JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BILLY – BUDDY AGAINST CYBERBULLYING

Dr. H M Manjula, Yashraj, Nayeem Laheji, Shivakumar Ramu Kamate, Bharath G
Department of Computer Science and Engineering,
Presidency University, Bengaluru, India

Abstract : This research addresses the growing concern of cyberbullying in the digital age, proposing a comprehensive solution through a chatbot-driven detection and support system. The system integrates a GPT-3.5 Turbo-powered chatbot named "Billy" for victim support and evidence submission, along with real-time communication features such as WebRTC video calls and Socket.IO-powered chats to foster community support. Built on a robust backend using Express.js and PostgreSQL and an intuitive React.js frontend, the platform ensures secure data handling and user anonymity. This multi-faceted approach aims to not only detect but also prevent cyberbullying, providing victims with immediate relief and fostering a safer online environment. Utilizing a manually labeled dataset of 20,000 tweets, a Logistic Regression model was developed for accurate detection of harmful content, achieving an accuracy of 92%.

Keywords **- Cyberbullying Detection, Chatbot Development, Machine Learning.**

I. INTRODUCTION

The number of people using social media is increasing every year. According to reports, 266 million new users started using social media for the first time over the year 2023, bringing the global total to over 5 billion users in 2024 [1]. The most popular social media includes WhatsApp, YouTube, Facebook, Instagram. The rise of the internet has also contributed to a surge in cybercrimes like phishing, spread of malwares and cyberbullying [2]. Reducing cybercrimes specially cyberbullying is very much crucial because various negative health outcomes have been observed among the victims who were affected by cyberbullying. Cyberbullying is increasingly recognized as a serious public health issue, with victims facing a significantly higher risk of suicidal thoughts. According to a 2023 study by the national anti-bullying charity, Ditch the Label, two out of three individuals aged 13 to 22 surveyed reported experiencing cyberbullying [3].

Detecting cyberbullying requires intelligent systems capable of automatically identifying potential threats. The primary approach involves utilizing natural language processing (NLP), machine learning (ML), and deep learning (DL) methods. These algorithms analyze text messages received by users to identify negative comments. Cyberbullying occurs across various platforms and in multiple forms, making its detection on social media particularly challenging. This is due to the short, noisy, and unstructured nature of social media content, along with users often being unaware they are being cyberbullied. Additionally, cyberbullying spans wide range of data types, such as text, images, audio, video, and emojis.

Text classification with machine learning involves assigning a text document to one of a set of predefined categories using a machine learning approach. This process typically relies on analyzing selected features extracted from the text documents [4]. Naive Bayes is frequently employed in text classification tasks due to its simplicity and effectiveness [5]. However, its performance can suffer because it does not inherently capture the complexities of textual data. Schneider identified these limitations and demonstrated that they can be mitigated with a few straightforward adjustments [6]. Lim proposed a method to improve the efficiency of KNN-based text classification by employing precisely determined parameters [7]. Ensemble methods such as random forest classifiers are also suitable for dealing with noisy and highly dimensional data in text classification. A Random Forest model consists of numerous decision trees, each constructed using randomly chosen subsets of features. Given the instance, the prediction from Random Forest Model is obtained by majority voting of all trees in the forest [8]. Logistic Regression is also considered a powerful choice when the outcome variable is “dichotomous” [9].

The primary goal of this research is to design and implement a chatbot named "Billy" leveraging machine learning algorithms, designed to provide immediate support to victims, assist in evidence submission, and ensure anonymity. Moreover, an intuitive web interface will offer resources for reporting incidents, accessing educational content, and joining support communities. By integrating educational and community-building features, the proposed solution aims not only to detect but also to prevent cyberbullying, fostering a safer online environment.

The structure of the paper is as follows: Section II presents a review of related research, Section III details the proposed methodology, Section IV provides an analysis of the results, and Section V concludes with key insights and potential future research directions.

