IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REALTIME CYBERBULLYING DETECTION USING ML AND NLP

[1]Mrs. Aarti Jadhav Patil, [2]Meghna Nakhate, [3]Ruhika Bulani, [4]Reshu Verma, [5]Astha Jain

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1]Department of Information Technology,

[1]D. Y. Patil College of Engineering, Akurdi, Pune, Maharashtra, India

Abstract: The use of social media has grown exponentially over time with the growth of the Internet and has become the most effective networking platform in the 21st century. However, enhancing social connectivity often creates negative impacts on the society that contribute to a couple of destructive phenomena such as online abuse, harassment, cyberbullying, cybercrime, and online trolling. Cyberbullying frequently leads to serious mental and physical distress, particularly for women and children, and even sometimes forces them to attempt suicide. Online harassment attracts attention due to its strong negative social impact. Many incidents have recently occurred worldwide due to online harassment, such as sharing private chats, rumors, and sexual remarks. Therefore, the identification of bullying texts or messages on social media has gained a growing amount of attention among researchers. The purpose of this research is to design and develop an effective technique to detect online abusive and bullying messages by merging natural language processing and machine learning. Term frequency-inverse text frequency (TFIDF) is used to analyze the accuracy level of four distinct machine learning algorithms.

Keywords: Cyberbullying, machine learning, natural language processing, term frequency-inverse document frequency, N-gram

I. INTRODUCTION

The Internet itself has been transformed. In its early days—which from a historical perspective are still relatively recent—it was a static network designed to shuttle a small freight of bytes or a short message between two terminals; it was a repository of information where the content was published and maintained only by expert coders. Today, however, immense quantities of information are uploaded and downloaded over this electronic leviathan, and the content is very much our own, for now, we are all commentators, publishers, and creators. People no longer spend hours gazing at a computer screen after work or class; instead, they use their mobile devices to stay online everywhere, all the time. The Internet has become embedded in every aspect of our day-to-day lives, changing the way we interact with others. The Internet has clearly impacted all levels of education by providing unbounded possibilities for learning. We believe that the future of education is a networked future. The network of networks is an inexhaustible source of information. What's more, the Internet has enabled users to move away from their former passive role as mere recipients of messages conveyed by conventional media to an active role, choosing what information to receive, how, and when. The information recipient even decides whether or not they want to stay informed.

II. BACKGROUND

Cyberbullying is a pervasive and growing problem in the digital age. The widespread use of social media, online chat rooms, and other digital communication channels has led to an increase in the number of incidents of cyberbullying, which can have serious and long-lasting effects on the mental health and well-being of victims. Cyberbullying involves the use of electronic communication to harass, intimidate, or bully others, often anonymously or from behind a screen.

The challenge of detecting and preventing cyberbullying has received increasing attention from researchers, online platforms, and policymakers. The scale and complexity of online communication make it difficult to manually monitor and moderate content, and traditional keyword-based filters are often ineffective due to the evolving nature of cyberbullying tactics and the use of subtle and indirect forms of bullying. As a result, there is a need for automated approaches to cyberbullying detection that can accurately identify and flag instances of cyberbullying in real-time.

Recent research has shown that machine learning techniques, including natural language processing and text classification, can be effective in detecting cyberbullying in online communication. In particular, the use of n-gram and TF-IDF features, combined with supervised machine learning models such as Support Vector Machines, Naive Bayes, and Random Forest, has shown promising results in the detection of cyberbullying. However, there is still a need for further research to explore the effectiveness of different feature sets and machine learning models for cyberbullying detection, as well as to better understand the characteristics and dynamics of cyberbullying in online communication.

III. LITERATURE SURVEY

While cyberbullying is a well-studied social problem, it has only lately attracted the attention of computer scientists, particularly in terms of autonomous detection tasks. Yin et al. [1] used supervised learning to detect harassment by training an SVM classifier using a bag of words model based on content, sentiment, and contextual aspects of texts. The authors employed a hybrid model based on sentiment and contextual data to achieve a recall level of 61.9% using a support vector machine learner.

