8th IEEE International Conference and Workshops on Recent Advances and Innovations in Engineering- ICRAIE 2023
(IEEE Record #59459)

Building Prediction Model for Detecting Cyberbullying using TikTok Comments

Bunga Aura Prameswari
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bunga.prameswari@binus.ac.id

Haliza Syafa Oktaviani
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
haliza.oktaviani@binus.ac.id

Titus Rangga Wicaksono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
titus.wicaksono@binus.ac.id

Biben Pieter Leonard
International Business Management Program Management Department
BINUS Business School Undergraduate Program
Bina Nusantara University
Jakarta, Indonesia 11480
biben.leonard @binus.ac.id

Said Achmad
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
said.achmad@binus.edu

Rhio Sutoyo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
rsutoyo@edu

***Abstract*—Easily accessible Internet and the need for communication led to the widespread use of social media platforms, e.g., TikTok. Social media platforms offer opportunities for social interaction and enter- tainment but also introduce risks such as fake news, fraud, and cyberbullying. The latter severely threatens the victim's mental health and overall well-being. This research collects, annotates, and analyzes 1,508 TikTok comments concerning cyberbullying behavior. Then, the comments were used to build a Deep Learning BERT (Bidirectional Encoder Representations from Transformers) architecture prediction model. This ex- periment will use pre-trained model BERT and fine- tuning BERT. After fine-tuning, the prediction model achieves 0.63 validation accuracy. This research high- lights the importance of building a prediction model to detect cyberbullying and contribute to a better understanding and prevention of this behavior on social media platforms.**

Keywords— Cyberbullying, TikTok, BERT, Deep Learning, Sentiment Analysis

I. INTRODUCTION

Internet accessibility has become commonplace across the globe due to advancements in technology. The internet covers various fields, from economics, entertainment, and security to social. In the social area, users can interact and expand social networks with each other and even with various countries that have internet access through social media. We Are Social, a global creative agency, reports that active social media users in Indonesia as of January 2022 have reached 191.4 million people or around 68.9% of Indonesia's population [1].The increasing use of social media increases the avail- ability of natural language data. These data can be utilized for research to build various prediction models, such as sentiment analysis for customer review [2], comparative study using Twitter data [3], and many more. However, social media can be a place for spreading fake news, fraud, and cyberbullying. Cyberbullying is aggressive behavior directed at victims and carried out by a group or indi- vidual using electronic media repeatedly. It can affect the mental health of victims [4]. This behavior risks damaging the victim's self and mental health, lowering their self- confidence, increasing the risk of psychological disorders, and even causing the victim's desire to end their life [5].

Indonesia ranks second with the most TikTok social media users, namely around 109.9 million users, after the United States with 113.25 million users, according to We Are Social [1]. Until the third quarter of 2022, TikTok gained 1.53 billion monthly active users, up around 4.64% from the previous month worldwide. When viewed from the number of TikTok users in Indonesia and statistical data, the age range of most TikTok users is 16-24 years, and the average user spends 52 minutes in one visit. TikTok is a social media platform that allows short videos to be created and uploaded with a vast network. This application has become popular among all ages due to its easy use. TikTok is a platform that is a place for its users to show their creativity.TikTok already has application features to prevent cyberbullying, including Comment Filters, Report and Block, Privacy Settings, Emotional Support, Awareness Campaigns, and TikTok Safety Team as its security team. In addition, TikTok also has online safety experts and psychologists who continuously monitor content and user behavior to prevent bullying and other online crimes. This shows that TikTok cares about its users by taking anti-cyberbullying preventive measures. However, these cyber-bullying prevention features still require human touch because it is done manually, and TikTok usually takes about 24 to 48 hours to review users' reports.Sentiment analysis can help to detect cyberbullying behavior because it can help determine specific texts that contain positive, neutral, or negative sentiments. This sen- timent determination has helped several

2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 979-8-3503-1551-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRAIE59459.2023.10468424

studies to detect cyberbullying behavior in social media [6]–[8].

This study uses a similar approach and aims to analyze cyberbullying behavior on TikTok by utilizing the com- ments column as a dataset that aims to detect and prevent cyberbullying on TikTok with a high degree of accuracy. This work independently collects and utilizes comments on the TikTok platform as a dataset. Moreover, the deep learning approach, specifically the Bidirectional Encoder Representations from Transformers (BERT) method, will be used for sentiment analysis. By applying this method, this study seeks to gain insight into the prevalence and characteristics of cyberbullying on TikTok.

II. RELATED WORKS

A. Sentiment Analysis

Sentiment analysis is mining text contextually to extract and identify subjective information, referring to various natural language processing, computational linguistics, and text mining to determine the tone of a text, whether it is a sentence, a comment, or an entire document, and returning a score that describes which contains positive, neutral, or harmful content, and eventually can lead to the main focus of this study to detect cyberbullying behavior.