II. REVIEW OF LITERATURE

Over the past few years, numerous studies have been proposed concerning the detection of cyberbullying. This research specifically focuses on development of a machine learning model to for textual cyberbullying detection and reporting, as written content remains the dominant mode of interaction on social media platforms. The discussion of related work in this area emphasizes the datasets utilized, the detection approaches adopted, and the strategies implemented for cyberbullying prevention.

Many researchers have leveraged datasets from platforms such as Fromspring.me, ask.fm, Twitter, YouTube, and others for studying cyberbullying. [10] investigated various datasets to explore effects of text analysis methods on the automatic detection of cyberbullying. [11] utilized Chinese Weibo and English Twitter datasets and introduced a Character-Level Convolutional Neural Network with Shortcuts model for identifying cyberbullying in social media posts. [12] proposed an automated system for detecting cyberbullying on social media through analyzing posts from harassers, harmed individuals, and bystanders, using English and Dutch datasets. [13] applied Instagram datasets for cyberbullying detection. [14] employed deep neural networks for identifying cyberbullying text in Twitter datasets.

Machine learning techniques are extensively used to learn from data efficiently, enabling systems to make informed decisions on complex issues. These algorithms help in the timely and automatic detection of cyberbullying. [15] developed a multi-classifier based on machine learning to categorize cyberbullying severity into distinct levels. [16] suggested the Random Forest algorithm as a means to classify cyberbullying into various categories, including aggressor, spammer, bully, and normal. [17] and [18] also employed machine learning methods to identify cyberbullying. Feature extraction plays an important role in classification, with new predictors identified to enhance detection. However, the abundance of features that can emerge necessitates careful extraction and selection, posing computational difficulties.

Deep learning provides the advantage of detecting patterns in unstructured data like text, images, sound, and video, with the ability to automatically learn features from data. Recent research has proposed various deep learning models for cyberbullying detection. [19] applied deep neural networks for this purpose. [20] developed an LSTM model for detecting and reporting cyberbullying. [21] used Twitter data for cyberbullying detection by incorporating both textual and socio-contextual features into the predictive model. [22] worked with Ask.fm and crowdsourced datasets for detecting cyberbullying. [23] presented deep learning models for the same purpose. [24][25] implemented simple CNN, hybrid CNN-LSTM, and combined these three models, trained using Word2vec on datasets from Twitter, Google and Formspring.

The current project introduces several significant contributions. First, it presents a manually labeled cyberbullying dataset, which was created by scraping tweets from Twitter to ensure the data's accuracy and relevance. A Logistic Regression model was trained to efficiently detect instances of cyberbullying. Additionally, an easy-to-use chatbot powered by GPT-3.5 Turbo was created to facilitate smooth user interactions. The paper also demonstrates a method for detecting live instances of cyberbullying texts and comments, highlighting its practical application. Lastly, a comprehensive system was developed to integrate all these components, offering a robust solution for combating cyberbullying.

III. METHODOLOGY

The text classification process typically involves several key steps: data collection, text pre-processing, feature extraction, text representation, and the training and evaluation of classifiers [26].

3.1 Data Collection:-

The dataset was acquired from Twitter by searching for commonly used slurs and terms associated with religious, sexual, gender, and ethnic minorities across different regions worldwide [27]. It comprises 47,692 annotated tweets, of which 39,747 are labeled as positive for cyberbullying, while the remaining tweets are categorized as non-sexist and non-racist. Due to limited computational resources, this study focuses on a subset of 20,000 samples, with 8,000 of them annotated as one of the cyberbullying types. This data is manually labelled. The distribution of labels is shown in Fig 1.1. For simplification of the cyberbullying detection task, the output labels were transformed into binary numbers where 0 represent the text does not indicate cyberbullying and 1 indicate the text indicates cyberbullying. Fig 1.2 represents commonly used terms in the dataset.