The Formspring data was utilized in the study [2], Using Machine Learning to Identify Cyberbullying, by Kelly Reynolds, April Kontostathis, and Lynne Edwards. Data has been preprocessed by deleting redundant and irrelevant terms. As characteristics, the number of bad words and the density of bad words are employed. To detect cyberbullying, a supervised machine-learning technique was applied.

Dinakar et al. [3] at MIT used several binary and multiclass classifiers on a manually labeled corpus of YouTube comments. This method was 66.7% accurate. In addition, the authors employed an SVM learner in this situation.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong [4] proposed a suitable method integrating responsibilities characteristic of cyberbullying, content-based, and user-based. The findings revealed that using all features together resulted in higher performance. As a result, taking into account user characteristics such as age and gender, as well as the substance of the post, will increase the accuracy of cyberbullying detection.

In Unsupervised Cyberbullying Detection in Social Networks [5], Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino extracted a hybrid collection of features based on both traditional textual characteristics and the so-called "social features" as well. By using an unsupervised approach, they offered a novel method for identifying cyberbullying; nevertheless, they applied the classifiers inconsistently throughout their dataset.

The effectiveness of SVM and Neural Network classifiers is examined in the paper Social Media Cyberbullying Detection using Machine Learning [6] using both TFIDF and sentiment analysis feature extraction approaches. Moreover, trials with several n-gram language models were conducted. While assessing the model created by the classifiers, consideration was given to the 2-gram, 3-gram, and 4-gram sizes.

In the paper [7], the authors used natural language processing and machine learning to create and construct an efficient method for identifying online abusive and bullying texts. The accuracy level of four different machine learning methods is examined using two separate characteristics, namely Bag-of-Words (BoW) and term frequency-inverse text frequency (TFIDF).

IV. PROPOSED SYSTEM

The proposed system contains a Sender and a Receiver who interact through the chat application which processes abusive messages through a special system that we developed, which is built with many ML algorithms fed to the dataset of abusive words. We did feature extraction by preprocessing the data through these blogs. We used NLP for this purpose. Now the system shows an alert message if any abusive language is detected.

1. **NATURAL LANGUAGE PROCESSING (NLP)** NLP is a subfield of artificial intelligence that deals with the interactions between computers and human languages. NLP models are designed to analyze, understand, and generate human language, which enables a wide range of real-world applications like chatbots, virtual assistants, and language translation tools. NLP techniques are also used in text analysis and sentiment analysis to identify the underlying emotions and sentiments behind the words.
2. NLP techniques can be used for various preprocessing tasks to prepare text data for analysis. Tokenization, for example, is a common NLP technique used to split text data into individual words or phrases. Stemming and lemmatization are other techniques used to normalize words by reducing them to their base form, which helps to group similar words together. Additionally, stop words (common words that are unlikely to be relevant to the analysis) can be removed using NLP techniques.
3. **MACHINE LEARNING** (ML) Machine learning is a subset of artificial intelligence that enables machines to learn from data, without being explicitly programmed. ML algorithms are designed to find patterns and relationships in data, which can then be used to make predictions or take actions.
4. In the context of bullying detection, ML models can be trained on labeled data (data that has been manually annotated as abusive or non-abusive) to learn patterns that indicate abusive behavior. Once trained, these models can then be used to classify new messages as abusive or non-abusive.