Machine learning, lexicon-based, and hybrid approaches are some basic techniques that can be used for sentiment analysis [9], [10]. This sentiment determination has helped several studies to get an overview of public opinion on par- ticular topics, analyze customer satisfaction with certain products, brands, or services [2], detect crime prediction [11], and detect cyberbullying behavior [6]–[8], [12]–[15] in various social media, including Instagram, Twitter, and Facebook. Sentiment analysis can also create a system that uses textual, visual, and audio elements of multimodal data processing to detect cyberbullying in online conver- sations [16]. The same sentiment determination approach is dis- cussed in this study to analyze cyberbullying behavior on TikTok comments.

B. Cyberbullying in Social Media

The use of digital technology for intimidation or tyranny is known as cyberbullying. Social media, chat platforms, entertainment platforms, and mobile devices are all sus- ceptible to this. Cyberbullying is the repeated or irregular use of electronic media for hostile or intentional behavior against someone unable to defend themselves. Therefore, there is a disparity in authority between the offender and the victim. In this instance, the perception of one's phys- ical and cerebral capabilities is the difference in strength [5]. However, it is necessary to consider the characteristics to detect this behavior because cyberbullying cannot be detected by considering only one characteristic.Detecting cyberbullying is different from simply de- tecting aggressive content. Detecting offensive language, insulting words, or hate speech is not a valid reason to indicate that this is a cyberbullying behavior. It is necessary to analyze the words that are said to hurt someone, repeatedly and from time to time, and whether there are situations where there is an imbalance of power in situations where the victims cannot defend themself.

Systems that can adapt to language changes must be developed to enable the accurate detection of cyberbully- ing. Therefore, a new hypothesis was made that in cases of cyberbullying, people tend to use highly harsh words rather than idiots, etc. Next, the writer deletes some unwanted words from the data and converts the data in the form of informal words into formal words.The results of [17] reveal that profanity is not a key- word for detecting cyberbullying, so the authors manually compiled a list of the frequency of themes or categories of cyberbullying. Experimental studies such as machine learning and deep learning models are used to detect this behavior. Another study uses basic qualitative research methods, an open-ended approach that does not involve hypotheses to give in-depth insight into problems [18].

C. Cyberbullying Detection Methods.

One of the cyberbullying detection methods is ma- chine learning which helps with language detection and cyberbullying research analysis, identifies the issues, and suggests new lines of inquiry [19]. Various studies fre- quently use machine learning to detect cyberbullying behavior with algorithms such as Support Vector Ma- chine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF) method [7], [8], [20]–[22] which is a statistical measure that can evaluate how relevant a word is to a document in a collection of documents, a lexicon-based technique such as Sentiment Lexicon for Standard Arabic (SLSA) [23], Word2Vec technique [6], to Bag of Words (BoW) [22]; Naive Bayes (NB) [7], [22], [24]; Decision Tree [22]; and Random Forest (RF) [22], [24], [25]. The results show that these algorithms can au- tomatically distinguish texts containing cyberbullying and non-cyberbullying behavior. SVM is the best-performing classifier [20], providing a good outline that shapes the methods for detecting online bullying from a screenshot with design and implementation details. This work will help curb cyberbullying so that the users can stay at bay from victimization [26]. A study [27] compares Single Learner Approach (SML) classifiers such as KNeighbors (KNN), Logistic Regression (LR), and Linear Support Vector (Linear SVC) with Ensemble Machine Learning Approach (EML) classifiers such as Bagging (RF), Boost- ing (AdaBoost), and Voting (Linear SVC, LR, and KNN). The results show that EML with Voting classifiers out- perform other methods in accuracy for detecting offensive language and cyberbullying in Arabic text. Another study[28] compares machine learning to transfer learning meth- ods. The latter model's capacity to derive more meaningful text representations from extensive training data leads to a more effective approach for detecting cyberbullying in social networks, resulting in improved accuracy. Machine learning classifiers are also used in [29]. Six supervised multiclassification models are being trained to make the predictions, such as Random Forest, Decision Tree, SVM, XG Boost, Neural Network, and MLP. Based on the result, the MLP neural network is the best model to train and evaluate further. Another study [30] is clustered abusive text using Multinomial Naïve Bayes, LinearSVC, Logistic Regression, and K-Nearest neighbor to build a classifier from training datasets. The result is machine learning methods are very scalable for the identification of cyberbullying since they can quickly process and analyze massive volumes of data. A study [31] that combined NLP and MLP to classify and detect Urdu hate speech shows that LR and Multinomial models had minimum training time, but the Extra Tree classifier took longer time to build both word and character n-gram models. LR, Multinomial NB, and Extra Tree classifiers showed good performance in detecting offensive tweets from the dataset.

