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

Automated Cyberbullying Detection Leveraging Data Science and Machine Learning Techniques

[1]Vishwa K. Patel, [2] Gaurav D. Tivari, [3]Paras Narkhede, [4]Dr. Satvik Khara

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Head of Department
[1]Department of Computer Engineering
[1]Silver Oak College of Engineering & Technology, Silver Oak University, Ahmedabad, Gujarat

Abstract : Social media has become a dominant platform in modern culture but has also seen a rise in issues like cyberbullying and online harassment, impacting people's mental health. Researchers are increasingly focusing on identifying indicators of online abuse, aiming to develop a system that can detect such instances using Natural Language Processing and Random Forest regression. The COVID-19 pandemic has exacerbated cyberbullying, especially among young people, due to changes in social interactions online. Proposed solutions include data cleansing, text mining with lemmatization, feature extraction with VADER emotion analysis, and data classification using Naive Bayes and VADER sentiment for word vector generation.

Key Words: Cyberbullying Detection, Machine learning algorithms, social media, Twitter, Feature engineering, Rumour detection, natural language processing

INTRODUCTION:

Social media can be both beneficial and detrimental. While it is fantastic for keeping in touch with friends and family, it also has a dark side with issues like rude comments and online bullying, particularly affecting young people. This report delves into how these negative online experiences impact mental health. We have developed a clever system using computers to identify and block those harmful messages. This initiative became even more crucial during the COVID-19 pandemic when online bullying escalated as people spent more time at home and online. The rapid spread of COVID-19 and subsequent shift in social norms have led to an increase in cyberbullying, especially among the youth. The rise in popularity of various online communication platforms has contributed to the growing number of reported cyberbullying cases. The pandemic-induced shift to remote work and online social interactions has not only facilitated social contact but also contributed to the ongoing digitization of bullying behaviors.

AIM OF THE RESEARCH:

The primary aim of the cyberbullying detection model is to enhance manual monitoring of cyberbullying on social networks. In this project, we will collect tweets from Twitter accounts, preprocess the tweets, and apply the generated model to detect instances of cyberbullying. The objectives of the system's development and implementation include:

1. Dataset Collection and Preprocessing: Collecting a comprehensive dataset of bullying-related words and phrases from various sources. This dataset will then be cleaned and preprocessed to ensure it is suitable for analysis.

2. Natural Language Processing (NLP): Utilizing NLP techniques to analyze the textual data. This involves tokenization, lemmatization, and sentiment analysis to understand the context and emotions conveyed in the tweets.

3. Machine Learning Model Development: Implementing and comparing various machine learning algorithms, such as Random Forest, Naive Bayes, and Support Vector Machines (SVM), to identify the most effective model for detecting cyberbullying.

4. Data Collection from Twitter: Fetching tweets from selected Twitter accounts using Twitter's API. This step involves setting up automated scripts to collect real-time data and ensuring the data is representative of different demographics and regions.

5. Tweet Preprocessing: Preprocessing the collected tweets to remove noise, such as special characters, links, and stop words. This step is crucial for improving the accuracy of the detection model.

6. Model Application and Evaluation: Applying the trained machine learning model to the preprocessed tweets to detect instances of cyberbullying. The results will be evaluated to measure the model's accuracy, precision, recall, and F1 score.

7. System Integration and Deployment: Integrating the detection model into a user-friendly system that can be used by social network moderators to monitor and address cyberbullying in real-time. This system will include features such as alert generation, report generation, and user feedback.

8. Continuous Improvement: Continuously updating the dataset and retraining the model to adapt to new trends and slang in cyberbullying. This ensures the system remains effective over time.

By achieving these objectives, this research aims to provide a robust solution for identifying and mitigating cyberbullying on social networks, ultimately contributing to a safer online environment for users.

PROBLEM STATEMENTS:

Lack of High-Quality Labeled Data: One of the major challenges in developing an effective cyberbullying detection system is the scarcity of high-quality labeled data. Labeled data is essential for supervised learning algorithms, which are commonly used in cyberbullying detection. Without sufficient labeled data, it becomes challenging to train accurate and reliable models.

Limited Language Coverage: One of the significant challenges in developing effective automatic cyberbullying detection systems is their limited language coverage. Most current detection systems are primarily trained on English language data, which restricts their applicability and effectiveness in non-English contexts.

Difficulty in Detecting Sarcasm and Irony One of the notable challenges in developing effective cyberbullying detection systems is the detection of sarcasm and irony in messages. Cyberbullies often employ these language nuances to disguise their intent, making it challenging for machine learning algorithms to accurately interpret their meaning. These complexities are not only difficult for machines but also pose challenges for human understanding.

While significant strides have been made in using data science and machine learning for automatic cyberbullying detection, challenges like detecting sarcasm and irony persist. Addressing these challenges requires interdisciplinary efforts and innovative approaches to develop more effective and reliable detection systems in the future.

