Scientific and Practical Cyber Security Journal (SPCSJ) 8(2): 8 – 18 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)

BERT-BASED DETECTION OF CYBERBULLYING IN ONLINE TEXTS

Amrutha Muralidhar
B. M. S College of Engineering, Bangalore, India

ABSTRACT: Social media has experienced exponential growth in recent years, becoming integral to daily communication and interaction. However, along with this growth, cyberbullying has emerged as a significant issue, causing harm and distress to individuals online. This paper investigates the effectiveness of utilizing BERT-based models for identifying cyberbullying behavior in online text. A BERT classifier was trained on a labeled dataset containing instances of cyberbullying and assessed for its performance in accurately detecting such behavior. Results indicate that the BERT classifier achieves a strong accuracy rate of 94% on the test dataset. These findings suggest the potential of BERT-based models in bolstering online safety efforts and combating cyberbullying. The aim of this study is to contribute to the advancement of tools aimed at fostering digital well-being and cultivating safer online communities.

KEYWORDS*: Cyberbullying, Online Safety, Sentiment Analysis, Deep Learning, Text Classification*

1. INTRODUCTION

The social media landscape has undergone a seismic shift in recent years, witnessing unprecedented growth and becoming an indispensable aspect of daily life for millions worldwide. With over 62% of the world's population actively engaged in social media platforms (Petrosyan, 2024), these platforms have become integral to modern communication and social interaction.

Among its diverse user base, a significant contingent comprises children and adolescents, who leverage social media for communication, entertainment, and social interaction. Platforms like Instagram and Snapchat are particularly popular among young adults under 30 (Auxier & Anderson, 2021), revolutionizing interpersonal communication and providing novel avenues for connectivity and community engagement.

While the proliferation of social media offers myriad benefits in facilitating connections and fostering community, it also brings to light critical challenges, the most important of which is the occurrence of cyberbullying. Defined as the use of electronic communication to harass, intimidate, or demean others, cyberbullying poses a significant threat to the psychological well-being of young individuals navigating the digital landscape. Studies have underscored the alarming correlation between extensive social media use and increased vulnerability to cyberbullying victimization, emphasizing the urgent need for proactive intervention strategies (Craig et al., 2020; Horner et al., 2015). Research indicates that cyberbullying victimization rates vary widely, ranging from 5.3% to 66.2% for perpetration and 1.9% to 84.0% for victimization (Camerini et al., 2020).

On April 15th, 2020, United Nations Children's Fund (2020) issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyberbullying are outright alarming: 36.5% of middle and high school students have felt cyberbullied and 87% have observed cyberbullying, with effects ranging from decreased academic performance to depression to suicidal thoughts.

In addition to cyberbullying, the spread of hate speech on social media platforms poses a serious threat, increasing the potential harm that can be done to vulnerable individuals. Though most platforms promote transparency and freedom of speech, the unrestrained spread of hate speech can worsen mental health conditions and prolong social divisions by creating a toxic online atmosphere (Calpbinici et al.,

2019). Ganson et al. (2024) analyzed data from 12,031 adolescents across six countries, finding that increased weekday screen time and use of various social media platforms were associated with higher prevalence of weight-related bullying victimization. Each additional hour of social media was linked to a 13% increase in bullying, with Twitter use showing a 69% increase. Their findings show the significance of addressing social media bullying among adolescents.

Various forms of online conduct, such as insults, threats, harassment, exclusion, and mockery, are indicative of the prevalence of cyberbullying. These forms of cyberbullying can manifest through text-based communication on social media platforms, messaging apps, and online forums. For instance, individuals may resort to insults and name-calling, such as calling someone derogatory names or belittling their character. Threats can take the form of intimidating messages, where individuals express intent to cause harm or distress. Harassment involves persistent and offensive communication aimed at causing emotional harm. Exclusion occurs when individuals purposefully exclude others from online interactions or communities, leading to feelings of isolation. Mockery involves ridiculing or mocking someone's appearance, intelligence, or behavior. Figure 1 illustrates the various forms of cyberbullying, which can have detrimental effects on the psychological well-being of individuals.