Nevertheless, despite the invaluable contributions of these studies to the detection of such behavior, one study [24]

comparing SVM, NB, RF, and a specialized form of machine learning, deep learning, indicates that deep learn- ing has the best performance with an accuracy of 92.7%, followed by SVM with 89.3%. In contrast, the remaining algorithms have lower performance. By this outcome, the study suggests that the deep learning approach improves over machine learning for detecting cyberbullying behav- ior. This is in accordance with [22], which uses the BERT model for sentiment analysis, resulting in more precise results than other machine learning methods. When used to the Twitter dataset for sentimental analysis, the result of the suggested model produced a higher accuracy of 91.90%, which can be deemed a better result when com- pared to the conventional machine learning models em- ployed on similar datasets. The BERT algorithm method has high sensitivity and is very suitable for sentiment analysis research.

Moreover, cyberbullying can be detected using a deep learning method. Deep learning offers a workable solution to the issue of cyberbullying that social media platforms can employ to stop and lessen the negative impacts of cyberbullying on users [17]. A study that focuses on using a deep learning algorithm, Deep Neural Network (DNN), compares its algorithm models such as Convolutional Neu- ral Networks (CNN), Long Short-Term Memory Networks (LSTM), Bidirectional LSTM (BLSTM), and BLSTM with attention to detect the same behavior. The results show that these models, coupled with transfer learning, beat state-of-the-art results in detecting cyberbullying behavior on Formspring, Twitter, and Wikipedia [32].Another study compares two deep learning algorithms, Bidirectional Encoder Representations from Transformers (BERT) and Recurrent Neural Network (RNN) combined with LSTM, showing that BERT performs higher at 92.82% accuracy. However, it requires more time than the latter, which has the most balanced option between accuracy and complexity [33]. Another comparative study on deep learning compares the hybrid approach of BERT, CNN, GRU, and capsule as a whole model over machine learning models such as SVM, LR, NB, RF, and deep learning models, likely CNN and LSTM as the baselines. As a result, the research model outperforms the baselines with overall accuracy, precision, recall, and F1-measure values of 79.28%, 78.67%, 81.99%, and 80.30%, respec- tively [34]. Increasing the size of the dataset will also help the deep learning models to learn more and predict the outcome more accurately [35]. A study used Unsupervised Cyberbullying Detection (UCD) with Hierarchical Atten- tion Network (HAN) for textual features and a Graph Auto-Encoder (GAE) for the representation of learning network models in social media sessions. The result shows that the existing system considers only the text, user, and network-related features for detecting cyberbullying. However, overall, the BERT deep learning method has advantages over the UCD-HAN-GAE method regarding data requirements and computing time. The proposed work marks linguistic attributes such as idioms, sarcasm, irony, and active or passive voice along with existing system features and supports unsupervised cyberbullying detection in Instagram [21].

Another study uses basic qualitative research methods, an open-ended approach that does not involve hypotheses but gives an in-depth insight into problems [36]. Besides using machine learning and deep learning, cyberbullying can be detected using the content analysis method. Con- tent analysis is a method that is systematic and objective. Content analysis helps examine different kinds of support offered in Facebook groups created to combat cyberbul- lying. The study discovered that the most common help offered in Facebook groups was emotional support and emphasizing the help provided to victims and bullies [37]. Overall, the BERT deep learning method has advantages over content analysis methods in terms of detection ability and sensitivity. Using the content analysis method tends to be easily biased and less sensitive.The results show that these algorithms can automatically distinguish texts containing cyberbullying and non-cyberbullying behavior. SVM is the best-performing classifier [20], providing a good outline that shapes the methods for detecting online bullying from a screenshot with design and implementation details. This work will help curb cyberbullying so that the users can stay at bay from victimization [26]. However, although they help detect this behavior, one study comparing SVM, NB, RF, and a specialized form of machine learning, deep learning, indicates that deep learning has the best performance with an accuracy of 92.7%, followed by SVM with 89.3%. In contrast, the remaining algorithms have lower performance [24]. By this outcome, the study suggests that the deep learning approach improves over machine learning for detecting cyberbullying behavior.

