International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

[1]NAYAKAM HRUDAYA,[2]SHAIK MOHAMMAD TANEEM,[3]G M KAUSHIK,[4]PALLE GANESH,[5]DR.K.PHALGUNA RAO

[1,2,3,4]Students, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100
[5]Professor, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100

ABSTRACT

The increasing use of the internet and the widespread access to online communities, such as social media platforms, have contributed to the rise of cybercrime. Cyberbullying, a relatively new form of bullying that emerged alongside the growth of social networks, involves sending harmful messages or verbally abusing individuals in front of an online audience. The unique features of online social networks allow cyberbullies to reach places and communities that were previously inaccessible. This project aims to identify cyberbullying on Twitter using Support Vector Machines (SVM). The objectives of this implementation are outlined in the objective section. Additionally, we will use Optical Character Recognition (OCR) to detect image-based cyberbullying, examining its impact on individuals through a dummy system. We will employ machine learning and natural language processing (NLP) techniques to identify the characteristics of cyberbullying and automatically detect such behavior by matching textual data to these traits. Based on an extensive literature review, we classify existing approaches into four categories: supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning methods, such as SVM and Naïve Bayes, are commonly used to build predictive models for cyberbullying detection. In this project, we will apply various machine learning techniques, including Bayesian logistic regression, random forest, and SVM, to identify cyberbullying.

Keywords: Machine Learning, Cyberbullying, Social Media, Twitter.

INTRODUCTION

Modern youth, often referred to as "digital natives," have grown up in an era where new technologies dominate communication, enabling real-time interactions and facilitating relationships across vast networks. The rapid rise of social networking sites among teenagers has unfortunately made them more vulnerable to bullying. Abusive comments, often directed at teens, can severely impact their mental health, demoralizing them and affecting their psychological well-being. In this work, we propose methods to detect cyberbullying using supervised learning techniques. Cyberbullying refers to the use of digital platforms to harass, intimidate, or harm others. While it has been a concern for years, its recognition as a major issue, particularly among young people, has gained more attention in recent times. By applying machine learning, we can identify language patterns used by bullies and victims, creating automated systems to detect cyberbullying content.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Social media platforms serve as spaces for people to engage in social interaction, form new relationships, and maintain existing friendships. However, on the downside, these platforms also increase the risks for young users, exposing them to harmful situations such as grooming, sexually inappropriate behavior, depression, suicidal thoughts, and cyberbullying. Social media provides 24/7 access and often allows anonymity, making it a convenient medium for bullies to target their victims beyond the school environment.

Detecting cyberbullying and online harassment is commonly approached as a classification problem. Techniques used for document classification, topic detection, and sentiment analysis can be applied to identify electronic bullying by analyzing message content, sender characteristics, and recipient profiles. However, cyberbullying detection presents unique challenges, as it requires more than just identifying abusive content. Context is crucial to determine whether a single abusive message is part of a broader pattern of online harassment aimed at an individual.

The rise of cyberbullying parallels the growth of social networks, and it poses a significant threat to the mental and physical health of its victims. While there are existing projects aimed at detecting bullying, fewer focus on monitoring social networks for such activities. Therefore, the proposed system aims to identify and detect cyberbullying activities in social networks using natural language processing techniques.

LITERATURE SURVEY

M. Di Capua et al. [1] explored an unsupervised approach to developing an online bullying model using a combination of traditional textual features and "social features." These features were categorized into four groups: Syntactic features, Semantic features, Sentiment features, and Community features. The authors used the Growing Hierarchical Self Map (GHSOM) network with a 50x50 grid of neurons and 20 elements as the insertion layer. To enhance the clustering of data, the k-means algorithm was integrated with GHSOM. This hybrid approach outperformed previous models in tests with the Formspring database. However, when applied to the YouTube database, the model showed lower accuracy due to differences in text analysis and syntactical features between the two platforms. Additionally, the method performed weakly with Twitter data in terms of memory and F1 Score. Despite these challenges, the proposed model could be further improved for developing applications aimed at mitigating cyberbullying.

J. Yadav et al. [2] proposed a novel approach for detecting internet cyberbullying on social media using the BERT model with a single-layer neural network. The model was tested on the Formspring forum and Wikipedia databases. The results demonstrated 98% accuracy on the Formspring database and 96% on the larger Wikipedia database. The BERT model performed better on the Wikipedia database, likely due to its size and the absence of excessive sampling requirements, while the Formspring data required multiple sampling attempts.

R. R. Dalvi et al. [3] introduced a method to detect and prevent online exploitation on Twitter using supervised machine learning algorithms. The study utilized the live Twitter API to collect tweet data. Both Support Vector Machine (SVM) and Naive Bayes were tested on the collected datasets, with the TFIDF vectorizer used to remove irrelevant features. The results showed that the SVM model outperformed Naive Bayes by achieving an accuracy of 71.25%, compared to Naive Bayes' 52.75%.

Trana R.E. et al. [4] aimed to design a machine learning model to detect special events, including text extracted from image memes. The study used a dataset of approximately 19,000 text views published on YouTube. Three machine learning models were explored: Naive Bayes, Support Vector Machine (SVM), and Convolutional Neural Networks (CNN). The results indicated that Naive Bayes outperformed both SVM and CNN in categories such as race, nationality, politics, and general content. While SVM performed better in some categories, the study highlighted the need for further research to build a more effective system for detecting cyberbullying in image-based content on platforms like YouTube.