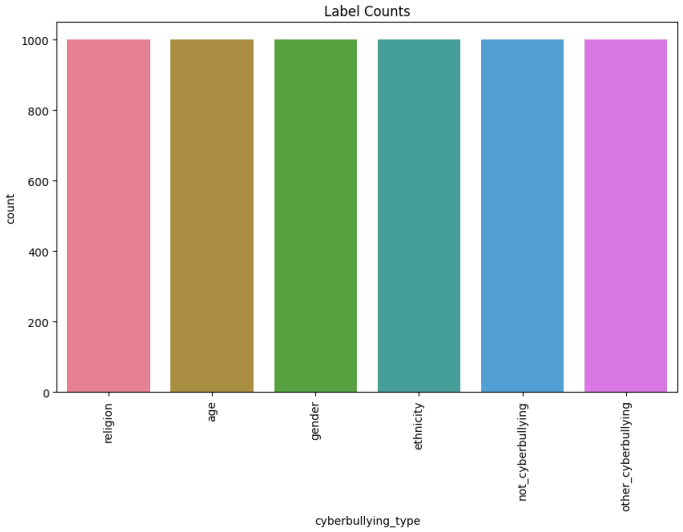

Figure 1.1 Distribution of output labels

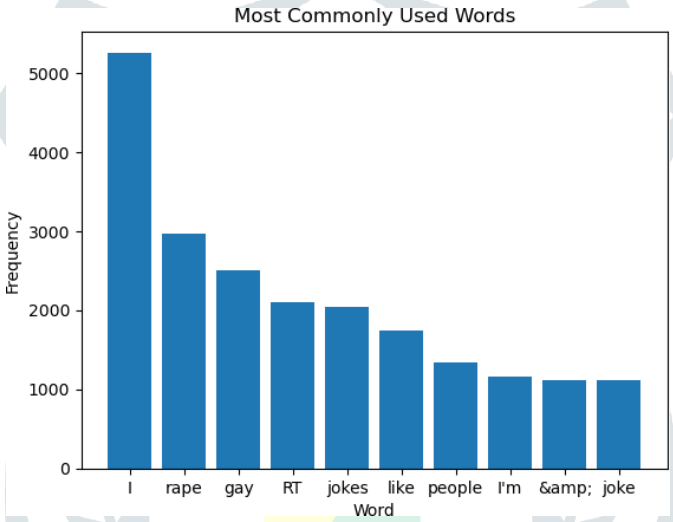

Figure 1.2 Frequent terms in the dataset

3.2 Text Pre-processing:-

Before performing feature extraction, the data was preprocessed using a number of methods described below in order.

Lowercasing: It ensures consistency by treating words with the same meaning but different cases as identical [28] and also reduce the vocabulary size. For example, the word 'Happy' and 'HAPPY' might be considered different and would be assigned different vectors when the text is vectorized for feature extraction.

Removal of non-alphabetic characters, URL, hashtags and mentions: Raw data contains a lot of instances of punctuations or special characters which are neither of much importance nor understood by the machine. Hence, its presence in the data acts as noise and must be removed [28].

Acronym and Slang Expansion: The rise of digital communication has led to the emergence of slang and acronyms, commonly known as microtext [29]. Example "OMG u ok?" instead of "Oh my god, are you okay?". Studies indicate that expanding acronyms enhances the precision of sentiment analysis by 4% [29]. We manually collected list of acronyms and slangs to appropriately convert the raw text into general English form.

Stop-word removal: Stop words (prepositions, articles, pronouns and conjunctions) [30] are typically excluded to minimize the computational load during data training [31]. In this study, we use the stop-word list provided by the NLTK library [32], a widely recognized tool for text processing.

Tokenization: This is the method of breaking text into smaller units known as tokens while excluding irrelevant characters [33]. The word_tokenize() function from the NLTK library has been employed in this research to extract tokens from the text.