Moreover, cyberbullying can be detected using a deep learning method. Deep learning offers a workable solution to the issue of cyberbullying that social media platforms can employ to stop and lessen the negative impacts of cyberbullying on users [17]. A study that focuses on using a deep learning algorithm, Deep Neural Network (DNN), compares its algorithm models such as Convolutional Neu- ral Networks (CNN), Long Short-Term Memory Networks (LSTM), Bidirectional LSTM (BLSTM), and BLSTM with attention to detect the same behavior. The results show that these models, coupled with transfer learning, beat state-of-the-art results in detecting cyberbullying behavior on Formspring, Twitter, and Wikipedia [32].Another study compares two deep learning algorithms, Bidirectional Encoder Representations from Transformers (BERT) and Recurrent Neural Network (RNN) combined with LSTM, showing that BERT performs higher at 92.82% accuracy. However, it requires more time than the latter, which has the most balanced option between accuracy and complexity [33]. Another comparative study on deep learning compares the hybrid approach of BERT, CNN, GRU, and capsule as a whole model over machine learning models such as SVM, LR, NB, RF, and deep learning models, likely CNN and LSTM as the baselines. As a result, the research model outperforms the baselines with overall accuracy, precision, recall, and F1-measure values of 79.28%, 78.67%, 81.99%, and 80.30%, respec- tively [34]. Increasing the size of the dataset will also help the deep learning models to learn more and predict the outcome more accurately [35]. A study used Unsupervised Cyberbullying Detection (UCD) with Hierarchical Atten- tion Network (HAN) for textual features and a Graph Auto-Encoder (GAE) for the representation of learning network models in social media sessions. The result shows that the existing system considers only the text, user, and network-related features for detecting cyberbullying. The proposed work marks linguistic attributes such as idioms, sarcasm, irony, and active or passive voice along with existing system features and supports unsupervised cyberbullying detection in Instagram [21].

Another study uses basic qualitative research methods, an open-ended approach that does not involve hypotheses but

gives an in-depth insight into problems [36]. Besides using machine learning and deep learning, cyberbullying can be detected using the content analysis method. Con- tent analysis is a method that is systematic and objective. Content analysis helps examine different kinds of support offered in Facebook groups created to combat cyberbul- lying. The study discovered that the most common help offered in Facebook groups was emotional support and emphasizing the help offered to victims and bullies [37].

D. Cyberbullying Detection using BERT Approach

Aside from the previous comparative study, a study focusing mainly on the deep learning model BERT states that the proposed model achieved significant improve- ment compared to the slot-gated or attention-based DNN models [38]. Another study on the same model proposed that the BERT-based pipeline can be used effectively for sentiment analysis. This pipeline achieved an accuracy of 88.56%, gaining higher accuracy than the previous studies [39]. TikTok cyberbullying detection using BERT has been conducted. Sankar et al. [40] utilized BERT to detect cyberbullying in TikTok user comments by extracting relevant features and performing classification using a deep neural network.Andini et al. [41] employed BERT and Convolutional Neural Networks (CNNs) to detect cyberbullying on Tik- Tok by extracting audio and textual features from videos. Shahrukh et al. [42] proposed a BERT-based sentiment analysis approach for cyberbullying detection on TikTok. This research successfully conveyed feature extraction based on the emotions and sentiments conveyed in user comments. Jahan et al. [43] conducted a comparative study of deep learning techniques for cyberbullying detec- tion on TikTok, including BERT, and found that BERT outperformed other models regarding the accuracy and F1 score.

Ahmed et al. [44] combined BERT with multimodal textual, audio, and visual features to detect cyberbullying on TikTok. Wang and Wang [45] utilized a sentiment analysis approach based on BERT to detect cyberbullying on TikTok by analyzing user comments' sentiments and emotions. Liu et al. [46] proposed a machine learning approach that uses BERT to detect cyberbullying on TikTok by analyzing the user comments' textual features. Chen et al. [47] utilized BERT for sentiment analysis and cyberbullying detection on TikTok by analyzing user comments' emotions and sentiments.Zhang et al. [48] utilized sentiment analysis to detect cyberbullying on TikTok by analyzing the emotions con- veyed in user comments using BERT. Finally, Liu et al. [49] proposed a deep learning approach that utilizes BERT to detect cyberbullying on TikTok by analyzing user com- ments' textual features and achieving higher accuracy than other deep learning models. In conclusion, the studies cited above demonstrate the effectiveness of BERT in cyberbul- lying detection on TikTok. Researchers have explored var- ious approaches to improve detec tion accuracy, including sentiment analysis, multimodal features, and deep learning techniques. These studies provide a solid foundation for future research to develop more sophisticated methods for detecting cyberbullying on TikTok.

III. METHODOLOGY

Cyberbullying has become a primary concern on various social media platforms, including TikTok. This section will delve deeply into the approaches and techniques used in this research to identify and analyze cyberbullying acts through comments on the TikTok application. The methodology of this work is illustrated in Fig. 1.