N. Tsapatsoulis et al. [5] provided a detailed review of cyberbullying detection on Twitter. The paper discussed the importance of identifying various abusers on the platform and the practical steps required to develop an effective application for internet traffic detection. It covered data classification methods, machine learning models, feature types, and model evaluations. This paper serves as an important foundational reference for projects focused on developing cyberbullying detection technologies using machine learning.

G. A. León-Paredes et al. [6] described the development of an online bullying detection model using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The Spanish Cyberbullying Prevention (SPC) system was built using machine learning strategies like Naïve Bayes, Support Vector Machine, and Logistic Regression. The dataset used in this study was sourced from Twitter, and the model achieved an accuracy of 93%. This work highlights the potential of combining NLP and ML for effective cyberbullying detection in social media platforms.

PROPOSED METHODOLOGY

This project will be developed using Python and web technologies. The process begins by collecting and loading the dataset. Once the dataset is loaded, we will preprocess the data, followed by applying the Term Frequency-Inverse Document Frequency (TF-IDF) technique. We will then train the dataset using three different algorithms: Naive Bayes, Support Vector Machine (SVM), and Deep Neural Networks (DNN). These models will be trained separately to evaluate their individual performances.

Next, we will develop a web application using the FLASK framework. The application will fetch live tweets from Twitter, apply the trained models, and analyze whether the content (text or images) involves cyberbullying. The backend will be powered by Python, with MySQL as the database. The frontend will be developed using HTML, CSS, and JavaScript.

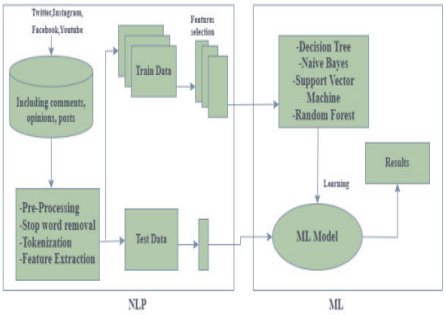
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Cyberbullying detection and online harassment are often approached as classification problems. Techniques typically used for document classification, topic detection, and sentiment analysis will be leveraged to detect bullying behaviors based on message characteristics, sender, and recipient traits. However, detecting cyberbullying is inherently more challenging than merely identifying offensive content. Additional context is often necessary to confirm that a single offensive message is part of a series of targeted online harassment.

Cyberbullying and social media usage are both on the rise, which has heightened the urgency of developing automated solutions for detection. We have categorized the features extracted from the dataset into five groups:

- **Sentimental Features**
- **Sarcastic Features**
- **Syntactic Features**
- **Semantic Features**
- **Social Features**

While several projects aim to detect bullying, there are few that focus on social media monitoring to detect cyberbullying. The proposed system aims to fill this gap by using natural language processing (NLP) techniques. However, a significant challenge remains in acquiring a sufficiently large and diverse training dataset. Existing datasets often fail to capture the vast volume of messages being sent daily, with many reports of bullying being a small fraction of the messages circulating on social media. Random sampling may only yield a limited number of aggressive messages, making the task of gathering enough training data a major obstacle.



CONCLUSION

In this work, we proposed a semi-supervised approach for detecting cyberbullying using five distinct features that define a cyberbullying post or message, with the BERT model as the core. Focusing solely on sentimental features, the BERT model achieved an impressive accuracy of 91.90% after training for two cycles, outperforming traditional machine learning models. The accuracy could be further improved with a larger dataset, and we anticipate even better results when all five features proposed in this study are considered. By integrating these features, we envision the development of an application capable of detecting and reporting cyberbullying posts effectively. Future work could explore combining other models with BERT to create a more robust detection system tailored specifically to NLP tasks related to cyberbullying detection.

REFERENCES

[1] M. Di Capua, E. Di Nardo, and A. Petrosino, “Unsupervised Cyberbullying Detection in Social Networks,” Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 432–437, 2016. doi: 10.1109/ICPR.2016.7899672

[2] D. Poeter, "Study: A Quarter of Parents Say Their Child Involved in Cyberbullying," PCMag, 2011. [Online]. Available: http://www.pcmag.com/article2/0,2817,2388540,00.asp

[3] J. W. Patchin and S. Hinduja, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," Youth Violence and Juvenile Justice, vol. 4, no. 2, pp. 148–169, 2006.

[4] Anti-Defamation League, "Glossary of Cyberbullying Terms," ADL.org, 2011. [Online]. Available: http://www.adl.org/education/curriculum_connections/cyberbullying/glossary.pdf

[5] N. E. Willard, Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress, Research Press, 2007.

[6] D. Maher, "Cyberbullying: An Ethnographic Case Study of One Australian Upper Primary School Class," Youth Studies Australia, vol. 27, no. 4, pp. 50–57, 2008.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in Proceedings of the Content Analysis of Web 2.0 Workshop (CAW 2.0), Madrid, Spain, 2009.

[8] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11), Barcelona, Spain, 2011.

[9] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., San Francisco, CA: Morgan Kaufmann, 2005.

[10] R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1993.

[11] W. W. Cohen, "Fast Effective Rule Induction," in Proceedings of the Twelfth International Conference on Machine Learning (ICML'95), Tahoe City, CA, 1995, pp. 115–123.

[12] D. W. Aha and D. Kibler, "Instance-Based Learning Algorithms," Machine Learning, vol. 6, pp. 37–66, 1991.

[13] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in Advances in Kernel Methods, pp. 185–208, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?id=299094.299105