3.3 Feature Extraction:-

To address the need for fixed-length inputs in machine learning algorithms, the Bag of Words (BoW) method was employed. This method is a simple and widely used approach for creating fixed-length feature representations. To maintain the semantic significance

of the text, Term Frequency-Inverse Document Frequency (TF-IDF) was applied as a statistical technique to determine the significance of keywords within a document [34]. Term Frequency (TF) measures how often a term appears in a document [35]. For instance, if the document "D1" contains 150 words and the term "Artificial" occurs 5 times, the TF for "Artificial" in "D1" would be calculated as follows:

$$\text{Term Frequency} = 5/150 = 0.033$$

Inverse Document Frequency (IDF) assigns greater weights to rare terms and lower weights to terms that occur frequently across documents. For instance, if "tech" appears in 3 out of 10 documents, the IDF for the word "tech" would be:

$$\text{Inverse Document Frequency} = \log_e(10/3) = 1.204$$

TF-IDF is calculated by multiplying the Term Frequency (TF) of a word in a document by its Inverse Document Frequency (IDF) across the entire corpus. The resulting feature matrix is then input into a supervised machine learning algorithm for further processing.

3.4 Model training and evaluation:-

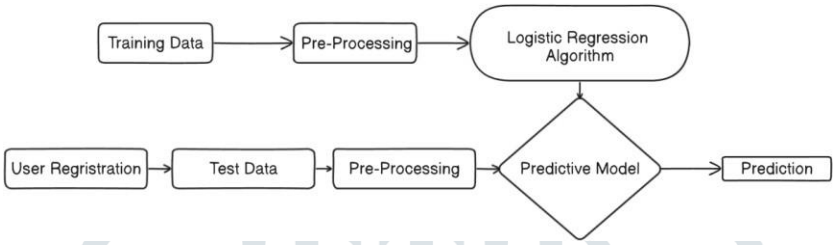

Figure 1.3 Proposed Cyberbullying Detection Method

In the first phase of the process, the model development begins with importing the necessary libraries. NLTK is used for pre-processing the vocabulary, while required python libraries such as scikit-learn, pandas, numpy are imported to facilitate the building of the model. Next, the dataset is loaded, consisting of Twitter tweet data that has been manually labelled into different categories. The dataset is divided into training and testing subsets, with 20% allocated for testing. After the dataset is prepared, different machine learning classifiers are used from the Scikit-Learn library. The output classes are defined, with the first class labeled as "not bullying", and the remaining classes are considered instances of bullying. Each classifier is trained using a baseline model. The model's performance is assessed using key metrics, including accuracy, precision, recall, and F1 score.

3.5 K-fold Cross Validation:-

It is a technique used in statistics to assess the effectiveness of a machine learning model's performance [30]. This research employs K-fold cross validation using k for 5 folds. The data is randomized and divided into k groups. k−1 folds are utilized to train the model, while the remaining fold is used to evaluate its performance. After the completion of this process, the mean score is obtained from score samples. The result for k-fold cross validation is presented in section IV.

3.6 System Design:-

The system is designed to support victims of cyberbullying by integrating advanced technologies for assistance, evidence analysis, and secure data handling. At the heart of the system is a chatbot supported by the OpenAI GPT-3.5 Turbo API, embedded in a React-based frontend. This chatbot provides emotional comfort to victims while enabling them to report incidents of cyberbullying by submitting evidence, such as links, images, or text, for further analysis. The submitted evidence is processed by a logistic regression classifier that determines whether the content qualifies as cyberbullying. This predictive capability ensures timely and reliable detection of harmful behavior.

Additionally, the platform fosters a sense of community by providing features for real-time interaction, such as WebRTC-based video calls and Socket.IO-powered chat functionality. These tools allow victims to connect with support groups or other affected individuals, offering both emotional and practical assistance.

The backend, developed using Express.js, ensures secure management of user authentication and evidence submission. It acts as the central communication layer, connecting the frontend, classifier, and database. All evidence and related data are securely stored in a PostgreSQL database, which prioritizes user confidentiality while enabling authorized access for further analysis or action. This seamless integration of AI-driven support, real-time communication, and robust backend infrastructure makes the system an effective tool for addressing cyberbullying.