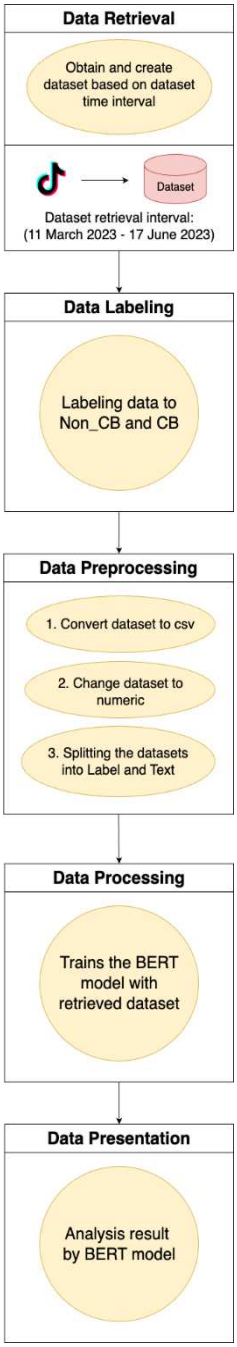

Fig. 1. The methodology of this research.

A. Data Retrieval

The goal of data retrieval is to identify and analyze the presence of cyberbullying on TikTok. This research retrieved TikTok user comment data with 1,508 comments from 11 March - 17 June 2023. The data is shared publicly1 for research purposes in the form of a CSV file. It was taken by considering the relevance of the topic and suitability for the research objectives.

B. Data Labeling

The comments are then classified into CB (Cyberbully- ing) and Non_CB (Non-Cyberbullying). The authors label

each comment by analyzing the texts and the meaning behind the texts. The labeling process uses several criteria to distinguish between the two categories. For instance, identifying words or sentences that contain demeaning language, discomfort, or direct threats directed at indi- viduals or groups. Moreover, cyberbullying often involves humiliation related to a person's physical appearance. Comments that include ridicule, insults, or body-shaming are considered examples of CB sentences.

C. Data Preprocessing

After retrieving data from TikTok user comments, this research preprocessed the data. This step involves cleaning and organizing the data before further analysis. This step includes removing irrelevant characters, combining similar words, and removing unnecessary stop words and punctu- ation. This process is essential to reduce the dimensions of the data and ensure good quality data before analysis. However, it is essential to note that not all comments containing harsh words or negative emotions are auto- matically classified as CB. Non_CB sentences include comments that are not personally offensive but express disapproval or constructive criticism. For example, "pas naik sombongnya minta ampun" or "percaya diri memang penting tapi sadar diri lebih penting".

Furthermore, there are also satirical comments that do not involve direct attacks. Such comments typically fall into the Non_CB category. These comments usually include sentences with subtle satire, prolonged taunts, or the use of humor to convey critical messages. For example, "iya bang keren tapi janji ya ini video terakhir" might express disappointment with someone's presence. Though these comments do not directly insult or demean individuals or groups, they may still contain elements of criticism or ridicule that can sometimes be hurtful to the recipient. The assessment of this type of comment becomes more subjective as it relies on the perception and sensitivity of the individual receiving it.From 1,508 datasets, 676 comments (44.83%) were la- beled as CB (Cyberbullying), and 832 comments (55.17%) were labeled as Non_CB (Non-Cyberbullying).

D. Data Processing

After preprocessing the data, this study trains a deep learning model using the BERT architecture. BERT is a model proven effective in various natural language process- ing tasks, including sentiment analysis.The Cyberbullying BERT model will be trained using the classified data from the previous step. The BERT algo- rithm model used in this study uses several important hy- perparameters in the BERT model for sentiment analysis of TikTok comments, including batch size, learning rate, number of epochs, dropout rate, and maximum length of the processed text. These hyperparameters influence model behavior and performance, such as training speed, adaptability to task data, and overfitting control.

E. Data Presentation

After the Cyberbullying BERT model is trained and tested, the results of the data analysis will be presented in the form of visualizations and reports explaining the research findings. This report will also discuss the impli- cations of these findings and provide recommendations for measures to prevent and intervene in cyberbullying on the TikTok platform. The following Table I is a sample dataset from several Indonesian TikTok comments.

TABLE I. EXAMPLE OF TIKTOK COMMENT

TikTok Comment	Annotation Result
Itam bangett muka lu	CB
Pantes badan nya gede, makan uang orang	CB
Pasangan norakkk	CB
Ku kira muka ternyata ampela😠	CB
Ngeliat mukanya aja udah gedeg	CB
Percaya diri memang penting tapi sadar diri lebih penting 😇	Non_CB
Pas naik sombongnya minta ampun	Non_CB
Pede dulu,glow upnya belakangan	Non_CB
Iya bang keren tapi janji ya ini video terakhir	Non_CB
Awalnya gue suka dia...ehh makin kesini makin kesanah	Non_CB

*CB = Cyberbullying, Non_CB = Non-Cyberbullying

IV. RESULTS AND DISCUSSION

This research trains the BERT algorithm using the preprocessed dataset. The BERT algorithm model used in this study goes through two main stages: pre-training and fine-tuning. At the pre-training stage, BERT is trained on unlabeled text data using "masking" and "next sentence prediction" techniques to study patterns and structures in the text. After that, in the fine-tuning stage, BERT was adapted to the TikTok comment sentiment analysis task using the tagged data.

