Analysing and Detecting Cyberbullying in Social Platforms

Zara Ashreen[1], Mohammed Waheeduddin Hussain[2], Sridhar Gummalla[3]
[1]*Research Scholar, Department of Computer Science and Engineering, SCET, Hyderabad, Telangana*
[2]*Professor, Department of Information Technology, SCET, Hyderabad, Telangana*
[3]*Professor, Department of Computer Science and Engineering, SCET, Hyderabad, Telangana*

Abstract- Social networking and communication have been accelerated by information and communication technologies, yet cyberbullying presents serious problems. The cumbersome and ineffective processes currently in place for reporting and prohibiting cyberbullying are relied on the user. For automated cyberbullying identification, traditional machine learning and transfer learning techniques were investigated. An organized annotation procedure and an extensive dataset were employed in the study. The Conventional Machine Learning technique used term lists, psycholinguistics, textual, sentiment and emotive, static and contextual word embeddings, and toxicity characteristics. The use of toxicity features for cyberbullying identification was first demonstrated by this study. The word Convolutional Neural Network (Word CNN) showed similar performance when its contextual embeddings were selected based on their higher F-measure. When supplied separately, toxicity characteristics, embeddings, and textual features raise the bar. In this case, linear SVC was not as effective of handling high-dimensionality characteristics and training time. By using Word CNN for fine-tuning, Transfer Learning was able to achieve a faster training computation than the base models. Furthermore, the implementation of Flask web for cyberbullying detection produced the maximum accuracy. For reasons of privacy, the reference to the particular dataset name was removed.

INTRODUCTION

With their subtle evolution over time, information and communication technologies (ICT) have become an essential element of everyone's life and have accelerated online contact between individuals. With the increasing usage of internet platforms, communication has become as simple as clicking a button, which has aided in the development of social networking. The prevalence of ICTs has a negative side when people abuse them with ease, such in the case of cyberbullying. Cyberbullying is the extension of traditional or direct bullying onto digital media. In order to safeguard online communities, social media becomes the virtual medium for bullying, hiding the identity of the aggressor and making cyberbullying detection a difficult and demanding task. Because cyberbullying may be readily conducted anonymously, its incidence rise with increased Internet usage. This poses a serious public health risk and has numerous detrimental effects, including as social, psychological, and mental health issues. While despair, anxiety, loneliness, and anhedonia are common mental health issues among cyberbullying victims, some have also been documented to engage in self-harming behaviors and entertain suicidal thoughts.

The anticipated result of this research is the creation of classification models that, by utilizing the state-of-the-art in natural language processing and deep learning, can efficiently distinguish between instances of cyberbullying and non-cyberbullying from disorderly posts. This work combines word CNN model building, feature engineering, and text pre-processing.

OBJECTIVE

Investigate and apply state-of-the-art NLP and Deep Learning techniques to enhance the detection of cyberbullying on digital platforms. Examine the addition of new factors, such toxicity indicators, to standard textual and sentiment data in order to improve the accuracy of cyberbullying detection systems.

With Word CNN, you can create robust classification models that perform better than conventional machine learning methods for contextual embeddings. As a practical application of the developed models, incorporate cyberbullying detection into a Flask web platform to attain high accuracy in real-time identification and prevention.

PROBLEM STATEMENT

The emergence of cyberbullying poses a serious threat to online communities in the Information and Communication Technologies (ICT) era, as online communication has become pervasive. Cyberbullying is the continuation of traditional bullying via electronic means, and it commonly takes the shape of anonymous posts on social media and other websites. This anonymity makes it more difficult to identify and stop cyberbullying, which has detrimental effects on the social, psychological, and mental health of victims.

The goal of this project is to create efficient categorization models by utilizing the most recent Deep Learning and Natural Language Processing (NLP) methods. The objective of these models is to precisely identify instances of cyberbullying in textual data and differentiate them from non-cyberbullying material. Text pre-processing, feature engineering, and model creation using word CNN (Convolutional Neural Networks) are all included in the research. The best possible the development of effective tools and tactics to stop cyberbullying and promote safer online communities is the main objective.