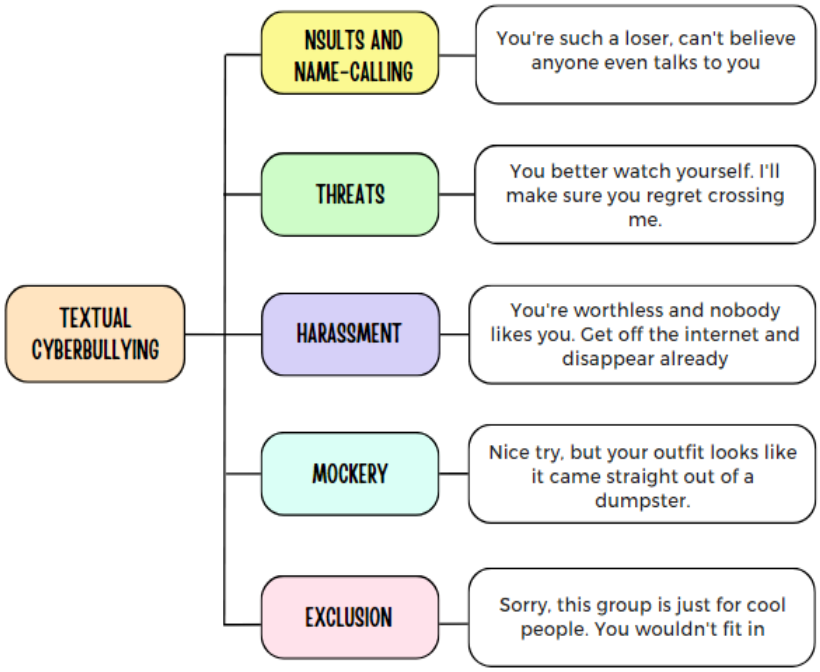

Fig 1. Forms of Cyberbullying

Unlike traditional bullying, which predominantly unfolds in physical spaces like schools, cyberbullying breaches geographical and temporal constraints, reaching victims at any time through their digital devices (Horner et al., 2015). The detrimental effects of cyberbullying on the emotional and psychological well-being of young individuals are well-documented, leading to anxiety, depression, low self-esteem, academic underperformance, and even suicidal ideation. To address this issue, we propose automating text analysis for cyberbullying detection on online platforms. Our research aims to answer the following questions:

R1: How to effectively identify patterns indicative of cyberbullying behavior?
R2: How accurate can we be in distinguishing between normal online interactions and cyberbullying instances?

In response to the escalating prevalence and impact of cyberbullying, we propose automating text analysis for cyberbullying detection on online platforms using a BERT based model. Through this research, we aim to cultivate safer online environments for youth and mitigate the adverse impacts of

digital harassment. By addressing these research questions, we strive to advance our understanding of cyberbullying detection and contribute to the development of effective preventive measures and interventions.

To address these questions comprehensively, this paper is structured as follows: Section 2 provides a review of existing literature and related work. Section 3 delineates the methodology employed in this study, encompassing the utilization of advanced classifiers such as LSTM with attention, Naive Bayes, and BERT. Section 4 offers a detailed exposition of our experimental results, evaluating the performance of each classifier against established benchmarks. Subsequently, in Section 5, we engage in a nuanced discussion of our findings, exploring their implications and potential avenues for future research. Finally, Section 6 encapsulates our conclusions, summarizing key insights and delineating the significance of our contributions to the broader discourse on cyberbullying prevention and mitigation strategies.

2. RELATED WORK

The proliferation of social networks and microblogging platforms has significantly increased instances of "cyber" conflicts and hate speech, posing considerable challenges for online moderation. Despite regulations prohibiting hate speech on most online platforms, the sheer volume of content makes manual moderation impractical, necessitating automated detection and filtering mechanisms. However, existing approaches to detecting cyberbullying and hate speech have shown varying levels of effectiveness, leaving room for improvement in terms of accuracy and efficiency.

One approach to addressing this gap is the Lexical Syntactic Feature (LSF) architecture proposed by Chen et al. (2012), which aims to identify potentially offensive individuals and content on social media. Their framework incorporates pejoratives, profanities, and syntactic rules to predict users' potential to send out offensive content based on writing style and specific cyberbullying content. While this approach offers insight into individual behaviors, it may not capture the full spectrum of offensive language and context present in online texts.

Özel et al. (2017) conducted a study aimed at detecting cyberbullying in Turkish social media messages. They employed information gain and chi-square feature selection methods to enhance classifier accuracy. Their findings indicated that considering both words and emoticons as features improved cyberbully detection accuracy, with Naïve Bayes Multinomial exhibiting the highest classification accuracy among the classifiers tested. Feature selection further improved classification accuracy up to 84%.

Martins et al. (2018) found that incorporating emotional information from text significantly improved the accuracy of hate speech detection. Their research demonstrated a precision rate increase from 41% in previous studies to 80.64% in their tests. However, their study did not address user characterization or the potential use of coding to circumvent anti-hate speech policies and detection systems. Watanabe et al. (2018) proposed an approach to detect hate expressions on Twitter, using unigrams and patterns collected from a training set. Their method achieved an accuracy of 87.4% in binary classification (offensive or not) and 78.4% in ternary classification (hateful, offensive, or clean).

