International Journal of Multidisciplinary Engineering in Current Research - IJMEC

Volume 10, Issue 5, May-2025, http://ijmec.com/, ISSN: 2456-4265

Cyber Harassment Prediction In Social Media Using Word CNN

Manik Rao Patil , Avula Vaishnavi, Jaini Swathi, Jolla Jai Chandra

[1]Assistant Professor, Department Of IT, Guru Nanak Institutions Technical Campus (Autonomous), India.

[2,3,4]b.Tech Students, Department Of IT, Guru Nanak Institutions Technical Campus (Autonomous), India

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.

1- INTRODUCTION

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.

EXISTING SYSTEM

- The existing system focuses on addressing the resource-intensive nature of machine learning (ML) classifier training, particularly with the rising challenge posed by large datasets and the prevalence of Deep Neural Networks (DNN). Feature Density (FD) is analyzed as a means to estimate ML classifier performance before training, aiming to optimize the training process.
- The study underscores the environmental impact of resource-intensive training, especially concerning the escalating CO2 emissions associated with large-scale ML models. The research aims to minimize the demands for powerful computational resources and enhance efficiency in Natural Language Processing, with a specific emphasis on dialog classification, such as cyber bullying detection.

PROPOSED SYSTEM

- The automatic detection method of the proposed system tackles the issue of cyber bullying in social networks. The system uses a combination of textual, sentiment, emotional, static, and contextual information, using a big dataset and structured annotations. This method is distinct in that it incorporates toxicity factors to improve the identification of cyber bullying.

  When it comes to managing high-dimensionality features and training time, the system performs better than Linear SVC. When compared to base models, Transfer Learning enhances performance and speeds up training calculations by fine-tuning WORD CNN. Furthermore, a 97.06% accuracy rate in actual usage is guaranteed by a Flask web implementation

2-LITERATURE SURVEY

- **Title:** Cyber Bullying Detection Using Machine Learning.
- **Author:** K. Siddhartha, K. Raj Kumar, K. Jayanth Varma, M. Amogh, Mamatha Samson,
- **Year:** 2022.
- **Description:** 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.

- **Title:** Cyber Bullying Detection for Twitter Using ML Classification Algorithms.
- **Author:** Muskan Patidar, Mahak Lathi, Manali Jain, Monika Dhakad,
- **Year:** 2021.
- **Description**: 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.

- **Title:** Detecting A Twitter Cyberbullying Using Machine Learning.
- **Author:** Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe,
- **Year:** 2020.
- **Description:** 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.

- **Title:** Brute-Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection.
- **Author:** M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and G. Leliwa,
- **Year:** 2019.
- **Description**: Cyberbullying, or humiliating people using the Internet, has existed almost since the beginning of Internet communication. The relatively recent introduction of smartphones and tablet computers has caused cyberbullying to evolve into a serious social problem. In Japan, members of a parent-teacher association (PTA) attempted to address the problem by scanning the Internet for cyberbullying entries. To help these PTA members and other interested parties confront this difficult task we propose a novel method for automatic detection of malicious Internet content. This method is based on a combinatorial approach resembling brute-force search algorithms, but applied in language classification. The method extracts sophisticated patterns from sentences and uses them in classification. The experiments performed on actual cyberbullying data reveal an advantage of our method vis-à-vis previous methods. Next, we implemented the method into an application for Android smartphones to automatically detect possible harmful content in messages. The method performed well in the Android environment, but still

needs to be optimized for time efficiency in order to be used in practice.

- **Title:** An Abusive Text Detection System Based on Enhanced Abusive and Non-Abusive Word Lists.
- **Author**: H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han,
- **Year:** 2018**.**
- **Description:** Abusive text (indiscriminate slang, abusive language, and profanity) on the Internet is not just a message but rather a tool for very serious and brutal cyber violence. It has become an important problem to devise a method for detecting and preventing abusive text online. However, the intentional obfuscation of words and phrases makes this task very difficult and challenging. We design a decision system that successfully detects (obfuscated) abusive text using an unsupervised learning of abusive words based on word2vec's skip-gram and the cosine similarity. The system also deploys several efficient gadgets for filtering abusive text such as blacklists, n-grams, edit-distance metrics, mixed languages, abbreviations, punctuation, and words with special characters to detect the intentional obfuscation of abusive words. We integrate both an unsupervised learning method and efficient gadgets into a single system that enhances abusive and non-abusive word lists. The integrated decision system based on the enhanced word lists shows a precision of 94.08%, a recall of 80.79%, and an f-score of 86.93% in malicious word detection for news article comments, a precision of 89.97%, a recall of 80.55%, and an f-score 85.00% for online community comments, and a precision of 90.65%, a recall of 93.57%, and an f-score 92.09% for Twitter tweets. We expect that our approach can help to improve the current abusive word detection system, which is crucial for several web-based services including social networking services and online games.

3-PROJECT DESCRIPTION

This research employs state-of-the-art Natural Language Processing (NLP) and Deep Learning techniques to fight the growing threat of cyberbullying in online environments. By means of a detailed study of several attributes, such as signs of toxicity, the study aims to construct strong categorization models. These models are designed to outperform conventional machine learning techniques in precisely detecting instances of cyberbullying in textual online communication. They are based on Word CNN for contextual embeddings. A Flask online platform for real-time cyberbullying detection that aims for a high degree of identification and preventive accuracy serves as an example of how these improvements are put into practice.