EXISTING SYSTEM

The current method is designed to tackle the resource-intensive nature of training machine learning (ML) classifiers, especially in light of the growing difficulty presented by massive datasets and the widespread use of Deep Neural Networks (DNN). In order to maximize the training process, feature density (FD), a technique for estimating machine learning classifier performance prior to training, is examined.

The study emphasizes how resource-intensive training affects the environment, particularly in light of the growing CO2 emissions linked to large-scale machine learning models. The goal of the project is to improve Natural Language Processing efficiency and reduce the need for heavy computer resources. Particular focus will be placed on dialog classification, which includes cyber bullying detection.

Disadvantage of Existing System

- Narrow Focus: While Feature Density (FD) analysis may provide light on classifier performance estimation, it may also obscure other vital components of machine learning model efficiency and optimization.
- Complexity: Using FD analysis and refining ML classifiers based on this parameter may make the training process more difficult, including more knowledge and processing power.
- Dependency on Data: The quality and features of the dataset can have a significant impact on the efficacy of FD analysis and optimization techniques, which can limit their applicability to different datasets.

PROPOSED SYSTEM

The proposed approach uses Word CNN's cyber bullying detection features. Transfer Learning adapts the WORD CNN to the distinct features of cyber bullying in the dataset. Through this process, training computation is sped up compared to starting from zero, and the model's ability to recognize subtle patterns is enhanced.

Most importantly, with bigger F-measures, the contextual embeddings generated by the WORD CNN function similarly. This approach builds upon the conventional use of embeddings and offers a novel perspective on toxicity features, contributing to a more comprehensive model for cyber bullying identification. The system's utilization of WORD CNN not only increases accuracy but also demonstrates a progressive approach to addressing the evolving challenges related to online social interactions.

Advantages of Proposed System

Transfer Learning: Leveraging Transfer Learning with Word CNN allows the model to benefit from pre-trained knowledge and adapt to the unique characteristics of cyber bullying in the dataset. This approach can lead to faster training times and improved performance compared to training from scratch.

Improved Pattern Recognition: By fine-tuning Word CNN for cyberbullying detection, the model's ability to identify subtle patterns associated with cyberbullying behaviors is enhanced. This can result in higher accuracy and sensitivity in detecting instances of cyber bullying.

Efficient Training: The use of Transfer Learning and contextual embeddings helps accelerate the training process, reducing computational resources and time required for model training, making it more scalable for large datasets.

RELATED WORKS

In the field of cyber bullying detection, numerous study directions have been investigated. Text-based methods are a fundamental component that use sentiment analysis, language patterns, and contextual indicators to identify abusive content. Deep learning methods, including as transformers, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), have drawn interest because of their capacity to identify complex correlations in textual data, improving the identification of cyber bullying. Transfer learning has been a useful approach that lets models make use of pre-trained knowledge and fine-tune to the unique characteristics of cyber bullying. By utilizing the combined strength of several classifiers, ensemble methods increase the precision and resilience of predictions. Furthermore, the shift towards multimodal techniques that incorporate textual, visual, and behavioral characteristics holds the potential to provide a more comprehensive knowledge of the behaviors associated with cyber bullying on various platforms. An examination of social networks methods provide information about the spread and effects of cyber bullying in virtual networks by highlighting prominent nodes and malevolent conduct trends.

METHODLOGY OF PROJECT

Data collection from many web platforms is the first step in the methodology for this project on cyberbullying detection using Word CNN and Transfer Learning. Next, the text data is cleaned and prepared by preprocessing. In order to capture semantic subtleties, a bespoke Word CNN architecture is created that incorporates transfer learning by starting with pre-trained word embeddings. To adjust its capabilities, the model is trained on a split dataset and refined using data particular to cyberbullying. Model performance is measured using evaluation measures like accuracy, precision, recall, and F1-score; cross-validation and hyper parameter adjustment maximize efficacy. After training, the model is implemented and integrated into a real-time detection environment in a production setting, all the while taking ethical issues like privacy, bias, and responsible AI deployment into account. thorough records and to ensure openness and usability in the fight against cyber bullying, reporting should include the project's goals, methods, conclusions, constraints, and suggestions for the future.