Basak et al. (2019) categorized shaming tweets into six types and developed a classification system to identify shamers and nonshamers. Their findings revealed that most users participating in shaming events were likely to shame the victim, and shamers experienced faster growth in follower counts compared to nonshamers. Rodríguez, Argueta, and Chen (2019) proposed an approach to automatically detect hate speech on Facebook, employing graph analysis, sentiment analysis, and emotion analysis techniques to identify pages promoting hate speech and uncover associated topics.

Yadav et al. (2020) utilized contextual embeddings to generate task-specific embeddings for classification. They trained and evaluated their BERT model on two social media datasets. Roy et al. (2020) utilized tweet text with GloVe embedding vectors to capture semantic information, achieving precision, recall, and F1-score values of 0.97, 0.88, and 0.92, respectively.

Zhou et al. (2020) presented a study exploring fusion techniques of ELMo, BERT, and CNN text classification methods to enhance hate speech detection performance. Their results demonstrated improved classification accuracy.

Alam, Bhowmik, and Prosun (2021) developed ensemble-based model for classifying content into offensive and non-offensive categories. Akter et al. (2022) introduced machine-translated data to address data unavailability issues and evaluated various deep learning models' performance, like LSTM, BiLSTM, LSTM-Autoencoder, word2vec, BERT, and GPT-2. Their study showcased the effectiveness of the BERT model on both semi-noisy and fully machine-translated datasets.

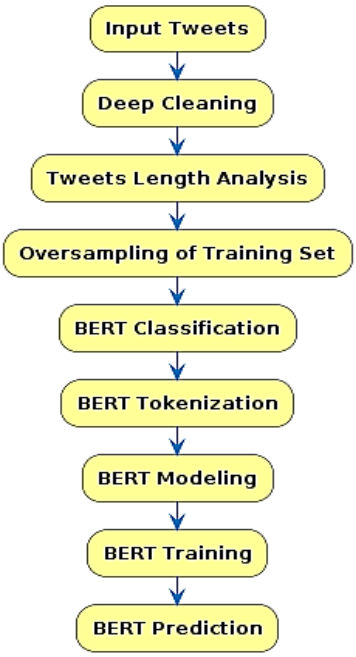

Fig.2 *Methodology Flowchart*

3. METHODOLOGY

This section outlines our methodology its flowchart is depicted in ***Fig.2***.

3.1. Data Preprocessing and Cleaning

The process is crucial for removing noise and irrelevant information, thus enhancing the effectiveness of subsequent classification algorithms, the flowchart is shown in ***Fig.3***. The following functions were applied for data preprocessing:

1. **Emoji Stripping:** Emojis were removed from the text using regular expressions to eliminate non-textual elements that may not contribute to the classification process.
2. **Contractions Expansion:** Contractions were expanded to their full forms to standardize the text and improve consistency in language usage.
3. **Language Filtering:** A language detection algorithm was applied to filter out non-English tweets, as the analysis focuses on English language text. This step ensures that only relevant data is considered for cyberbullying detection.
4. **Entity Stripping:** Various entities such as URLs, mentions, and non-ASCII characters were removed from the text to eliminate noise and irrelevant information.
5. **Hashtag Cleaning:** Hashtags were processed to remove redundant '#' symbols and hashtags occurring at the end of sentences, while retaining those occurring within the text.
6. **Character Filtering:** Special characters such as '$' and '&' present within words were filtered out to ensure uniformity in text representation.
7. **Whitespace Removal:** Extra whitespaces were removed from the text to improve readability and consistency.
8. **URL Shortener Removal:** Shortened URLs commonly used in tweets were removed to prevent misleading information.
9. **Numeric Removal:** Numeric characters were removed from the text to focus solely on textual content.
10. **Word Lemmatization:** Words were lemmatized to reduce inflectional forms and variants to their base or dictionary form, aiding in feature reduction and normalization.
11. **Short Word Removal:** Short words were filtered out from the text to eliminate noise and improve the relevance of the remaining words for classification.