PROPOSED ARCHITECTURE

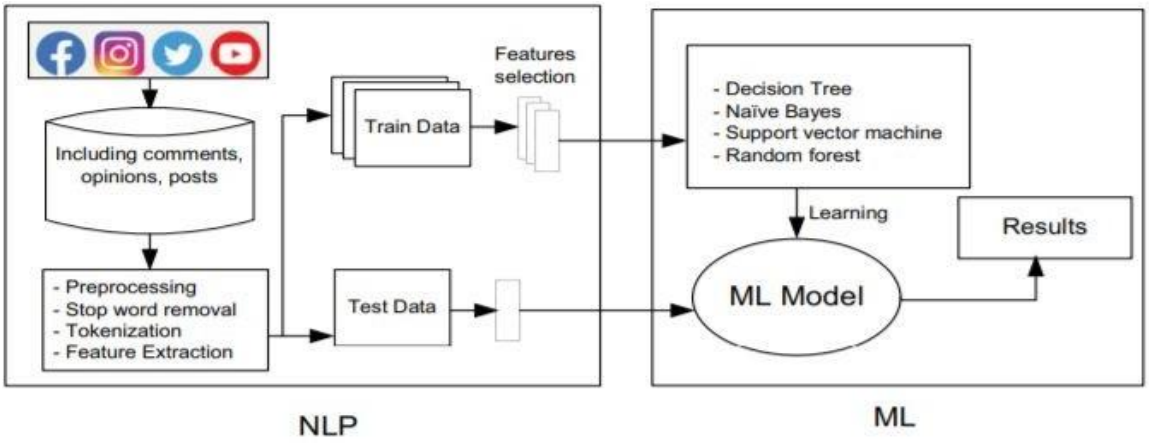

Fig: proposed architecture for real time cyberbullying detection

V. METHODOLOGY

1. We collected a dataset from diverse sources that included examples of both cyberbullying and non-bullying content.
2. The dataset was preprocessed using NLP techniques such as lemmatization, stemming, and tokenization.
3. We extracted features from the preprocessed dataset using the term frequency-inverse document frequency (TF-IDF) method.
4. The dataset was trained using several ML techniques, including a decision tree, random forest, support vector machine (SVM), and naive Bayes.
5. We analyzed the performance of each model and deployed the most accurate one.

- **ALGORITHMS:**

The referenced research publications took a variety of techniques. The data sources also varied widely and were obtained from social media platforms such as Twitter, Youtube, and Formspring. There are two parts/phases to the technique. The first is known as NLP (Natural Language Processing), while the second is known as ML (Machine Learning) (Machine learning). In the initial step, datasets including bullying texts, messages, or posts are gathered and processed for use by machine learning algorithms using natural language processing.TF-IDF is extracted feature. Decision Tree (DT), Random Forest, Support Vector Machine, and Naive Bayes were the machine learning algorithms employed. With two separate datasets, it was discovered that SVM outperformed other techniques. We discovered that, despite varied methodologies, most studies reported that Support Vector Machine (SVM) was the most successful.

VI. DETAILED METHODOLOGICAL APPROACH

Data Collection and Preprocessing: We collected a large dataset of online communication containing instances of cyberbullying using several sources, including social media platforms such as Kaggle, Twitter, and YouTube. The dataset consisted of a total of 33,457 rows and 2 columns. The dataset was preprocessed by removing any duplicate or irrelevant data and then cleaned by performing tokenization, lemmatization, stop-word removal, and other text normalization techniques to ensure consistency and accuracy in the data.

1. Tokenization: Break the text into individual words or tokens. This can be done using libraries such as NLTK or spaCy.

2. Stop word removal: Remove common words that do not add much meaning to the text, such as "the," "and," and "is." This can be done using libraries such as NLTK or spacy.

3. Punctuation removal: Remove any punctuation marks from the text, such as periods, commas, and exclamation marks. This can be done using regular expressions or Python's built-in string functions.

4. Lemmatization or stemming: Convert words to their base form, such as converting "running" to "run." This can be done using libraries such as NLTK or spaCy.

Feature Extraction: We extracted several features from the preprocessed dataset, including unigrams, bigrams, and trigrams, as well as TF-IDF scores for each feature. These features were selected based on their ability to capture the characteristics and dynamics of cyberbullying behavior in online communication.