.

LITERATURE REVIEW:

The explosion of social media in the 21st century has revolutionized communication, yet it has brought significant drawbacks. Cyberbullying and online abuse have become prevalent issues, causing harm to mental health and social harmony. In response, researchers are leveraging technology to combat these challenges. Techniques such as Natural Language Processing (NLP) and machine learning, including Random Forest and deep learning approaches, are proving effective in detecting cyberbullying. From simple methods like analyzing word lists to sophisticated neural networks, researchers continue to refine techniques for identifying harmful online behaviors. Emerging approaches that integrate analysis of text, images, and videos are showing promise in addressing the dynamic nature of cyberbullying. These efforts not only underscore the intricacies of online harassment but also emphasize the necessity for robust detection models to safeguard individuals' well-being in our increasingly digital world. The fight against cyberbullying is an ongoing endeavor, demanding constant adaptation to stay ahead of evolving tactics and strategies. Collaborative efforts across disciplines are essential to developing comprehensive solutions that protect users and promote a safer online environment.

METHODOLOGY:

This project was developed using Python and web technology. Initially, we searched for and found the dataset, then downloaded it to train the model. After downloading, we pre-processed the data and transferred it to Tf-Idf. Using Naive Bayes and SVM (Support Vector Machine), we trained the dataset and generated models separately. Subsequently, we developed a web-based application using the Streamlit framework. We fetched tweets from Twitter and applied the generated models to these fetched tweets to check if the text indicated cyberbullying or not.

1. **Requirements Gathering**: This phase involved gathering information about the problem of cyberbullying, the available data sources, and the desired outcomes of the detection system.

2. **Data Collection and Preprocessing:** In this phase, relevant data related to cyberbullying was collected from various sources such as Twitter, social media platforms, online forums, and messaging applications. The collected data was then preprocessed by removing noise, irrelevant information, and performing feature engineering.

3. **Model Selection and Training:** In this phase, appropriate machine learning algorithms such as Support Vector Machines (SVM), Random Forests, or Deep Learning models were selected and trained using the pre-processed data.

The models were evaluated using cross-validation techniques and hyperparameter tuning to achieve optimal performance.

4. **Model Integration and Testing:** In this phase, the trained models were integrated into a detection system that automatically detected cyberbullying in real-time. The system was tested using a dataset containing cyberbullying instances to assess its accuracy, precision, and recall.

5. **Deployment and Maintenance:** Once the detection system was deployed, it was continuously monitored to ensure its effectiveness in detecting cyberbullying instances. The system was also updated regularly to adapt to new trends and patterns of cyberbullying.

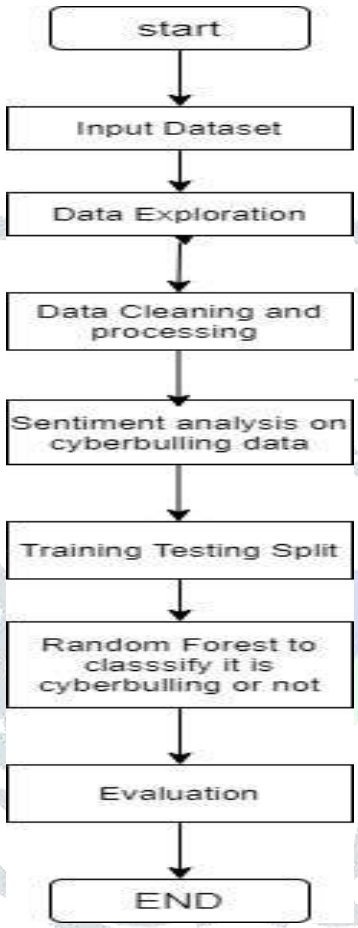


Fig 1: Flow Chart

RESULT:

The evaluation results of Multinomial Naive Bayes are given in Following figure:

Accuracy: 0.7594087430548275
Classification Report:

	precision	recall	f1-score	support
1	0.83	0.96	0.89	1566
2	0.80	0.96	0.87	1603
3	0.86	0.91	0.88	1603
4	0.76	0.80	0.78	1531
5	0.58	0.51	0.55	1612
6	0.64	0.42	0.51	1624
accuracy			0.76	9539
macro avg	0.75	0.76	0.75	9539
weighted avg	0.74	0.76	0.75	9539

Fig.2 – Result of Multinomial Naive Bayes

The evaluation results of SVM (Support Vector machine) are given in following figure.