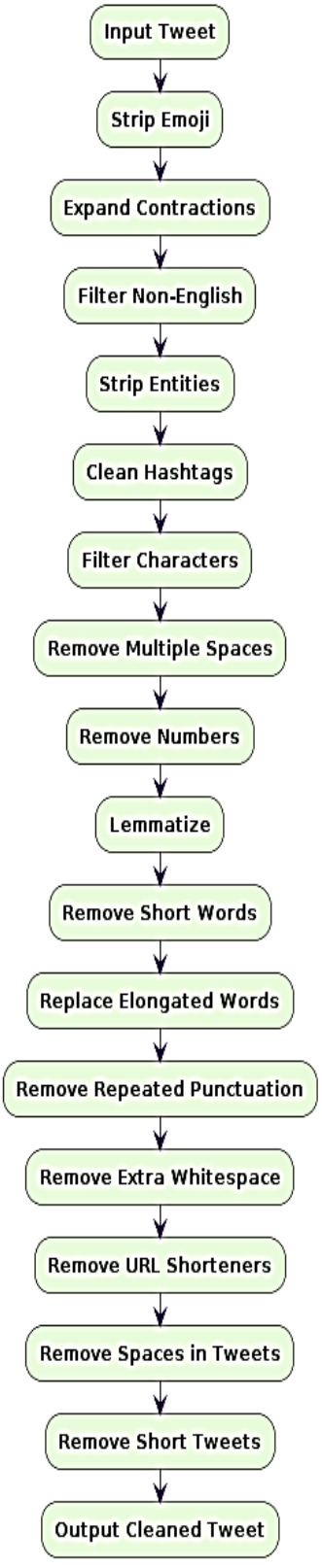

Fig.3 *Deep Cleaning Process*

12. Elongated Word Replacement: Elongated words, characterized by repeated letters, were replaced with their base form to standardize the text representation.

13. Repeated Punctuation Removal: Redundant punctuation marks were removed to enhance readability and simplify text representation.

14. Tweet Length Filtering: Short tweets containing fewer than a predefined number of words were removed to ensure an adequate amount of textual content for analysis.

This cleaning process can be summarized as follows:

$$\text{Cleaned Tweet} = f_{14}\left(f_{13}\left(\ldots f_2\left(f_1(\text{Raw Tweet})\right)\ldots\right)\right)$$

where □□ represents each cleaning function applied sequentially to the raw tweet.

3.2. Tweet Length Analysis and Oversampling

The examination of the class distribution showed that there was an imbalance among the classes as shown in ***Fig.4***, meaning that some have fewer occurrences than others. Oversampling approaches are used to rectify this imbalance and avoid bias towards the dominant class during model training. Oversampling involves randomly duplicating instances from the minority classes or generating synthetic instances to balance the class distribution. In this study, we opted to oversample the training set such that all classes have the same count as the most populated one.

The oversampling process can be represented mathematically as follows: Let □ be the desired count of instances for each class, and □ be the number of classes. For each class □, where □ = 1, 2,.., □, we calculate the oversampling factor □□ as:

$$F_i = \frac{N}{n_i}$$

Where □□ represents the current count of instances for class □. Then, for each class □, we randomly select instances from the original dataset to achieve the desired count □ using the calculated oversampling factor □□. This process is repeated until all classes have the same count of instances.

After oversampling, the class distribution is rebalanced as shown in ***Fig.5***, ensuring that each class contributes equally to the training process. This mitigates the risk of model bias towards the majority class and improves the overall performance of the classification model.

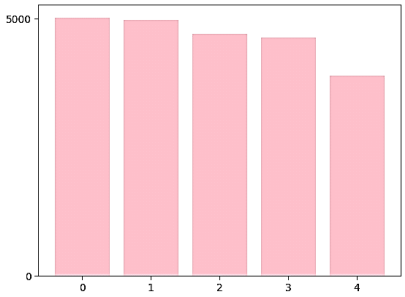

Fig.4 *Class Distribution before oversampling*

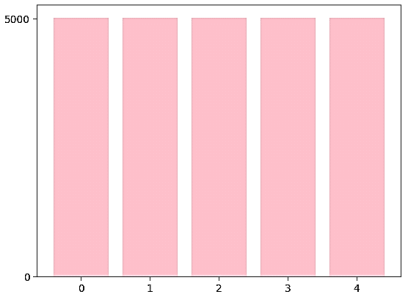

Fig.5 *Class Distribution after oversampling*

3.3. BERT Classification

We loaded a pre-trained BERT model and fine-tuned it on our sentiment classification task. This classification process is shown in ***Fig.6*** To enable reliable model assessment, the dataset was divided into training, validation, and test sets. Gradient descent optimization was then used to optimize the BERT model's parameters. This can be represented as the process of updating the parameters □ of the pre-trained BERT model using gradient descent optimization:

$$\theta_{fine-tuned} = \text{argmin}_{\theta} \sum_{i=1}^{N} \text{L}oss(\text{B}ERT(X_i), y_i)$$

where Loss represents the cross-entropy loss, □□ denotes the input data, □□ represents the corresponding true labels, and □ is the number of samples in the training set. Cross-entropy loss measures the dissimilarity between the predicted probability distribution (output of the model) and the true probability distribution (ground truth labels). This can be expressed as:

$$CrossEntropyLoss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} log(p_{i,c}) (y_{i,c})$$