MODULE NAMES:

1) Dataset:

In order to identify cyberbullying in the early stages of the project, we acquired a dataset. The compilation, derived from "cyberbullying_tweets.csv," consists of various text items, or tweets. There are 47,692 records in the dataset that have been categorized as "not_cyberbullying," "gender," "religion," "other_cyberbullying," "age," or "ethnicity."

2) Importing the Necessary Libraries:

We imported the required project libraries and made the decision to develop in Python. Important libraries include scikit-learn, which divides the data into training and testing sets, PIL, which creates arrays from pictures, and other standard libraries such as matplotlib, pandas, numpy, and TensorFlow. The main model is built with Keras.

3) Data Pre-processing:

To read the CSV file, we utilized pandas. We also handled any missing values and performed an initial data examination using info(). Next, using Label Encoding, the textual data was modified for the 'cyberbullying_type' labels. We then separated the dataset into training, validation, and testing sets.

4) Model Creation for Word CNN:

We refer to them as Convolutional Neural Networks (CNNs) as their effectiveness in solving document classification issues has been demonstrated. With 128 filters (parallel fields for word processing) and a kernel size of 5, a rectified linear (or "relu") activation function, a conservative word CNN configuration is employed. A pooling layer that lowers the convolutional layer's output comes next.

It is evident that the Embedding layer encodes each word in a document as an 11-element vector and expects documents containing words as input.

Since the problem we are learning is a categorical classification problem, we employ a categorical cross entropy loss function. We employ the effective Adam implementation of stochastic gradient descent, and we monitor both loss and accuracy in the course of training. A total of 35 epochs, or 64 iterations through the training set, are used to train the model. Six nodes are output as "not_cyberbullying," "gender," "religion," "other_cyberbullying," "age," or "ethnicity" by the final thick layer. This layer forecasts which of the six possibilities has the highest likelihood using the softmax activation function, which provides a probability value.

5) Training and Evaluation:
We trained the Word CNN model by adjusting hyperparameters like batch size and epochs using the fit function. In training, an average of 97.06% was reached, however in validation, an average of 99.92% was attained. After that, we evaluated the model with the test set, and the accuracy was 97.7%.

6) Saving the Trained Model:
We then saved the model for use in the final UI detection and testing. The model was stored in the Hierarchical Data Format (HDF5) using the Tensorflow/Keras library. Moreover, the tokenizer used for text preparation was serialized using the pickle library and saved as "tokenizer.pickle".

Benefits:
Ethical Considerations: Including ethical considerations at every stage of the project guarantees the responsible use of AI, addresses privacy issues, reduces bias, and encourages just and equitable detection procedures.
Thorough Documentation: By facilitating information transfer and promoting transparency, documentation and reporting help stakeholders comprehend the approach, findings, constraints, and future suggestions for cyber bullying detection study and application.

Impediments of DL
There may be a number of challenges when implementing Word CNN and Transfer Learning for the detection of cyberbullying. First off, skewed findings could emerge from biased or insufficient training data, which is why the quality and diversity of the datasets are so important to the effectiveness of the model. Furthermore, deep learning model training can involve significant computational demands, particularly when Transfer Learning is used. This means that high-performance computing resources are necessary. Another issue is model interpretability because deep learning models frequently lack transparency, making it challenging to comprehend and evaluate their conclusions. Additionally, there may be limitations to the model's adaptation to various cyber bullying behaviors and circumstances, therefore model robustness and adaptability must be carefully considered. Privacy issues pertaining to the gathering and examination of private text data about cyberbullying, in addition to Implementation becomes even more challenging when ethical issues like algorithmic fairness and regulatory compliance are involved. To ensure the efficient and responsible deployment of cyberbullying detection systems, overcoming these obstacles calls for a comprehensive strategy that takes into account data quality, computational resources, model interpretability, generalization, privacy protection, ethical considerations, and regulatory compliance.