IV. RESULTS AND DISCUSSION

Four different classification algorithms were trained against the dataset, Random Forest Classifier, Multinomial NB, Logistic Regression and K-Nearest Neighbor. Each model was assessed using accuracy as the primary metric. The Logistic Regression classifier attained the highest accuracy of 91.73%, followed closely by the Random Forest Classifier with 91.62%. The Multinomial Naive Bayes model attained 85.55%, while the KNN model achieved an accuracy of 79.73%. A bar chart illustrating these results is presented in Fig 1.4.

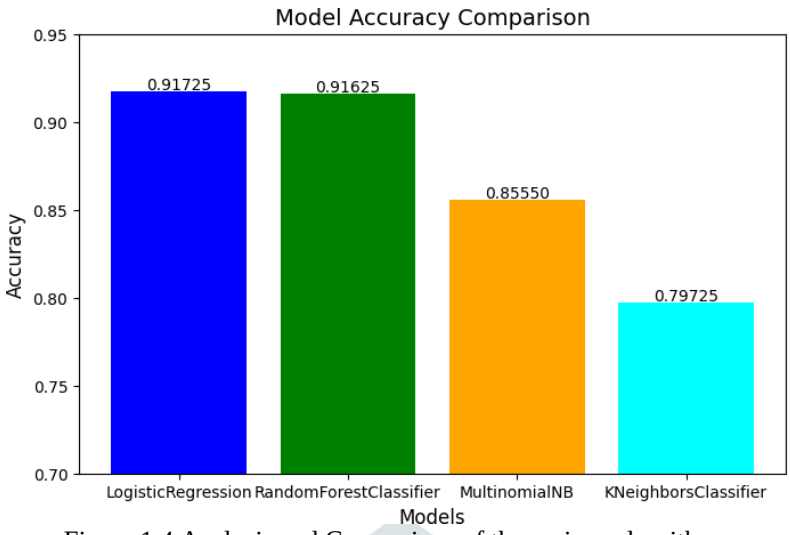

Figure 1.4 Analysis and Comparison of the various algorithms

The confusion matrix illustrates the model's performance in a binary classification problem. The matrix highlights the following metrics:

- **True Negatives (1461)**: The model accurately identified 1461 instances as negative.
- **False Positives (139)**: The model mistakenly classified 139 negative instances as positive.
- **False Negatives (257)**: The model failed to recognize 257 positive instances.
- **True Positives (2143)**: The model correctly classified 2143 instances as positive.

These values indicate that the model has relatively strong predictive accuracy, particularly for the positive class, as evident by the high number of true positives. The false negative count, while not negligible, may require attention to improve sensitivity, especially if the application demands minimizing missed detections.

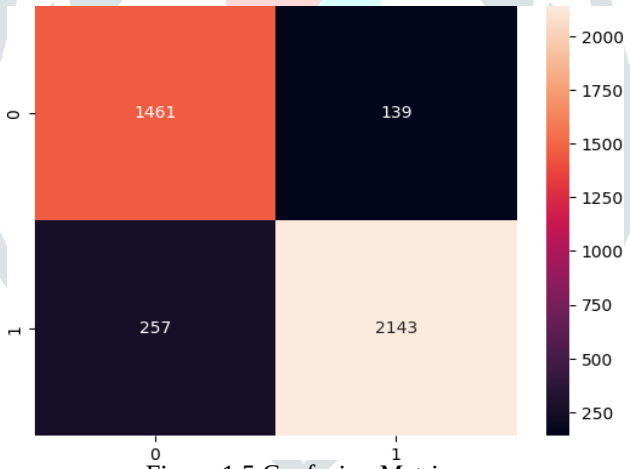

Figure 1.5 Confusion Matrix

Table 1.1 summarizes the results of a k-fold cross-validation procedure with k=5. Each subset was used for testing once, while the remaining subsets were used for training. The training and testing sizes for each fold were 16,000 and 4,000, respectively. The accuracy values obtained across the folds were consistent, ranging from 89.2% to 90.1%. The average accuracy across all folds was 89.7%, demonstrating the robustness of the model.