The training process is carried out using the supervised learning method, dividing the dataset into training data and testing data in a specific ratio. This research tested the trained model on test data to measure the accuracy of cyberbullying detection. The training loss value shows how well the BERT algorithm model predicts the output during the training phase. If the training loss drops from 0.76 to 0.08, BERT's algorithm model is getting better at predicting the correct output on the training data, which is generally reasonable. Validation loss measures how well the model can predict the correct output on validation data, which is not used during training and is usually used to evaluate overfitting. If the validation loss increases from 0.69 to 1.41, the algorithm model is getting worse at predicting the correct output on the validation data. An increase in validation loss usually indicates overfitting, in which the model learns too well from the training data and fails to generalize to data it has never seen before. The training and validation loss is shown in Fig. 2.

The training accuracy value indicates how well BERT's algorithm model predicts the correct class during the training phase. If the training accuracy increases from 0.53 to 0.97, the model is better at predicting the correct class in the training data, which is generally reasonable. Validation accuracy 1.0 means the model can predict the class perfectly on the training data. This measures how well the model can predict the correct class of validation data, which is not used during training and is usually used to evaluate overfitting. If the validation accuracy increases from 0.50 to 0.63, the BERT Algorithm model is better at predicting the correct class in the validation data. This is also usually a good thing, although this increase is not as significant as the increase in training accuracy, which

could indicate that the model may be over fitting the training data. The training and validation accuracy is shown in Fig. 3.

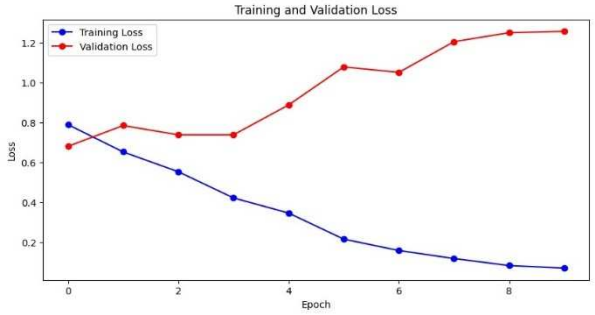

Fig. 2. Training and Validation Loss

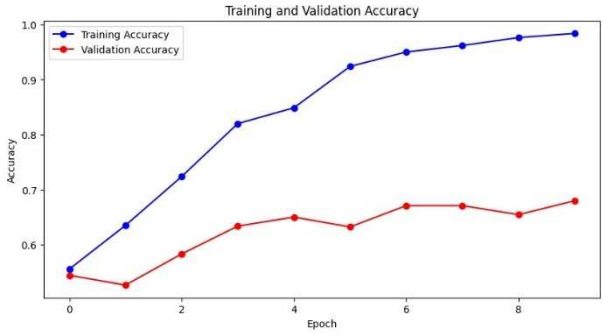

Fig. 3. Training and Validation Accuracy

V. CONCLUSIONS AND FUTURE WORKS

This work builds a prediction model for detecting cyber-bullying using TikTok comments. Comments from TikTok related to cyberbullying were collected independently by the authors. Then, the comments are preprocessed before it was passed for the learning process. This work utilizes BERT architecture for building the prediction model. The experiment results achieved promising performances, resulting 63% validation accuracy in epoch nine.This study shows that the BERT architecture can be utilized to identify cyberbullying on TikTok. The work successfully classified comments into cyberbullying (CB) and non-cyberbullying (Non_CB) categories using a Tik- Tok dataset to train the Cyberbullying BERT model. The results show how the BERT model can effectively identify cases of cyberbullying by detecting patterns and contextual information within comments.Besides the prediction model, looking into efficient in- tervention techniques and assistance programs for TikTok cyberbullying victims is critical. To lessen the harmful effects of cyberbullying, this can entail the development of user-centric features, including reporting systems, privacy restrictions, and algorithmic interventions.The knowledge and prevention of cyberbullying on Tik- Tok and other comparable platforms can be improved by further study in many directions. Future research can concentrate on improving the performance of the BERT model by including a more extensive and varied dataset, further linguistic and contextual characteristics, and op- timal hyperparameters. Multi-modal analysis tools that include text, image, and audio data can also be integrated to provide a more thorough knowledge of cyberbullying behaviors and their context.By addressing these research directions, this paper can advance the understanding of cyberbullying on TikTok and develop more effective prevention, intervention, and support strategies. Ultimately, these efforts will contribute to fostering a safer and more inclusive online environment for all TikTok users.

REFERENCES

[1] S. Kemp, "Digital 2020: Indonesia—datareportal–global digital insights," *datareportal. com*, 2020.