DATA FLOW DIAGRAM

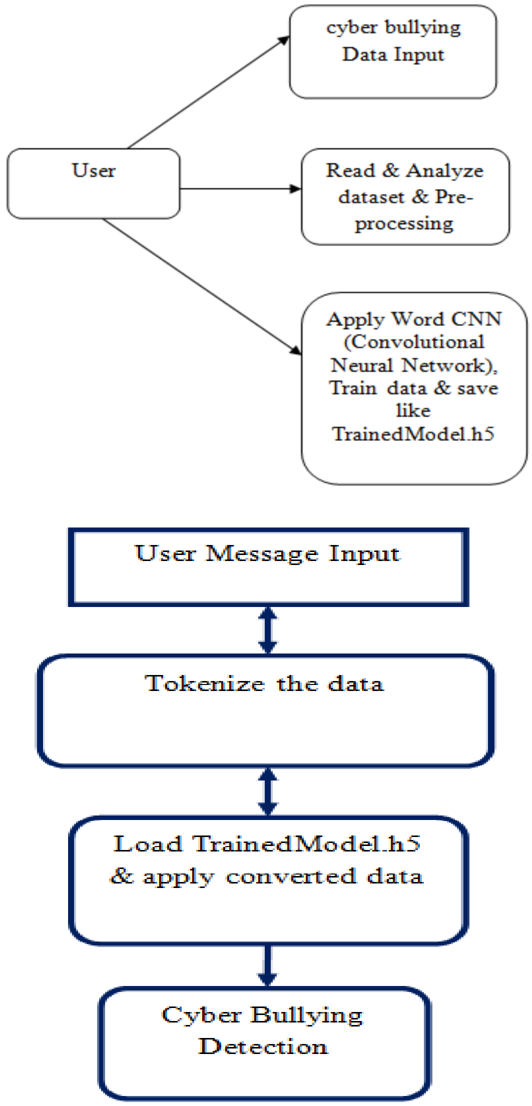

cyber bullying Data Input
User
Read & Analyze dataset & Pre-processing
Apply Word CNN (Convolutional Neural Network), Train data & save like TrainedModel.h5
User Message Input
Tokenize the data
Load TrainedModel.h5 & apply converted data
Cyber Bullying Detection

Fig: 7 Flow Diagrams of Modules

SYSTEM ARCHITECTURE

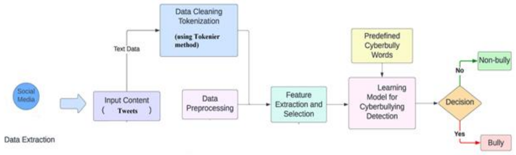
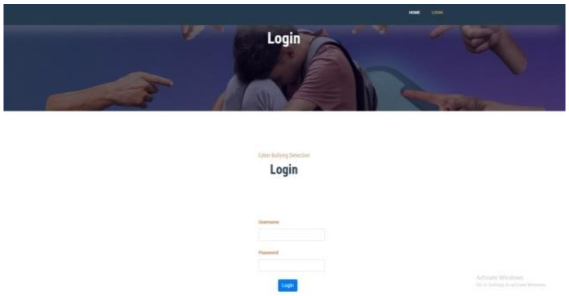
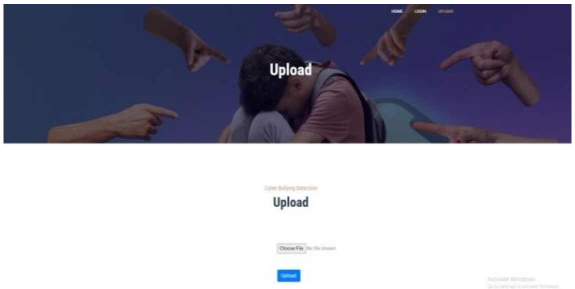
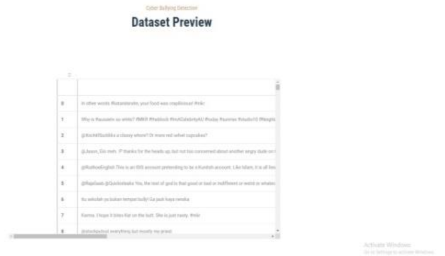

Social Media
Data Extraction
Input Content (Tweets)
Text Data
Data Cleaning Tokenization (using Tokenier method)
Data Preprocessing
Feature Extraction and Selection
Predefined Cyberbully Words
Learning Model for Cyberbullying Detection
Decision
No
Non-bully
Yes
Bully