1. N-grams: N-grams are contiguous sequences of N words from a text. In our implementation, we extracted unigrams, bigrams, and trigrams, which are sequences of 1, 2, and 3 words respectively. This helps to capture both the context and structure of the text.

2. TF-IDF: This is a statistical measure that reflects the importance of a word in a text corpus, taking into account the frequency of the word in the text and the frequency of the word in the entire corpus. We used the Tf Idf Vectorizer module from the sci-kit-learn library in Python to extract TF-IDF scores for each n-gram feature.

Model Selection and Training: We trained several machine learning models using the extracted features, including Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF). These models were selected based on their proven effectiveness in text classification tasks, including cyberbullying detection. We used a stratified k-fold cross-validation approach to evaluate the performance of each model and selected the best-performing model based on its overall accuracy, precision, recall, and F1-score.

1. Support Vector Machines (SVM): SVM is a popular machine learning algorithm used for classification tasks. It works by finding the hyperplane that maximally separates the data into different classes. In our implementation, we used the Radial Basis Function (RBF) kernel, which is a popular kernel function used with SVMs.

2. Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm that works by calculating the probability of each class given the input features. It assumes that the features are independent of each other, which is often not true in practice, hence the name "naive". Despite this limitation, Naive Bayes can be a very effective algorithm for certain types of text classification tasks.

3. Random Forest: Random Forest is an ensemble machine learning algorithm that works by creating a large number of decision trees and aggregating their results. It can be effective for text classification tasks because it is able to capture non-linear relationships between features.

VII. REAL-WORLD APPLICATION

The proposed system has real-world applications in online chat rooms, social media platforms, and other digital communication channels. It can be used to prevent and detect instances of cyberbullying and promote safer online communication. The system can also be extended to detect other forms of online harassment and hate speech.

VIII. CONCLUSION

In conclusion, this paper presents a comprehensive review of the existing research on cyberbullying detection using machine learning and natural language processing techniques. The study highlights the significance of the problem of cyberbullying and the need for effective detection methods to prevent its harmful consequences.

Through an analysis of various approaches and techniques, this paper shows that machine learning and natural language processing can be used to accurately detect and classify instances of cyberbullying. In particular, the study focuses on the use of tf-idf, which has been shown to achieve high levels of accuracy and precision in cyberbullying detection models.

The findings of this study have important implications for the development of effective cyberbullying detection systems, as well as for the design of online communities and social media platforms that prioritize user safety and well-being. By identifying instances of cyberbullying in real time, these systems can help prevent the spread of harmful behavior and promote a more positive online environment.

Overall, this paper contributes to the ongoing research efforts toward developing effective techniques for cyberbullying detection and prevention. The study highlights the importance of continued research and development in this area to ensure the safety and well-being of online users.

IX. REFERENCES

[1] Yin, Dawei & Xue, Zhenzhen & Hong, Liangjie & Davison, Brian & Edwards, April & Edwards, Lynne. (2009). Detection of harassment on Web 2.0.

[2] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning to Detect Cyberbullying," 2011 10th International Conference on Machine Learning and Applications and Workshops, 2011, pp. 241-244, doi: 10.1109/ICMLA.2011.152.

[3] Dinakar, Karthik, Roi Reichart and Henry Lieberman. "Modeling the Detection of Textual Cyberbullying." The Social Mobile Web (2011).

[4] Dadvar, Maral & Trieschnigg, Dolf & Ordelman, Roeland & de Jong, Franciska. (2013). Improving Cyberbullying Detection with User Context. In Proceedings of 35th European Conference on IR Research, ECIR 2013, Advances in Information Retrieval. pp 693-696. 10.1007/978-3-642-36973-5_62.

[5] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyberbullying detection in social networks," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 432-437, doi: 10.1109/ICPR.2016.7899672 (Publisher IEEE).

[6] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning" International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100587

[7] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411601.