus on expanding datasets with more diverse cyberbullying categories collected from multiple social media platforms, such as Instagram and Facebook, to improve model generalization. Furthermore, investigation of advanced transformer architectures, including RoBERTa, DeBERTa, and BigBird, together with explainable AI techniques, may provide deeper insights into model decision-making and support the development of more transparent and trustworthy cyberbullying detection solutions for real-world applications.

REFERENCES

[1] G. M. Abaido, "Cyberbullying on social media platforms among university students in the United Arab Emirates," Int. J. Adolesc. Youth, vol. 25, no. 1, pp. 407-420, 2020.

[2] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," Comput. Human Behav., vol. 63, pp. 433-443, 2016.

[3] D. Van Bruwaene, Q. Huang, and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," Lang. Resour. Eval., vol. 54, no. 4, pp. 851-874, 2020.

[4] I. Alanazi and J. Alves-Foss, "Cyber bullying and machine learning: A survey," Int. J. Comput. Sci. Inf. Secur. (IJCSIS), vol. 18, no. 10, 2020.

[5] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in 2016 European Modelling Symposium (EMS), IEEE, 2016, pp. 165-171.

[6] D. Maher, "Cyberbullying: An ethnographic case study of one Australian upper primary school class," Youth Studies Australia, vol. 27, no. 4, pp. 50-57, 2008.

[7] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, 2007.

[8] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," Multimedia Systems, pp. 1839-1852, Sep. 2023, [Online]. Available: https://doi.org/10.1007/s00530-020-00701-5.

[9] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification," Multimedia Systems, pp. 1897-1904, Dec. 2022, [Online]. Available: https://doi.org/10.1007/s00530-020-00710-4.

[10] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," Information (Switzerland), vol. 14, no. 8, Sep. 2023, [Online]. Available: https://doi.org/10.3390/info14080467.

[11] Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data," J. Big Data, vol. 8, no. 1, Dec. 2021, [Online]. Available: https://doi.org/10.1186/s40537-021-00550-7.

[12] M. Q. Saadi and B. N. Dhannoon, "Arabic Cyberbullying Detection Using Support Vector Machine with Cuckoo Search," Iraqi Journal of Science, vol. 64, no. 10, pp. 5322-5330, 2023, [Online]. Available: https://doi.org/10.24996/ijs.2023.64.10.37.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2019, pp. 4171-4186, [Online]. Available: https://doi.org/10.18653/v1/N19-1423.

[14] S. Ravichandiran, Getting Started with Google BERT: Build and Train State-of-the-Art Natural Language Processing Models Using BERT. Packt Publishing Ltd, 2021.

[15] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, Dive into Deep Learning. Cambridge University Press, 2023.

[16] Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," Trans. Assoc. Comput. Linguist., vol. 8, pp. 842-866, 2020.

[17] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 1699-1708.

[18] A. M. El Koshiry, E. H. I. Eliwa, T. Abd El-Hafeez, and M. Khairy, "Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique," PeerJ Comput. Sci., vol. 10, p. e1961, 2024.

[19] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham: Packt Publishing Ltd, 2019.

[20] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T. H. Kim, and I. Ashraf, "RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features," IEEE Access, vol. 12, no. May, pp. 58950-58959, 2024, [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3386637.