Where:

- □ is the number of samples in the dataset.
- □ is the number of classes.
- □□,□ is a binary indicator (0 or 1) of whether sample □ belongs to class □.
- □□,□ is the predicted probability that sample □ belongs to class *c* according to the model.

3.4. BERT Tokenization

In the tokenization process shown in ***Fig.7***, the dataset was initially split into training, validation, and test sets to facilitate subsequent model training and evaluation. A custom tokenizer function was utilized to tokenize the raw textual data into sequences of tokens. To ensure focused attention during model training and inference, attention masks were generated to differentiate between actual tokens and padding tokens. To ensure uniformity throughout the dataset, the longest tokenized tweet is used to determine the maximum token length.

3.5. BERT Modelling

In the BERT modeling phase, we construct a custom BERT classifier tailored to our sentiment classification task. This involves designing a model architecture that combines the power of BERT's transformer layers with additional dense layers for classification. The custom BERT classifier comprises a pre-trained BERT model, which serves as the foundation for capturing contextual embeddings and understanding intricate language patterns. These transformer layers are augmented with fully connected (dense) layers followed by ReLU activation functions, enabling the transformation of BERT's contextual embeddings into sentiment predictions. The initialization process

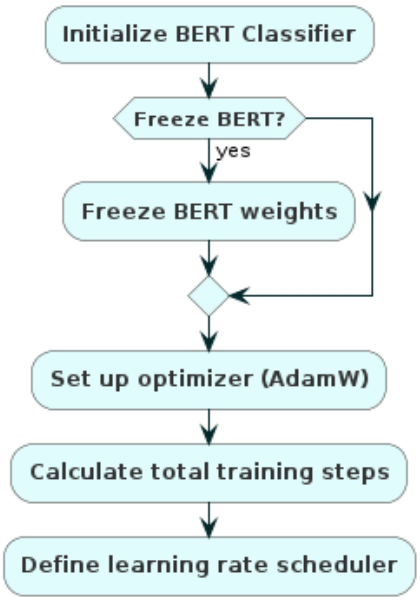

Fig.6 *BERT Classifier*

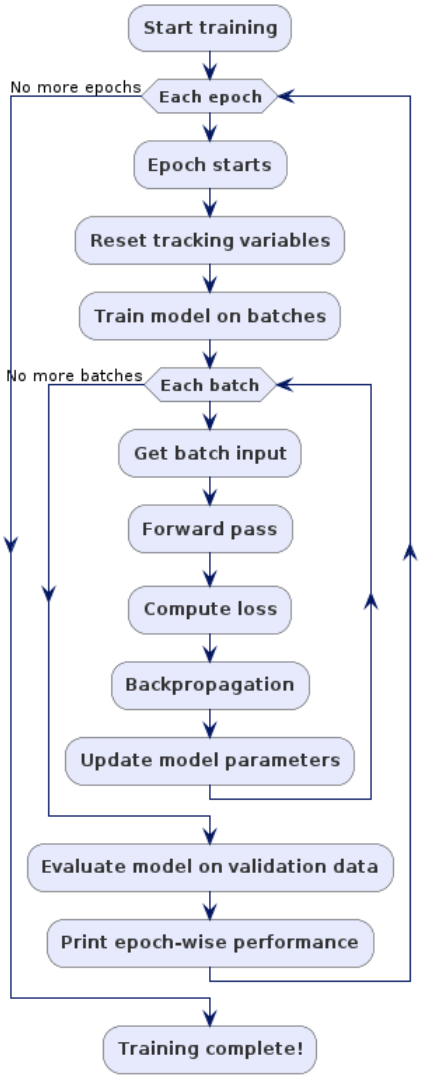

Fig.8 *BERT Training*

involves setting up the model, optimizer, and learning rate scheduler. The optimizer, typically AdamW, is employed to update model parameters during training with a specific learning rate. Additionally, a learning rate scheduler, such as the linear scheduler with warmup, is utilized to modulate the learning rate throughout training epochs, ensuring stability and efficiency. During fine-tuning, the BERT classifier is trained on the tokenized dataset using the AdamW optimizer and the designated learning rate scheduler. This iterative process involves updating model parameters to minimize the chosen loss function, typically cross-entropy loss, computed via backpropagation. GPU acceleration is leveraged to harness computational efficiency, particularly beneficial for complex deep learning architectures like BERT. Subsequently, the BERT model is initialized to configure the model, optimizer, and learning rate scheduler for effective fine-tuning on the sentiment classification task. This comprehensive process ensures the successful integration of BERT for sentiment analysis tasks.

