A SENTIMENT AND EMOTION ANALYSIS APPROACH FOR DETECTING CYBERBULLYING ON SOCIAL MEDIA

[1]JALA SHILPA, [2] M.VYSHNAVI

[1]Assistant Professor &HOD, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda – 506003.Email-id: jala.shilpa2@gmail.com.

[2]Research Scholar, H.no: 24UC1D5806, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda – 506003,Email-id: mattewadavyshnavi@gmail.com.

ABSTRACT

The rapid rise of social media platforms has transformed how young people communicate and interact, but it has also given rise to new challenges such as cyberbullying — the use of digital technologies to harass, threaten, or humiliate others. Cyberbullying has severe psychological and emotional effects, particularly among teenagers, leading to anxiety, depression, and in extreme cases, suicidal tendencies. This work proposes a machine learning-based cyberbullying detection system that leverages natural language processing (NLP) techniques to automatically identify and classify bullying-related content on social networks. By analyzing text patterns, sentiment, and linguistic features, the system can distinguish between normal conversations and harmful or abusive messages. Various supervised learning algorithms such as Support Vector Machines (SVM), Random Forest, and Logistic Regression are employed to train models on labeled datasets containing bullying and non-bullying content. The proposed approach aims not only to detect abusive language but also to consider contextual dependencies to identify ongoing harassment or targeted attacks. Experimental results demonstrate that the integration of NLP with supervised machine learning significantly enhances the accuracy of cyberbullying detection, offering a valuable tool for safer and more responsible use of online social platforms.

Index Terms — Cyberbullying Detection, Machine Learning, Natural Language Processing (NLP), Supervised Learning, Social Media Analysis, Text Classification, Sentiment Analysis, Online Harassment, Support Vector Machine (SVM), Random Forest, Logistic Regression.

1.INTRODUCTION

Modern young people ("digital natives") have grown up in an era dominated by new technologies where communication occurs almost in real time and poses no limits on establishing relationships with other people or communities (Prensky, 2001). The fast-growing use of social networking sites among teenagers has made them increasingly vulnerable to exposure to bullying. Comments containing abusive or offensive words negatively affect the psychology of teens and lead to demoralization (Patchin & Hinduja, 2010).In this work, we devise methods to detect cyberbullying using supervised learning techniques. Cyberbullying is the use of digital technologies as a medium to harass, threaten, or humiliate someone (Smith et al., 2008). Although this issue has existed for many years, its severe impact on the mental and emotional well-being of young people has recently gained greater recognition (Tokunaga, 2010). Through machine learning (ML), it becomes possible to detect language patterns commonly used by bullies and their victims, and to develop rules that automatically identify cyberbullying-related content (Dadvar et al., 2013).

Social media platforms enable individuals to engage in social interaction, form new relationships, and maintain existing friendships. However, on the negative side, they also increase the risk of exposure to harmful online behavior, including grooming, sexually transgressive actions, signs of depression or suicidal ideation, and cyberbullying (Livingstone & Smith, 2014). Users are reachable 24/7 and can often remain anonymous, making social media a convenient channel for bullies to target their victims even outside traditional environments such as schools (Beran & Li, 2007).The detection of cyberbullying and online harassment is typically formulated as a text classification problem (Reynolds, Kontostathis & Edwards, 2011). Techniques commonly used for document classification, topic detection, and sentiment analysis can be effectively applied to detect bullying behavior based on the linguistic and semantic characteristics of messages, senders, and recipients (Dinakar et al., 2011). However, cyberbullying detection is inherently more complex than simply identifying abusive or offensive content, as additional context is often required to establish that an individual message forms part of a larger sequence of harassment

directed at a specific user (Zhao et al., 2016).

The growth of cyberbullying activities continues to rise in parallel with the rapid expansion of social networks. Such activities pose significant threats to the mental and physical health of victims, leading to anxiety, social isolation, and even self-harm (Kowalski et al., 2014). Although several studies have explored automatic detection of bullying behavior, the implementation of real-time monitoring systems for social networks remains limited. Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using natural language processing (NLP) and supervised learning techniques.

2.LITERATURE SURVEY

[1] M. Di Capua et al. proposed an unsupervised approach to develop an online bullying detection model that integrates both traditional textual features and social features. The features were categorized into four main groups — syntactic, semantic, sentiment, and community features. The authors utilized the Growing Hierarchical Self-Organizing Map (GHSOM) network with a 50×50 neuron grid and 20 elements in the insertion layer. The model also incorporated a K-means clustering algorithm to segment the input database and integrate it with GHSOM on the Formspring dataset. Experimental results showed that this hybrid unsupervised method outperformed previous models. However, the model achieved lower accuracy when tested on the YouTube dataset, as syntactic and textual features behaved differently across platforms. When applied to the Twitter dataset, the model exhibited reduced memory efficiency and a lower F1-score. Despite these limitations, the proposed approach showed promise for future development of cyberbullying mitigation applications.