Fig: 8 System Architecture of Project

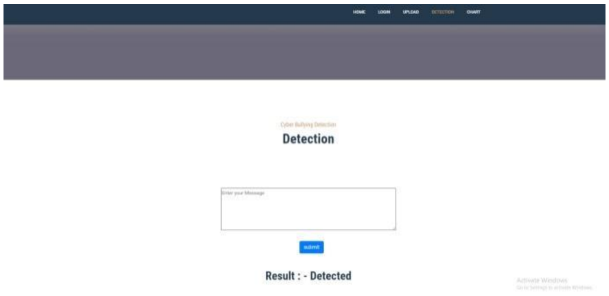
RESULTS AND DISCUSSION

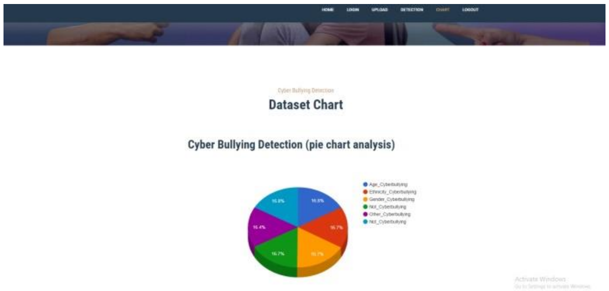






0	In other words #katandandre, your food was crapilicious! #mkr
1	Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise #studio10 #Neighb
2	@XochitlSuzzkk a classy whore? Or more red velvet cupcakes?
3	@Jason_Gio meh. :P thanks for the heads up, but not too concerned about another angry dude on t
4	@RudhoeEnglish This is an ISIS account pretending to be a Kurdish account. Like Islam, it is all lies
5	@Raja5aab @Quickieleaks Yes, the test of god is that good or bad or indifferent or weird or whatev
6	Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka
7	Karma. I hope it bites Kat on the butt. She is just nasty. #mkr
8	@stockputout everything but mostly my priest





FUTURE ENHANCEMENT

For the time being, the effort is restricted to binary text classification for cyber bullying detection, with the goal of determining whether or not a post includes cyber bullying content. Nonetheless, the cyber bullying corpus includes feedback from a range of roles in cyber bullying incidents, such as harassers, victims, onlookers, and non-bullies. You can think of categorizing participant roles in cyber bullying instances in order to expand the research. This would include creating models to recognize both the roles of the people participating in the conversation and the content that constitutes cyber bullying.

Moreover, expanding the research to take into consideration the relationships between posts may be advantageous, since the current method regards each post as a separate entity. You can learn more about the dynamics and patterns of cyber bullying by examining the interactions between users and the posts that make up an episode.

CONCLUSION

In summary, the unexpected increase in cyberbullying brought on by technological innovation has brought attention to the urgent need for effective preventive measures. Since automated detection techniques have the potential to have serious and widespread effects on internet users, they must be developed and implemented. This is a proactive approach that also significantly lowers the frequency of cyber bullying incidents. While previous approaches to categorize cyber bullying have primarily focused on textual traits, this research has adopted a more comprehensive approach by examining a wide range of feature categories. By looking at textual features, sentiment and emotional features, embeddings, psycholinguistic features, word list characteristics, and toxicity factors, we have expanded the set of potential indicators for cyber bullying identification. Word CNN's use in our models has demonstrated to be highly effective, as seen by their astounding 97.06% accuracy rate. This demonstrates how trustworthy and successful the recommended approach is at identifying and putting an end to cyber bullying incidents. The model's high accuracy rate indicates how versatile it is and how well it can identify a wide range of patterns and environments in the complex world of online communication.

REFERENCE

[1]. B. Cagirkan and G. Bilek, ‘‘Cyberbullying among Turkish high school students,’’ Scandin. J. Psychol., vol. 62, no. 4, pp. 608–616, Aug. 2021, doi: 10.1111/sjop.12720.

[2]. P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, ‘‘Online time, experience of cyber bullying and practices to cope with it among high school students in Hanoi,’’ Health Psychol. Open, vol. 7, no. 1, Jan. 2020, Art. no. 205510292093574, doi: 10.1177/2055102920935747.