The AdamW optimizer extends the original Adam optimizer by incorporating weight decay regularization, hence the name "Adam with weight decay" (AdamW). The weight decay term acts as a penalty on the magnitudes of model weights during optimization, helping to prevent overfitting by discouraging large parameter values. This rule can be expressed as follows:

$$\theta_{t+1} = \theta_t - \frac{\mathrm{lr} \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - lr \cdot wd \cdot \theta_t$$

Where:

- θ_t represents the model parameters at time step □.
- lr denotes the learning rate.
- $\hat{m}_t$ and $\hat{v}_t$ are the biased first and second moments estimators, respectively.
- □ is a small constant added to prevent division by zero.
- wd is the weight decay coefficient.

3.6. BERT Training

In the BERT training phase, the custom BERT classifier model is fine-tuned on the training dataset to optimize its performance for sentiment classification. The training process involves iterating over multiple epochs, where each epoch consists of several batches of training data. Within each epoch, the model computes the forward pass to generate predictions, calculates the loss function (cross-entropy loss) to measure prediction accuracy, and performs backpropagation to update the model parameters and minimize the loss. The AdamW optimizer is employed to update model parameters, while the learning rate scheduler dynamically adjusts the learning rate throughout training epochs for enhanced model convergence. To prevent the gradients from exploding, gradient clipping is applied. The model's performance is evaluated on the validation dataset after each epoch, computing metrics such as validation loss and accuracy. This iterative process continues until the specified number of epochs is reached, resulting in a fine-tuned BERT classifier model ready for deployment. This process is shown in ***Fig.8***.

3.7. BERT Prediction

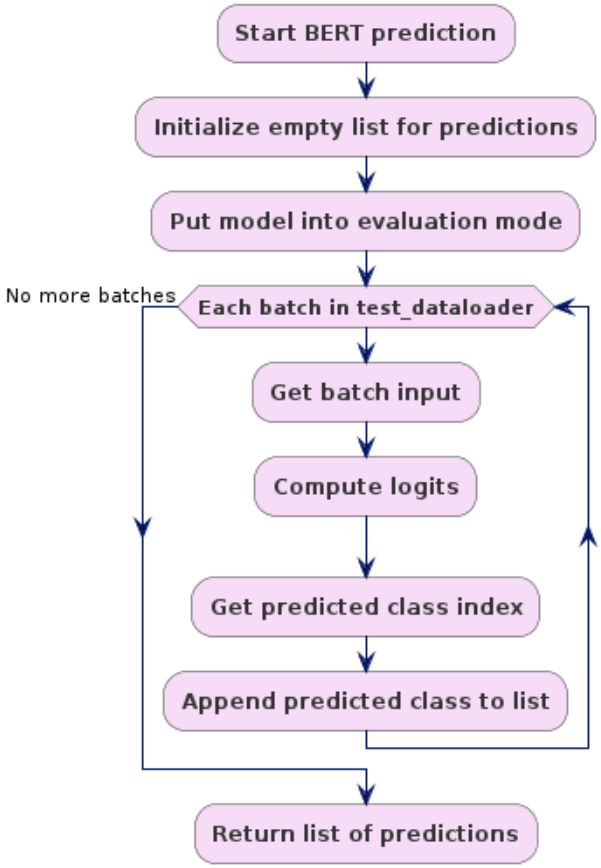

Fig.9 *Prediction*

In the BERT prediction phase, the fine-tuned BERT classifier model is utilized to generate sentiment predictions on the test dataset. A function, similar to the model evaluation

process, is defined to facilitate the prediction task as shown in ***Fig.9***. Within this function, the model is set to evaluation mode, ensuring that no gradient calculations are performed during inference. Subsequently, the test dataset is iterated over in batches. For each batch, the model generates predictions based on the input data, employing a forward pass to compute logits. The predicted class index is obtained by taking the argmax of the logits, and these predictions are aggregated and stored in a list.