[2] J. Yadav et al. introduced a BERT-based deep learning model for detecting cyberbullying on social media platforms. The model was tested on the Formspring and Wikipedia datasets. Results demonstrated 98% accuracy on Formspring data and 96% accuracy on Wikipedia data, outperforming previously used models. The improved results on Wikipedia were attributed to its larger dataset size, which minimized the need for oversampling, while Formspring required additional sampling. The study highlighted BERT's superior contextual understanding in identifying offensive and bullying language online.

[3] R. R. Dalvi et al. proposed a method to detect and prevent online exploitation on Twitter using supervised machine learning algorithms. Data were collected through the live Twitter API, and the authors applied both Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers. Feature extraction was performed using the TF-IDF vectorizer. The SVM model achieved an accuracy of 71.25%, outperforming Naïve Bayes, which achieved 52.75%. This demonstrated SVM's superior capability in identifying bullying-related content from social media posts.

[4] Trana R. E. et al. aimed to design a machine learning model to detect offensive or bullying text extracted from image memes. The dataset included approximately 19,000 text samples collected from YouTube. The study compared three algorithms — Naïve Bayes, Support Vector Machine (SVM), and Convolutional Neural Network (CNN) — across multiple bullying-related categories such as race, nationality, politics, and gender. Naïve Bayes outperformed SVM and CNN in most categories, except for gender-based bullying, where SVM performed better. The study concluded that incorporating contextual image-text information could enhance accuracy in detecting aggression-related content in multimedia posts.

[5] N. Tsapatsoulis et al. provided a comprehensive review on cyberbullying detection on Twitter. The study emphasized the importance of identifying various forms of abusive behavior and proposed a detailed framework for developing efficient Internet abuse detection systems. It outlined the data classification methods, platform-specific challenges, feature extraction techniques, and machine learning models used in previous studies. This review serves as a foundational step toward the creation of effective real-time cyberbullying detection tools using machine learning.

[6] G. A. León-Paredes et al. developed a cyberbullying detection model using Natural Language Processing (NLP) and Machine Learning (ML) techniques. Their system, named Spanish Cyberbullying Prevention (SPC), was trained using algorithms such as Naïve Bayes, Support Vector Machine, and Logistic Regression on Twitter data. The model achieved an impressive 93% accuracy, with cyberbullying detection rates ranging between 80–91%. The authors highlighted that stemming and lemmatization techniques in NLP significantly improved performance, and

that the model could be extended to English and regional languages.

[7] P. K. Roy et al. presented a hate speech detection framework on Twitter using deep neural networks. Traditional ML algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-Nearest Neighbors (KNN) were compared. Features were extracted using TF-IDF. Among traditional methods, SVM performed best, achieving 53% accuracy in detecting hate speech tweets. To improve results, deep learning models such as CNN, LSTM, and CLSTM were implemented, showing improved recall (0.88 for hate speech and 0.99 for non-hate speech). The study confirmed that k-fold cross-validation offers better robustness when dealing with imbalanced datasets.

[8] S. M. Kargutkar et al. proposed a dual-character classification model for cyberbullying detection using Convolutional Neural Networks (CNN) implemented through Keras. The system analyzed textual data to identify bullying characteristics and achieved promising results in recognizing offensive or aggressive communication patterns. This approach demonstrated the potential of deep learning for real-time cyberbullying detection in online environments.

3.EXISTING SYSTEM:

Social Networks give us great opportunities to communicate, and also increase the vulnerability of young people to threatening situations on the Internet. Cyberbullying on social media is a global phenomenon due to its large number of active users. The trend shows that social network cyberbullying is increasing rapidly day by day. Recent research shows that cyberbullying is a growing problem among young people. Successful prevention depends on the proper detection of potentially harmful messages, and the information overload of the Internet requires intelligent systems to automatically detect potential hazards. Therefore, in this project, we will focus on creating a model to automatically detect cyberbullying in social media text by simulating messages created by social media bullying.

DISADVANTAGES

Risk of misclassifying non-bullying messages as cyberbullying due to language complexity and context.

Possibility of failing to detect actual instances of cyberbullying, resulting in false negatives.

4.PROPOSED SYSTEM:

Cyberbullying detection is solved in this project as a binary classification problem where we are

detecting two majors form of Cyberbullying: hate speech on Twitter and Personal attacks on

Wikipedia and classifying them as containing Cyberbullying or not.

ADVANTAGES

The CBOW (Continuous Bag of Words) model can take one or multiple words as input and predict a target word based on context.

CBOW averages the context of input words to predict a word, allowing it to capture multiple meanings for a single word.