[3]. A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, ‘‘CyberDect. A novel approach for cyberbullying detection on Twitter,’’ in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989-9_9.

[4]. R. M. Kowalski and S. P. Limber, ‘‘Psychological, physical, and academic correlates of cyberbullying and traditional bullying,’’ J. Adolescent Health, vol. 53, no. 1, pp. S13–S20, Jul. 2013, doi: 10.1016/j.jadohealth.2012.09.018.

[5]. Y.-C. Huang, ‘‘Comparison and contrast of piaget and Vygotsky’s theo-ries,’’ in Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28–32, doi: 10.2991/assehr.k.210519.007.

[6]. A. Anwar, D. M. H. Kee, and A. Ahmed, ‘‘Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.

[7]. D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, ‘‘Cyberbul-lying on social media under the influence of COVID-19,’’ Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175.

[8]. I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.

[9]. R. Garett, L. R. Lord, and S. D. Young, ‘‘Associations between social media and cyberbullying: A review of the literature,’’ mHealth, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01.

[10]. M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, ‘‘Automatic extraction of harmful sentence patterns with application in cyberbullying detection,’’ in Proc. Lang. Technol. Conf. Poznań, Poland: Springer, 2015, pp. 349–362, doi: 10.1007/978-3-319-93782-3_25.

[11]. M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and G. Leliwa, ‘‘"Brute-force sentence pattern extortion from harmful messages for cyberbullying detection,’’’ J. Assoc. Inf. Syst., vol. 20, no. 8, pp. 1075–1127, 2019.

[12]. M. O. Raza, M. Memon, S. Bhatti, and R. Bux, ‘‘Detecting cyber-bullying in social commentary using supervised machine learning,’’ in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020, pp. 621–630.

[13]. D. Nguyen, M. Liakata, S. Dedeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, ‘‘How we do things with words: Analyzing text as social and cultural data,’’ Frontiers Artif. Intell., vol. 3, p. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.

[14]. J. Cai, J. Li, W. Li, and J. Wang, ‘‘Deeplearning model used in text classification,’’ in Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2018, pp. 123–126, doi: 10.1109/ICCWAMTIP.2018.8632592.

[15]. N. Tiku and C. Newton. Twitter CEO: We Suck at Dealing With Abuse. Verge. Accessed: Aug. 17, 2022. [Online]. Available: https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the

[16]. D. Noever, ‘‘Machine learning suites for online toxicity detection,’’ 2018, arXiv:1810.01869.

[17]. D. G. Krutka, S. Manca, S. M. Galvin, C. Greenhow, M. J. Koehler, and E. Askari, ‘‘Teaching ‘against’ social media: Confronting prob-lems of profit in the curriculum,’’ Teachers College Rec., Voice Scholarship Educ., vol. 121, no. 14, pp. 1–42, Dec. 2019, doi: 10.1177/016146811912101410.

[18]. H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. V. Simão, and I. Trancoso, ‘‘Automatic cyberbullying detection: A systematic review,’’ Comput. Hum. Behav., vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.

[19]. S. Bharti, A. K. Yadav, M. Kumar, and D. Yadav, ‘‘Cyberbullying detection from tweets using deep learning,’’ Kybernetes, vol. 51, no. 9, pp. 2695–2711, Sep. 2022.

[20]. A. Bozyiğit, S. Utku, and E. Nasibov, ‘‘Cyberbullying detection: Uti-lizing social media features,’’ Expert Syst. Appl., vol. 179, Oct. 2021, Art. no. 115001, doi: 10.1016/j.eswa.2021.115001.

[21]. H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, ‘‘An abusive text detection system based on enhanced abusive and non-abusive word lists,’’ Decis. Support Syst., vol. 113, pp. 22–31, Sep. 2018, doi: 10.1016/j.dss.2018.06.009.

[22]. Y. Fang, S. Yang, B. Zhao, and C. Huang, ‘‘Cyberbullying detection in social networks using bi-GRU with self-attention mechanism,’’ Informa-tion, vol. 12, no. 4, p. 171, Apr. 2021, doi: 10.3390/info12040171.