Let's denote the test dataset as *test_data*, the BERT classifier model as *model*, the batch size as □, and the number of classes as □. For each batch, the input data □ is passed

through the model, producing logits □ of size □ × □. The predicted class indices $\hat{Y}$ are obtained by taking the argmax along the class dimension:

$$\hat{Y} = argmax\,(Z, \text{dim} = 1)$$

These predicted class indices are then aggregated across batches, resulting in a list of predicted sentiment labels.

Finally, the predicted sentiment labels are compared against the ground truth labels □$_{\text{true}}$. A classification report is generated to assess the model's performance, providing insights into the precision, recall, F1-score, and support for each sentiment class. The F1-score is the harmonic mean of precision and recall, given by:

$$\text{F1 Score} = \frac{2 \times precision \times recall}{precision + recall}$$

Support refers to the number of occurrences of each class in the test dataset, providing an indication of the dataset's class distribution. Through this predictive analysis, the effectiveness of the BERT model in accurately classifying sentiments on unseen data is assessed.

4. RESULTS

The training process exhibited promising outcomes, as demonstrated by the training loss and validation accuracy metrics shown in ***Table.1***. Throughout the training iterations, the average training loss consistently decreased, indicating effective learning from the training data. Simultaneously, the validation loss remained consistently low, suggesting robust generalization to unseen data. The validation accuracy reached an impressive 94.85%, showing our fine-tuned BERT classifier's proficiency in accurately identifying sentiments across various categories.

Table.1. *Training Results*

BATCH NO.	TRAIN LOSS	ELAPSED (s)
100	0.660176	39.12
200	0.311569	38.77
300	0.241835	38.71
400	0.257079	38.71
500	0.240932	38.70
600	0.221965	38.72
700	0.178535	38.67
783	0.161037	31.99

AVG TRAIN LOSS	VAL LOSS	VAL ACCURACY (%)	ELAPSED (s)
0.287290	0.164714	94.85	325.67

Upon assessing the BERT classifier's performance on the test dataset, the classification report in ***Table.2*** showed notable precision, recall, and F1-score metrics across all sentiment classes. Precision scores ranged from 84% to 99%, indicating the model's adeptness in minimizing false positive predictions. Similarly, recall scores spanned from 90% to 98%, showing the model's capability to capture most positive instances for each sentiment class. And the F1-scores, representing the harmonic mean of precision and recall, exceeded 0.90 for all sentiment classes, reaffirming the model's balanced performance across these metrics.

Table.2. *Classification Report of Prediction*

Class	Precision	Recall	F1-score	Support
religion	0.95	0.98	0.96	1568
age	0.99	0.98	0.98	1552
ethnicity	0.99	0.99	0.99	1469
gender	0.92	0.90	0.91	1446
not bullying	0.84	0.84	0.84	1214
Accuracy	0.94			
Macro avg	0.94			
Weighted avg	0.94			

Thus, this BERT classifier demonstrates good accuracy and reliability in identifying cyberbullying instances, achieving an overall accuracy of 94% on the test dataset. These results show the effectiveness of utilizing BERT-based models for cyberbullying detection tasks.

5. CONCLUSION

In this research the effectiveness of utilizing BERT-based models for cyberbullying detection tasks was investigated. Through fine-tuning a pre-trained BERT classifier on a labeled dataset containing instances of cyberbullying, the model's ability to accurately identify and classify cyberbullying behavior in textual data was demonstrated. The results show the robust performance of the BERT classifier, as evidenced by high precision, recall, and F1-score metrics across various sentiment classes. The model's capacity to capture contextual information and semantic nuances within text enabled it to effectively discern instances of cyberbullying from non-cyberbullying content.

Nevertheless, the findings from this study underscore the potential of BERT-based models as valuable tools for enhancing online safety and combating cyberbullying. By leveraging advanced natural language processing techniques, we can better understand and address instances of harmful behavior in digital environments, ultimately contributing to the creation of safer and more inclusive online communities. Moving forward, further research and development efforts should focus on refining and optimizing cyberbullying detection systems, exploring multi-modal approaches, and addressing ethical considerations to ensure responsible and effective deployment of these technologies in fostering a comprehensive and sustainable approach to combating cyberbullying and promoting digital well-being for all.