[2] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of traveloka," *Procedia Computer Science*, vol. 216, pp. 682–690, 2023.

[3] A. J. Nair, G. Veena, and A. Vinayak, "Comparative study of twitter sentiment on covid-19 tweets," in *2021 5th InternationalConference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2021, pp. 1773–1778.

[4] S. I. Ali and N. B. Shahbuddin, "The relationship between cyberbullying and mental health among university students," *Sustainability*, vol. 14, no. 11, p. 6881, 2022.

[5] U. Indonesia, "Cyberbullying: What is it and how to stop it," *Unicef*, 2020.

[6] L. R. Halim and A. Suryadibrata, "Cyberbullying sentiment analysis with word2vec and one-against-all support vector ma- chine," *IJNMT (International Journal of New Media Technol- ogy)*, vol. 8, no. 1, pp. 57–64, 2021.

[7] M. Z. Naf'an, A. A. Bimantara, A. Larasati, E. M. Risondang, and N. A. S. Nugraha, "Sentiment analysis of cyberbullying on instagram user comments," *Journal of Data Science and Its Applications*, vol. 2, no. 1, pp. 38–48, 2019.

[8] W. A. Prabowo and F. Azizah, "Sentiment analysis for detecting cyberbullying using tf-idf and svm," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, 2020.

[9] M. R. Kurniawanda and F. A. T. Tobing, "Analysis sentiment cyberbullying in instagram comments with xgboost method," *IJNMT (International Journal of New Media Technology)*,vol. 9, 2022.

[10] D. Farid and N. El-Tazi, "Detection of cyberbullying in tweets in egyptian dialects," *International Journal of Computer Scienceand Information Security (IJCSIS)*, vol. 18, no. 7, pp. 34–41, 2020.

[11] M. Boukabous and M. Azizi, "Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 2, pp. 1131–1139, 2022.

[12] J. Hani, N. Mohamed, M. Ahmed, Z. Emad, E. Amer, and M. Ammar, "Social media cyberbullying detection using ma- chine learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019.

[13] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion classi- fication on indonesian twitter dataset," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 90–95.

[14] U. Khaira, R. Johanda, P. E. P. Utomo, and T. Suratno, "Senti- ment analysis of cyberbullying on twitter using sentistrength," *Indonesian Journal of Artificial Intelligence and Data Mining*,vol. 3, no. 1, pp. 21–27, 2020.

[15] V. R. Sari, N. Hayatin, and Y. Azhar, "Classifying cyberbul- lying data on indonesian social media feeds utilizing sentimentanalysis technique with decision tree model," in *AIP Conference Proceedings*, vol. 2453, no. 1. AIP Publishing LLC, 2022, p. 030011.

[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the twelfth acm international conference on web search and data mining*, 2019, pp. 339–347.

[17] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science*, vol. 181, pp. 605–611, 2021.

[18] D. E.-P. E. Iyobor, R. A. Birikorang, J. Acquah, J. Kankam, and E. A. Arhin, "Automated cyberbullying detection and prevention system," *International Journal of Innovative Science and Research Technology*, vol. 5, 2020.M. Arif, "A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges,"*Journal of Information Security and Cybercrimes Research*, vol. 4, no. 1, pp. 01–26, 2021.

[19] M. Arif, "A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges,"

Journal of Information Security and Cybercrimes Research, vol. 4, no. 1, pp. 01–26, 2021.

[20] X. M. Cuzcano and V. H. Ayma, "A comparison of classification models to detect cyberbullying in the peruvian spanish language on twitter," *International Journal of Advanced Computer Sci-ence and Applications*, vol. 11, no. 10, 2020.

[21] T. Sultan, N. Jahan, R. Basak, M. S. A. Jony, and R. H. Nabil, "Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition," *Int. J. Intell. Syst. Appl*, vol. 15, pp. 1–13, 2023.

[22] S. Mekala, Y. Nithin, B. Raghav, and T. Kumar, "Cyberbullying detection using machine learning," vol. 12, pp. 729–733, 06 2023.

[23] B. Y. AlHarbi, M. S. AlHarbi, N. J. AlZahrani, M. M. Alsheail, J. F. Alshobaili, and D. M. Ibrahim, "Automatic cyber bullying detection in arabic social media," *Int. J. Eng. Res. Technol*, vol. 12, no. 12, pp. 2330–2335, 2019.

[24] L. M. Al-Harigy, H. A. Al-Nuaim, N. Moradpoor, and Z. Tan, "Building towards automated cyberbullying detection: A comparative analysis," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[25] N. Novalita, A. Herdiani, I. Lukmana, and D. Puspandari, "Cyberbullying identification on twitter using random forest classifier," in *Journal of Physics: Conference Series*, vol. 1192, no. 1. IOP Publishing, 2019, p. 012029.

