Cyberbullying Identification in Social Media: A Comparative Analysis of Machine Learning and Transfer Learning Methodologies

[1]Annam Dinesh Reddy, [2]Ande Sai Vivek, [3]Kolluri Rakesh,
[4]Naredla Akshay Reddy, [5]Patil Manikrao
[1234]Under Graduate, Department of IT-GNITC-Hyderabad
[5]Assistant Professor, Department of IT-GNITC-Hyderabad

ABSTRACT:

Information and Communication Technologies have propelled social networking and communication, but cyber bullying poses significant challenges. Existing user-dependent mechanisms for reporting and blocking cyber bullying are manual and inefficient. Conventional Machine Learning and Transfer Learning approaches were explored for automatic cyber bullying detection. The study utilized a comprehensive dataset and structured annotation process. Textual, sentiment and emotional, static and contextual word embeddings, psycholinguistics, term lists, and toxicity features were employed in the Conventional Machine Learning approach. This research introduced the use of toxicity features for cyber bullying detection. Contextual embeddings of word Convolutional Neural Network (Word CNN) demonstrated comparable performance, with embeddings chosen for its higher F-measure. Textual features, embeddings, and toxicity features set new benchmarks when fed individually. This outperformed Linear SVC in terms of training time and handling high-dimensionality features. Transfer Learning utilized Word CNN for fine-tuning, achieving a faster training computation compared to the base models. Additionally, cyber bullying detection through Flask web was implemented, yielding an accuracy of 97.06%. The reference to the specific dataset name was omitted for privacy.

INRODUCTION:

Information and Communication Technologies (ICT) have become an integral part of everyone's life, evolving imperceptibly with time, catalyzing online communication between people. Communication has been just one button click with the widespread use of the online platform, facilitating the growth of social networking. ICT dominance has a dark side when people easily misuse technological advancement with abusive behaviors such as cyberbullying. Cyberbullying is the expanded form of direct or traditional bullying through electronic platforms. Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyberbullying a complex and challenging mission to protect online communities. Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously, leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems. While cyberbullying victims tend to suffer from mental health problems such as depression, anxiety, loneliness, and anhedonia, some are reported to be committing self-injurious behavior and suicidal ideation.

The expected outcome of this research is the development of classification models that can effectively detect cyberbullying and non-cyberbullying events from unruly posts by applying the knowledge of state of the art in NLP and Deep Learning. This work incorporates text pre-processing, feature engineering, model development using word CNN.

.

LITERATURE SURVEY

I. Cyber Bullying Detection Using Machine Learning.

Cyber bullying has evolved as a severe problem hurting children, teenagers, and young adults as a result of the increasing use of social media. Automatic detection of bullying communications in social media is now possible, thanks to machine learning techniques, which could aid in the creation of a healthy and safe social media environment. One major issue in this important research area is robust and discriminative numerical representation learning of text messages. To address this challenge, we offer a new representation learning method in this study. The Semantic-Enhanced Marginalized Denoising Auto-Encoder (SMSDA) is a semantic enhancement of the popular deep learning model stacked denoising Auto-Encoder. The semantic extension is made up of semantic dropout noise and sparsity constraints, with the semantic dropout noise being the most important.

II.Cyber Bullying Detection for Twitter Using ML Classification Algorithms.

Social networking platforms have given us incalculable opportunities than ever before, and its benefits are undeniable. Despite benefits, people may be humiliated, insulted, bullied, and harassed by anonymous users, strangers, or peers. Cyberbullying refers to the use of technology to humiliate and slander other people. It takes form of hate messages sent through social media and emails. With the exponential increase of social media users, cyberbullying has been emerged as a form of bullying through electronic messages. We have tried to propose a possible solution for the above problem, our project aims to detect cyberbullying in tweets using ML Classification algorithms like Naïve Bayes, KNN, Decision Tree, Random Forest, Support Vector etc. and also we will apply the NLTK (Natural language toolkit) which consist of bigram, trigram, n-gram and unigram on Naïve Bayes to check its accuracy. Finally, we will compare the results of proposed and baseline features with other machine learning algorithms. Findings of the comparison indicate the significance of the proposed features in cyberbullying detection. The study reviewed and identified Naïve bayes N-gram gives the best accuracy and also the system is able to identify the bullied and non-bullied statements.

III. Detecting A Twitter Cyberbullying Using Machine Learning.

Social media is a platform where many young people are getting bullied. As social networking sites are increasing, cyberbullying is increasing day by day. To identify word similarities in the tweets made by bullies and make use of machine learning and can develop an ML model automatically detect social media bullying actions. However, many social media bullying detection techniques have been implemented, but many of them were textual based. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of similar work on the same dataset. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not.