REFERENCE

1. Ani Petrosyan. 2024. “Worldwide Digital Population 2024.” Statista. May 7, 2024. Accessed May 8, 2024. https://www.statista.com/statistics/617136/digital-population-worldwide/.
2. Auxier, Brooke, and Monica Anderson. "Social media use in 2021." *Pew Research Center 1*, no. 1 (2021): 1-4.
3. Craig, Wendy, Meyran Boniel-Nissim, Nathan King, Sophie D. Walsh, Maartje Boer, Peter D. Donnelly, Yossi Harel-Fisch et al. "Social media use and cyber-bullying: A cross-national analysis of young people in 42 countries." *Journal of Adolescent Health* 66, no. 6 (2020): S100-S108. https://doi.org/10.1016/j.jadohealth.2020.03.006
4. Horner, Stacy, Yvonne Asher, and Gary D. Fireman. "The impact and response to electronic bullying and traditional bullying among adolescents." *Computers in human behavior* 49 (2015): 288-295. https://doi.org/10.1016/j.chb.2015.03.007
5. Camerini, Anne-Linda, Laura Marciano, Anna Carrara, and Peter Schulz. ‘Cyberbullying Perpetration and Victimization among Children and Adolescents: A Systematic Review of Longitudinal Studies’. *Telematics and Informatics* 49 (06 2020): 101362. https://doi.org/10.1016/j.tele.2020.101362.
6. Calpbinici, Pelin, and Fatma Tas Arslan. "Virtual behaviors affecting adolescent mental health: The usage of Internet and mobile phone and cyberbullying." *Journal of Child and Adolescent Psychiatric Nursing* 32, no. 3 (2019): 139-148.
7. United Nations Children’s Fund (UNICEF). 2020. “Children at Increased Risk of Harm Online During Global COVID-19 Pandemic.” Unicef.Org. April 14, 2020. Accessed May 8, 2024. https://www.unicef.org/press-releases/children-increased-risk-harm-online-during-global-covid-19-pandemic.
8. Ganson, Kyle T., Nelson Pang, Jason M. Nagata, Catrin Pedder Penn-Jones, Faye Mishna, Alexander Testa, Dylan B. Jackson, and David Hammond. 2024. “Screen Time, Social Media Use, and Weight-related Bullying Victimization: Findings From an International Sample of Adolescents.” *PloS One* 19 (4): e0299830. https://doi.org/10.1371/journal.pone.0299830.
9. Chavan, V. S., & Shylaja, S. S. "Machine learning approach for detection of cyber-aggressive comments by peers on social media network." In 2015 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2354-2358. Kochi, India, 2015. DOI: 10.1109/ICACCI.2015.7275970.
10. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 *International Conference on Social Computing*, pp. 71-80. Amsterdam, Netherlands, 2012. DOI: 10.1109/SocialCom-PASSAT.2012.55.
11. Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. "Detection of cyberbullying on social media messages in Turkish." In 2017 *International Conference on Computer Science and Engineering (UBMK)*, pp. 366-370. Antalya, Turkey, 2017. DOI: 10.1109/UBMK.2017.8093411.
12. Yadav, J., Kumar, D., & Chauhan, D. "Cyberbullying Detection using Pre-Trained BERT Model." In 2020 *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1096-1100. Coimbatore, India, 2020. DOI: 10.1109/ICESC48915.2020.9155700.
13. Basak, R., Sural, S., Ganguly, N., & Ghosh, S. K. "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation." In *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 208-220, April 2019. DOI: 10.1109/TCSS.2019.2895734.
14. Watanabe, H., Bouazizi, M., & Ohtsuki, T. "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection." In *IEEE Access*, vol. 6, pp. 13825-13835, 2018. DOI: 10.1109/ACCESS.2018.2806394.
15. Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X.-Z. "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network." In *IEEE Access*, vol. 8, pp. 204951-204962, 2020. DOI: 10.1109/ACCESS.2020.3037073.
16. Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. "Hate Speech Classification in Social Media Using Emotional Analysis." In 2018 *7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 61-66. Sao Paulo, Brazil, 2018. DOI: 10.1109/BRACIS.2018.00019.

17. Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach. In 2021 *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 710-715). Tirunelveli, India. DOI: 10.1109/ICICV50876.2021.9388499.
18. Rodríguez, A., Argueta, C., & Chen, Y.-L. (2019). Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In 2019 *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 169-174). Okinawa, Japan. DOI: 10.1109/ICAIIC.2019.8669073.
19. Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep Learning Based Fusion Approach for Hate Speech Detection. *IEEE Access*, 8, 128923-128929. DOI: 10.1109/ACCESS.2020.3009244.
20. Akter, M. S., Shahriar, H., Ahmed, N., & Cuzzocrea, A. (2022). Deep Learning Approach for Classifying Aggressive Comments on Social Media: Machine Translated Data Vs Real Life Data. 2022 *IEEE International Conference on Big Data (Big Data)*, 5646-5655. doi: 10.1109/BigData55660.2022.10020249.