METHODOLOGIES

1) Dataset:

We obtained a dataset for the purpose of detecting cyberbullying during the project's first phase. The collection, which comes from "cyberbullying_tweets.csv," is made up of different text entries i.e., tweets. Included in the dataset are 47,692 information that have been classified as either "not_cyberbullying," "gender," "religion," "other_cyberbullying," "age," or "ethnicity."

2) **Importing the Necessary Libraries:**

We decided to programme in Python and loaded the necessary project libraries. Key libraries include PIL for turning photos into arrays, scikit-learn for splitting the data into training and testing sets, and other common libraries like pandas, numpy, matplotlib, and TensorFlow. Keras is used to build the primary model.

3) Data Pre-processing:

We used pandas to read the CSV file, info() to do a first data inspection, and handling any missing values. The 'cyberbullying_type' labels were then encoded using Label Encoding, which changed the textual data. Next, we divided the dataset into sets for testing, validation, and training.

4) Model Creation for Word CNN:

We use a word Convolutional Neural Network (CNN) as they have proven to be successful at document classification problems. A conservative word CNN configuration is used with 128 filters (parallel fields for processing words) and a kernel size of 5 with a rectified linear ('relu') activation function. This is followed by a pooling layer that reduces the output of the convolutional layer.

5) Training and Evaluation:

Using the fit function, we trained the Word CNN model by setting hyperparameters such batch size and epochs. An average of 97.06% was achieved in training, while an average of 99.92% was achieved in validation. We then assessed the model using the test set, and 97.7% accuracy was obtained.

4-REQUIREMENTS ENGINEERING

We can see from the results that on each database, the error rates are very low due to the discriminatory power of features and the regression capabilities of classifiers. Comparing the highest accuracies (corresponding to the lowest error rates) to those of previous works, our results are very competitive.

3.2 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It should what the system do and not how it should be implemented.

- PROCESSOR : DUAL CORE 2 DUOS.
- RAM : 4GB DD RAM
- HARD DISK : 250 GB

SOFTWARE REQUIREMENTS

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

- Operating System : Windows 7/8/10
- Platform : Vs Code/ Spyder3
- Programming Language : Python
- Front End : HTML, CSS

FUNCTIONAL REQUIREMENTS

A functional requirement defines a function of a software-system or its component. A function is described as a set of inputs, the behavior, Firstly, the system is the first that achieves the standard notion of semantic security for data confidentiality in attribute-based deduplication systems by resorting to the hybrid cloud architecture.

NON-FUNCTIONAL REQUIREMENTS

The major non-functional Requirements of the system are as follows

Usability

The system is designed with completely automated process hence there is no or less user intervention.

Reliability

The system is more reliable because of the qualities that are inherited from the chosen platform python. The code built by using python is more reliable.

Performance

This system is developing in the high level languages and using the advanced back-end technologies it will give response to the end user on client system with in very less time.

5- DESIGN ENGINEERING

Design Engineering deals with the various UML [Unified Modelling language] diagrams for the implementation of project. Design is a meaningful engineering representation of a thing that is to be built. Software design is a process through which the requirements are translated into representation of the software. Design is the place where quality is rendered in software engineering.

SYSTEM ARCHITECTURE:

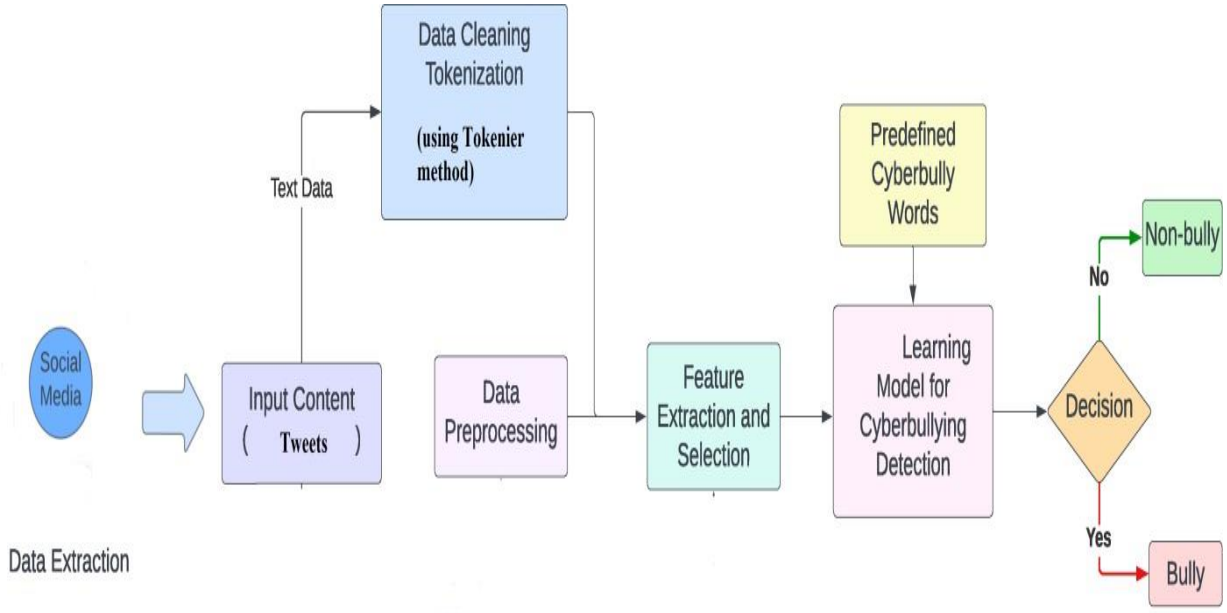

Fig 4.11: System Architecture

6- CONCLUSIONAND REFERENCES

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.

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,'' Information, 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.