PROPOSED TECHNIQUE USED OR ALGORITHM USED:

➢ Word Convolutional Neural Network (word CNN):

- The suggested method makes use of Word CNN's capabilities to identify cyber bullying. The WORD CNN is adjusted to the unique characteristics of cyber bullying in the dataset by Transfer Learning. By going through this procedure, the model's capacity to identify subtle patterns is improved and training computation is accelerated in comparison to beginning from zero.
- Crucially, the contextual embeddings produced by the WORD CNN show comparable performance with larger F-measures. This method adds to a more complete cyber bullying detection model by expanding on the traditional usage of embeddings and introducing a fresh take on toxicity characteristics. In addition to improving accuracy, the system's use of WORD CNN shows a progressive approach to tackling the changing difficulties associated with online social interactions.

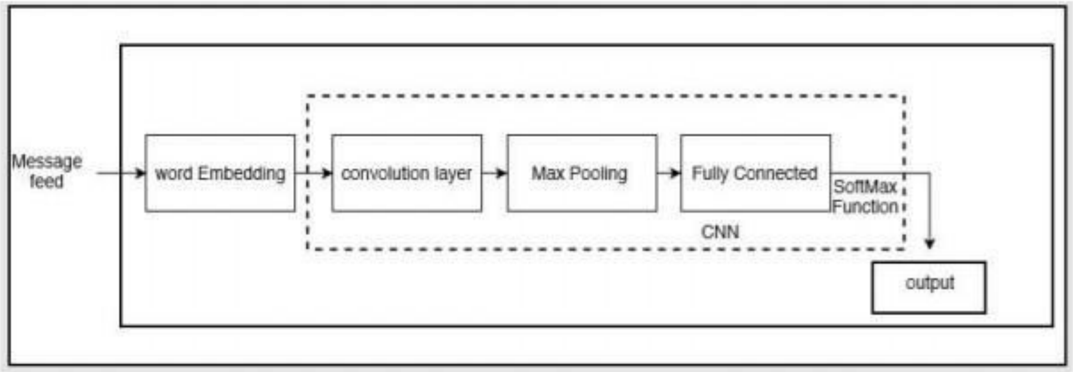

Fig 4.11: System Architecture

IMPLEMENTATION

Convolutional Neural Networks (CNNs) are known to have a good performance on data with a high locality when words get more care weight about the features surrounding them. We are trying to get high priority in the text given their short length and their tendency to focus on cyberbullying. We used CNNs that received input text in the form of sequences of integer representations are arising from the unigrams. The character processing included the conversion of emoticons into word and the removal of non-Latin characters. We also removed frequently occurring URL components (e.g., names of popular websites), metadata encoded in the main body text (e.g., 'RT: '), and a variety of social media platform-specific features. Hashtags and @- mentions were reduced to binary features. The text was then lower-cased and tokenized using NLTK's TweetTokenizer3.The tokenized text was next encoded using a dictionary of integers, with the original ordering of the tokens preserved

CONCLUSION

In conclusion, the unanticipated rise in cyberbullying as a result of technology advancement has highlighted the pressing need for efficient preventive measures. Automated detection methods must be developed and put into place since they have the potential to have severe and broad effects on Internet users. This is a preventative measure that also makes a substantial contribution to reducing the number of cyberbullying incidences. Although textual characteristics have been the mainstay of past techniques for classifying cyberbullying, this research has taken a more thorough approach by exploring many feature categories. We have broadened the range of possible indications for cyberbullying detection by examining textual features, sentiment and emotional features, embeddings, psycholinguistic features, word lists characteristics, and toxicity features. Our models' use of Word CNN has shown to be quite successful, as seen by their remarkable 97.06% accuracy rate. This illustrates how reliable and effective the suggested method is in locating and stopping instances of cyberbullying. The model's excellent accuracy rate demonstrates its adaptability and recognition of many patterns and settings in the intricate world of online communication.

10.2 REFERENCES

[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.

[23]. G. Jacobs, C. Van Hee, and V. Hoste, ‘‘Automatic classification of partici-pant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?’’ Natural Lang. Eng., vol. 28, no. 2, pp. 141–166, Mar. 2022, doi: 10.1017/S135132492000056X.

[24]. M. Gada, K. Damania, and S. Sankhe, ‘‘Cyberbullying detection using LSTM-CNN architecture and its applications,’’ in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2021, pp. 1–6, doi: 10.1109/ICCCI50826.2021.9402412.

[25]. H. H.-P. Vo, H. Trung Tran, and S. T. Luu, ‘‘Automatically detecting cyberbullying comments on online game forums,’’ in Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF), Aug. 2021, pp. 1–5, doi: 10.1109/RIVF51545.2021.9642116.