Table 1.1 K-fold Cross Validation result

Subset	Training Size	Testing Size	Accuracy (%)
1	16000	4000	90.1
2	16000	4000	89.8
3	16000	4000	89.2
4	16000	4000	89.6
5	16000	4000	90.0
Average			89.7

The performance of the proposed framework was assessed based on essential quality metrics, including precision, recall, F1 score, and accuracy. Accuracy reflects the percentage of correctly identified cyberbullying words, with the model demonstrating 92% effectiveness.

	precision	recall	f1-score	support
0	0.87	0.93	0.90	1600
1	0.95	0.91	0.93	2400
accuracy			0.92	4000
macro avg	0.91	0.92	0.91	4000
weighted avg	0.92	0.92	0.92	4000

Figure 1.6 Key performance metrics of Logistic Regression Model

The Logistic Regression classifier was ultimately integrated into the system due to its high accuracy and interpretability, ensuring reliable and timely detection of cyberbullying incidents.

V. CONCLUSION

The fight against cyberbullying requires innovative, comprehensive, and adaptive solutions. This research has highlighted existing gaps in cyberbullying detection systems, including limited dataset diversity, simplistic feature selection, lack of real-time intervention mechanisms, and insufficient multi-modal integration. In response to these challenges, the proposed system offers a robust, multi-faceted approach that integrates advanced machine learning models, a secure and scalable backend infrastructure, and an empathetic, real-time support chatbot, "Billy." Through these components, the system aims to address the nuanced and evolving nature of cyberbullying while prioritizing user confidentiality and providing immediate relief to victims. The methodology encompasses diverse dataset collection and preparation to ensure comprehensive coverage of various forms of online harassment. It incorporates cutting-edge technologies like logistic classifiers and LLM APIs for accurate detection, supported by real-time features such as WebRTC video calls and a community-driven platform. Additionally, the backend system, powered by Express.js and PostgreSQL, ensures the secure handling of user data and evidence while maintaining high scalability and reliability. The frontend, built with React.js, provides an intuitive interface that is accessible across demographics, further broadening the system's reach and usability.

By addressing these potential areas for improvement, the system could evolve into a state-of-the-art solution, capable of mitigating the pervasive issue of cyberbullying on a global scale. The research underscores the importance of a multi- pronged approach that blends technological innovation with empathy and user empowerment, offering a roadmap toward a safer and more inclusive digital world.

REFERENCES

[1] https://datareportal.com/reports/digital-2024-july-global-statshot.

[2] Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) 'Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network', Computers in Human Behavior, Vol. 63, pp.433–443.

[3] Zhao, Rui & Zhou, Anna & Mao, Kezhi. (2016). Automatic Detection of Cyberbullying on Social Networks based on Bullying Features. 10.1145/2833312.2849567.

[4] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401–3409. https://doi.org/10.1016/j.aej.2021.02.009.

[5] kim S. B., Rim H. C., Yook D. S. and Lim H. S., "Effective Methods for Improving Naive Bayes Text Classifiers", LNAI 2417, 2002, pp. 414-423.

[6] Schneider, K., Techniques for Improving the Performance of Naive Bayes for Text Classification, LNCS, Vol. 3406, 2005, 682-693.

[7] Heui Lim, Improving kNN Based Text Classification with Well Estimated Parameters, LNCS, Vol. 3316, Oct 2004, Pages 516 - 523.

[8] Islam, Md. Zahidul & Liu, Jixue & Li, Jiuyong & Liu, Lin & Kang, Wei. (2019). A Semantics Aware Random Forest for Text Classification. 1061-1070. 10.1145/3357384.3357891.

[9] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[10] Ç. ACI, E. ÇÜRÜK, and E. S. EŞSİZ, "Automatic Detection of Cyberbullying in Formspring.Me, Myspace and Youtube Social Networks," Turkish J. Eng., vol. 3, no. 4, pp. 168–178, 2019, doi: 10.31127/tuje.554417.

[11] N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K. K. R. Choo, "Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts," Concurr. Comput. , no. July 2019, pp. 1–11, 2020, doi: 10.1002/cpe.5627.