[26] R. Shah, S. Aparajit, R. Chopdekar, and R. Patil, "Machine learning based approach for detection of cyberbullying tweets," *Int. J. Comput. Appl*, vol. 175, no. 37, pp. 51–56, 2020.

[27] M. Khairy, T. Mahmoud, A. Omar, and T. Abd El-Hafeez, "Comparative performance of ensemble machine learning for arabic cyberbullying and offensive language detection," *Language Resources and Evaluation*, 08 2023.

[28] T. Teng and K. Varathan, "Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches," *IEEE Access*, vol. PP, pp. 1–1, 01 2023.

[29] Y. Silva, D. Fernando, L. Rupasinghe, S. Shehan, and R. Hiththatiya, "Machine learning algorithm based automated tool for cyberbullying detection in discord app," *InternationalJournal of Innovative Science and Research Technology*, vol. 8, pp. https://ijisrt.com/machine–learning, 06 2023.

[30] V. Bharadwaj, V. Likhitha, V. Vardhini, A. Asritha, S. Dhyani, and M. Kanth, "Automated cyberbullying activity detection using machine learning algorithm," *E3S Web of Conferences*, vol. 430, 10 2023.

[31] S. Khan and A. Qureshi, "Cyberbullying detection in urdu anguage using machine learning," 12 2022.

[32] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings. Springer, 2018, pp. 141–153.

[33] D. A. Andrade-Segarra, G. A. Le *et al.*, "Deep learning-based natural language processing methods comparison for presump- tive detection of cyberbullying in social networks," *Interna- tional Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[34] K. Maity and S. Saha, "Bert-capsule model for cyberbullying detection in code-mixed indian languages," in Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings. Springer, 2021, pp. 147–155.

[35] M. Vyawahare and S. Govilkar, "Identifying severity of cyberbullying using scalable labeled multi-platform dataset," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 4, pp. 201–210, 2022.

[36] I. Abishak, J. Sheeba, and D. Pradeep, "Unsupervised hybrid approaches for cyberbullying detection in instagram," *International Journal of Computer Applications*, vol. 174, no. 26, pp. 41–46, 2021.

[37] S. Alim and S. Khalid, "Support for cyberbullying victims and actors: A content analysis of facebook groups fighting against cyberbullyingng," *International Journal of Technoethics (IJT)*, vol. 10, no. 2, pp. 35–56, 2019

[38] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, 2022.

[39] M. Pota, M. Ventura, R. Catelli, and M. Esposito, "An effective bert-based pipeline for twitter sentiment analysis: A case studyin italian," *Sensors*, vol. 21, no. 1, p. 133, 2020.

[40] A. Sankar, R. Kaushal, and S. Sood, "Bert for cyberbullying detection on tiktok," in *IEEE International Conference on Signal Processing and Communication*, 2021, pp. 269–274.

[41] R. K. Andini, V. Christy, and I. Lombok, "Detecting cyberbullying on tiktok using bert and convolutional neural networks," in *IEEE International Conference on Smart Computing and Communication*, 2021, pp. 280–284.

[42] M. Shahrukh, M. Z. A. Bhuiyan, and M. R. Islam, "Bert-based sentiment analysis for cyberbullying detection on tiktok," in *International Conference on Computer and Information Science*, 2021, pp. 354–359.

[43] N. Jahan, M. I. Tanveer, and T. M. Rahaman, "A comparative study of deep learning techniques for cyberbullying detection on tiktok using bert," in *International Conference on Information Technology and Innovation*, 2020, pp. 1–6.

[44] T. Ahmed, M. Rafi, and J. Ferdous, "Cyberbullying detection on tiktok using bert and multimodal features," in *International Conference on Machine Learning and Data Engineering*, 2021, pp. 151–156.

[45] Z. Wang and K. Wang, "Cyberbullying detection on tiktok: A sentiment analysis approach," in *IEEE International Conference on Big Data and Smart Computing*, 2020, pp. 271–275.

[46] X. Liu, S. Liu, and X. Sun, "Detection of cyberbullying on tiktok: A machine learning approach," in *IEEE International Conference on Cyber Security and Cloud Computing*, 2020, pp. 89–94.

[47] C. Chen, Y. Liu, and X. Liu, "Sentiment analysis and cyberbul- lying detection on tiktok," in *IEEE International Conference on Computational Science and Engineering*, 2020, pp. 123–128.

[48] Y. Zhang, C. Wu, and H. Zhang, "Using sentiment analysis to detect cyberbullying on tiktok," in *IEEE International Conference on Information and Knowledge Management*, 2020, pp. 2556–2559.

[49] Z. Liu, H. Li, and J. Zhang, "A deep learning approach to cyberbullying detection on tiktok," in *IEEE International Conference on Intelligent Computing and Signal Processing*, 2021, pp. 306–311