5.SYSTEM MODEL

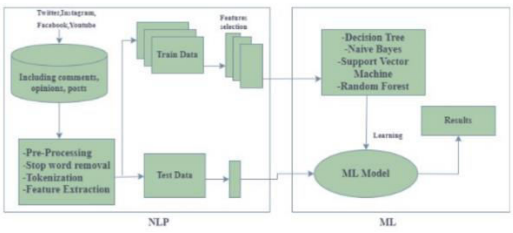


Fig.System Model

6.IMPLEMENTATION

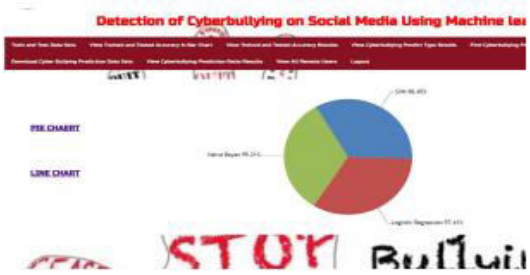
- **Import Libraries:** Import necessary libraries, including scikit-learn, for implementing the SVM model.
- **Load Dataset:** Load a sample dataset from scikit-learn's built-in datasets module.
- **Split Dataset:** Divide the dataset into training and testing sets using train_test_split from scikit-learn.
- **SVM Classifier:** Create an instance of the SVM classifier using the SVC class with a linear kernel.
- **Train the Model:** Train the SVM model on the training data using the fit() method.
- **Make Predictions:** Use the trained model to make predictions on the testing data with the predict() method.

7.SCREENSHOTS

User Login

User Register

PieChart



Predict Result

8.CONCLUSION

As the users of social media is increasing day by day along with-it cyber bullying related cases is also increasing on social media with its growing popularity and the increasing usage of social media by young people. It is necessary to devise an automated method of detecting cyberbullying in order to avoid the harmful effects of cyberbullying before it's too late. As sometimes the consequences of cyberbullying can be as bad as the suicide by the person that is bullied. So, keeping in mind the importance of a system which can detect the cyberbullying and online harassment, we are going to study different ML algorithms and their effectiveness in comparison with each other to predict their accuracy on a given data set to find out the best among them. After studying all the five algorithms and their results we come to a conclusion that SVM model perform best in detecting cyberbullying with an accuracy of 0.990145 and along with the second-best performing algorithm comes out to be naïve bayes algorithm with an accuracy of 0.986897. So, we can use any of these two algorithms to detect the cyberbullying and online harassment to get the highest accuracy while linear regression is the least accurate among all of them.

9.FUTURE ENHANCEMENTS

In the future, enhancing the cyberbullying detection system could involve the adoption of advanced NLP techniques like deep learning models to better capture nuanced linguistic patterns. Additionally, integrating semantic analysis and multimodal analysis to understand context and analyze diverse types of content could improve detection accuracy. Dynamic learning frameworks, user behavior analysis, and contextual awareness can further refine the system's adaptability and precision. Real-time monitoring, privacy-preserving techniques, and collaboration with social media platforms would enhance proactive intervention while ensuring user privacy and compliance. Lastly, rigorous evaluation and benchmarking against diverse datasets would ensure the system's effectiveness and generalization. These enhancements aim to create a more robust and proactive system for combating cyberbullying on social media platforms, fostering safer online environments.

10.REFERENCES

[1] PricewaterhouseCoopers LLP, 2016.

[2] W. Hochfeld, J. Riffell, N. Levinson, “Four trends that will transform healthcare in Europe in 2016,” European Pharmaceutical Review, vol. 21, no. 1, 2016.

[3] A. S. Mosa, I. Yoo, L. Sheets, “A systematic review of healthcare applications for smartphone,” BMC Med. Inform. Decis. Mak., vol. 12, p. 67, 2012.

[4] L. Bellina, E. Missoni, “Mobile cell-phones (M-phones) in telemicroscopy: Increasing connectivity of isolated laboratories,” Diagn. Pathol., vol. 4, p. 19, 2009.

[5] L. Dayer, S. Heldenbrand, P. Anderson, P. O. Gubbins, B. C. Martin, “Smartphone medication adherence apps: Potential benefits to patients and providers,” J. Am. Pharm. Assoc., vol. 53, pp. 172, 2013.

[6] N. Tripp, K. Hainey, A. Liu, A. Poulton, M. Peek, J. Kim, R. Nanan, “An emerging model of maternity care: Smartphone, midwife, doctor?,” Women Birth, vol. 27, pp. 64–67, 2014.

[7] A. P. Demidowich, K. Lu, R. Tamler, Z. Bloomgarden, “An evaluation of diabetes self-management applications for Android smartphones,” J. Telemed. Telecare, vol. 18, pp. 235–238, 2012.

[8] A. Rao, P. Hou, T. Golnik, J. Flaherty, S. Vu, “Evolution of data management tools

for managing self-monitoring of blood glucose results: A survey of iPhone applications," J. Diabetes Sci. Technol., vol. 4, pp. 949–957, 2010.

[9] S. Wallace, M. Clark, J. White, "'It's on my iPhone': Attitudes to the use of mobile computing devices in medical education, a mixed-methods study," BMJ Open, vol. 2, p. e001099, 2012.

[10] K. E. Muessig, E. C. Pike, S. Legrand, L. B. Hightow-Weidman, "Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: A review," J. Med. Internet Res., vol. 15, p. e1, 2013.