SVM Accuracy: 0.815703952196247
SVM Classification Report:

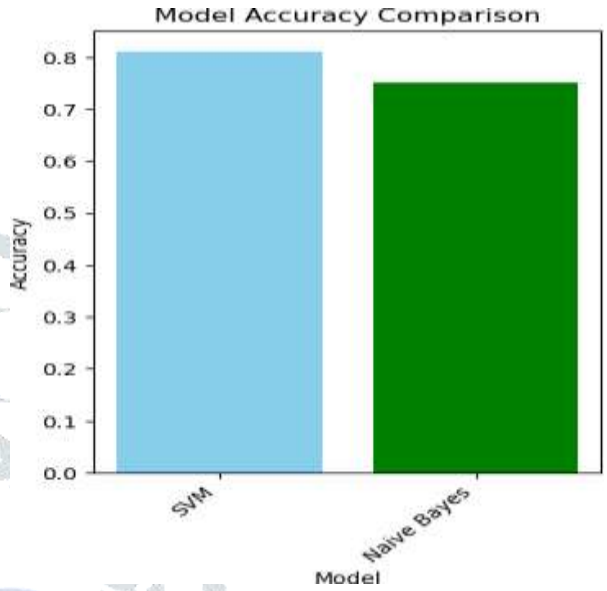
	precision	recall	f1-score	support
1	0.94	0.95	0.94	1566
2	0.95	0.97	0.96	1603
3	0.98	0.97	0.97	1603
4	0.87	0.83	0.85	1531
5	0.58	0.68	0.63	1612
6	0.59	0.52	0.55	1624
accuracy			0.82	9539
macro avg	0.82	0.82	0.82	9539
weighted avg	0.82	0.82	0.82	9539

Fig.3 – Result of Support Vector Machine (SVM)

Comparison Table

Algorithm Used	Support Vector Machine	Naive Bayes
Feature Extraction method	TextBlob	Vader sentiment
Accuracy	81 %	75%

Criteria	Naive Bayes	SVM
Complexity	Low	High
Speed	Faster	Slower
Implementation	Simple	Complex
Memory Usage	Low	High



FUTURE ENHANCEMENTS:

In the future, it is intended to improve the developed system by using a more accurate dataset to detect cyberbullying. We also plan to apply other machine learning algorithms to check the accuracy of the models. A higher accuracy model will help to detect bullying more accurately. Another interesting direction for future work would be the detection of fine-grained cyberbullying categories, such as threats, curses, and expressions of racism and hate. When applied in a cascaded model, the system could identify severe cases of cyberbullying with high precision, which would be particularly useful for monitoring purposes. Additionally, our dataset allows for the detection of participant roles typically involved in cyberbullying.

Multimodal Data Analysis: Cyberbullying involves different forms of data such as text, images, and videos. Developing models that analyze multiple types of data can enhance detection accuracy.

Incorporating Social Network Analysis: Cyberbullying often occurs on social media platforms, making it useful to analyze the social network to understand the relationship between the bully and the victim. Incorporating social network analysis techniques can enhance the detection of cyberbullying instances.

Active Learning: Developing models that learn from their mistakes and incorporate new labeled data can enhance the model's performance over time. Active learning is an approach that involves selecting the most informative data points to label and adding them to the training set to improve model accuracy.

CONCLUSION:

In conclusion, we conducted research on the identification of cyberbullying in tweets to identify cyberbullying language and actors. The Naive Bayes algorithm for classification and VADER Sentiment for feature extraction, along with suitable data preprocessing techniques, were used in this work. It successfully identified tweets engaging in cyberbullying. The proposed work focuses on detecting the occurrence of cyberbullying in Twitter networks using machine learning algorithms and type- and topic-specific classification. Additionally, the information of cyberbullies and rumor-spreading individuals will also be extracted. The integration of cyberbully detection and rumor detection in a single application simplifies the detection process. This proposed model may provide better results in preventing social network users from becoming victims compared to other existing techniques.

REFERENCES:

[1] Adnan, 2019. An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data , pp. 85- 91.
[2] Agarwal, 2015. Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. International Journal of Computer Applications , pp. 30-36.
[3] Agrawal, 2018. Deep learning for detecting cyberbullying across multiple social media platforms. s.l., Springer, pp. 141-153.
[4] Ajlan, 2018. Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications, pp. 1-9.
[5] Al-garadi, 2016. Cybercrime detection in online communications: The experimental case of cyberbully-lying detection in the twitter network. Computers in Human Behaviour, pp. 433-443.
[6] Anon, 2012. Stemming Algorithms: A Comparative Study and their Analysis. International Journal of Applied Information Systems, pp. 7-12.
[7] Breiman, 2001. RANDOM FORESTS, Berkeley, CA: Statistics Department University of California.
[8] Campbell, 2012. Online social networking and the experience of cyber- bullying. Studies in Health Technology and Informatics, pp. 212-217.
[9] Ghosh, 2017. Toward multimodal cyberbullying detection. s.l., s.n., pp. 2090-2099.
[10] Huang, 2018. Weakly supervised cyberbullying detection using co- trained ensembles of embedding models. s.l., IEEE, pp. 479-486.