[12] C. Van Hee et al., "Automatic detection of cyberbullying in social media text," PLoS One, vol. 13, no. 10, pp. 1–22, 2018, doi: 10.1371/journal.pone.0203794.

[13] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," SIAM Int. Conf. Data Mining, SDM 2019, pp. 235–243, 2019, doi: 10.1137/1.9781611975673.27.

[14] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019, pp. 604–607, 2019, doi: 10.1109/ICACCS.2019.8728378.

[15] Talpur BA, O'Sullivan D (2020) "Cyberbullying severity detection: A machine learning approach". PLoS ONE 15(10): e0240924. https://doi.org/10.1371/journal.pone.0240924.

[16] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using Big Five and Dark Triad features," Pers. Individ. Dif., vol. 141, no. January, pp. 252–257, 2019, doi: 10.1016/j.paid.2019.01.024.

[17] Bakshi and A. K. Patel, "Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection," vol. 2, no. January. Springer Singapore, 2019.

[18] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell, "Machine learning and semantic analysis of in-game chat for cyberbullying," Comput. Secur., vol. 76, pp. 197–213, 2018, doi: 10.1016/j.cose.2018.02.016.

[19] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019, pp. 604–607, 2019, doi: 10.1109/ICACCS.2019.8728378.

[20] Vijayakumar, V., & Hari Prasad, D. (Year). Intelligent Chatbot Development for Text-Based Cyberbullying Prevention. International Journal of New Innovations in Engineering and Technology, Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore, Tamilnadu, India.

[21] N. Tahmasbi and E. Rastegari, "A Socio-contextual Approach in Automated Detection of Cyberbullying," Proc. 51st Hawaii Int. Conf. Syst. Sci., vol. 9, pp. 2151–2160, 2018, doi: 10.24251/hicss.2018.269.

[22] Emmery, C., Verhoeven, B., De Pauw, G. et al. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Lang Resources & Evaluation (2020). https://doi.org/10.1007/s10579-020-09509-1.

[23] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models," vol. 12393 LNCS, pp. 245–255, 2020, doi: 10.1007/978-3-030-59065-9_20.

[24] H. Rosa et al., "Automatic cyberbullying detection: A systematic review," Comput. Human Behav., vol. 93, no. October 2018, pp. 333–345, 2019, doi: 10.1016/j.chb.2018.12.021.

[25] H. Rosa, D. Matos, R. Ribeiro, L. Coheur and J. P. Carvalho, "A "Deeper" Look at Detecting Cyberbullying in Social Networks," 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489211.

[26] C. Li, G. Zhan, and Z. Li, "News text classification based on improved Bi-LSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893: IEEE.

[27] Jason Wang, Kaiqun Fu, Chang-Tien Lu. (2020). Fine-Grained Balanced Cyberbullying Dataset. IEEE Dataport. https://dx.doi.org/10.21227/kn1c-zx22.

[28] Jakhotiya, A., Jain, H., Jain, B., & Chaniyara, C. (2022). Text pre-processing techniques in natural language processing: A review. International Research Journal of Engineering and Technology (IRJET), 9(2). e-ISSN: 2395-0056, p-ISSN: 2395-0072. https://www.irjet.net.

[29] Satapathy, R.; Guerreiro, C.; Chaturvedi, I.; Cambria, E. Phonetic-based Microtext Normalization for Twitter Sentiment Analysis. In Proceedings of the IEEE International Conference On Data Mining Workshops, New Orleans, LA, USA, 18–21 November 2017; pp. 407– 413.

[30] J. Brownlee, Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end. Machine Learning Mastery, 2016.

[31] Sinka, Mark & Corne, David. (2003). Evolving Better Stoplists for Document Clustering and Web Intelligence.. 1015-1023.

[32] Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.

[33] Manning, C.D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval; Cambridge University Press: Cambridge, UK, 2008.

[34] Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25. https://doi.org/10.5120/ijca2018917395.

[35] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2015). "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology, (ICITEE), 2014.