IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyberbullying Detection And Prevention Using Machine Learning

Mohammed Ahmad Raza Khan, Tejas S, Vikram Patel, Tharun Kumar M H, Dr Suresha D

Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru 560056, Karnataka, India.

Abstract: The abstract presents an overview of the strategies for preventing and detecting cyberbullying. Prevention efforts encompass a multi-faceted approach, addressing both individual and systemic factors. Additionally, the integration of technology-based solutions, such as content filtering and social media monitoring tools, can help mitigate the risk of cyberbullying incidents. Detection of cyberbullying involves the timely identification and assessment of harmful online behaviors. Machine learning algorithms offer promising avenues for detecting cyberbullying content across various digital platforms. Natural language processing (NLP) algorithms can analyze text-based communications to identify patterns indicative of bullying behavior, while image and video analysis algorithms can detect visual content containing abusive or threatening material. However, the ethical implications and potential biases associated with algorithmic detection methods impose careful consideration and ongoing refinement. Cultivating a culture of digital civility and promoting positive online behaviors are evident for creating safer and more inclusive digital environments. Continued research, collaboration, and community engagement are necessary for advancing the field of cyberbullying prevention and detection and mitigating its adverse impacts on individuals and society.

Keywords: Cyberbullying, Machine Learning, NLP.

1 Introduction

[1] Cyber bullying is emerging as a serious social problem, especially among teenagers. Cyber bullying is defined as “the use of information technology to harm or harass other people in a deliberate, repeated, and hostile manner” . With the advent of social media networks such as Twitter and Facebook, it has become more prevalent. Beyond detection, machine learning holds promise in facilitating proactive prevention strategies by identifying risk factors, predicting potential instances of cyberbullying, and empowering users with timely interventions and support mechanisms. Through a synthesis of empirical research, case studies, and theoretical frameworks, this paper aims to elucidate the opportunities and challenges inherent in leverages machine learning for prevention and cyberbullying detection .

1.1 Existing System:

Text bullying is the major part where cyber bullying happens. The existing cyberbully detection tools focus solely on text-based content.

o Perspective API by Google: Perspective API uses machine learning models to score the perceived impact of online comments. It can identify toxic language, including content related to cyberbullying, harassment, and abuse.

o Jigsaw's Conversation AI: Jigsaw, a subsidiary of Alphabet Inc. (Google's parent company), developed Conversation AI, which utilizes machine learning algorithms to detect toxic language in online conversations. It aims to improve online discussions by automatically flagging and filtering out harmful content, including cyberbullying.

o Sift Ninja: Sift Ninja is a tool that uses machine learning to automatically moderate and filter out toxic comments on social media platforms and websites. It can detect variety of harmful content, including cyberbullying, hate speech, and harassment.

o IBM Watson Natural Language Classifier: IBM Watson offers a Natural Language Classifier service which could be trained to categorize text into custom classes, including cyberbullying. By training the classifier on labeled examples of cyberbullying content, it could be used to automatically detect and flag potential instances of cyberbullying in text-based content .

Drawbacks or Limitations with the existing system are described below:

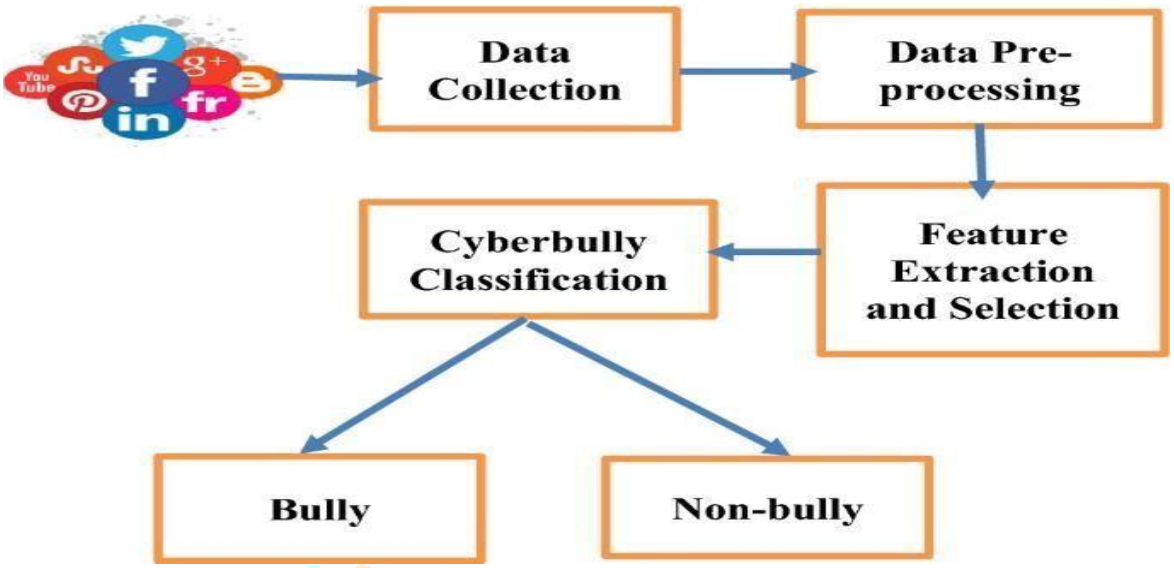
• Most models deals with text classification as a binary classification problem (Hate and Not hate), which is seen as a limitation.

• An attempt at multi-label classification got a low accuracy due to a huge amount of unbalanced data in the dataset.

• The current approaches use single classification technique. Majority of the approaches deal with feature extraction from the text using dictionaries and bag-of-words. It was observed that the word-order is ignored and causes misclassification as different words are used in different context.

1.2 Proposed System:

The proposed system is a desktop UI application where a user can input some textual content, and view the results of classification of the content if it is hate, non-hate or offensive and for this classification we will be building a model which employs N-gram and word embedding techniques of NLP and an ensemble machine learning approach by considering various single classification algorithms. such as WKNN, RVM and Naïve Bayes.



2 . Related Work

Literature survey**:**

- **In paper [2] the Advantages:** Machine Learning: Versatile, well-established techniques. Transfer Learning: Utilizes knowledge from one domain to enhance performancein another.

Limitations: - Machine Learning: Limited generalization to new data.

Transfer Learning: Dependency on the quality and relevance of pre-trained models.

Performance: Varied based on the specific techniques employed and datasets used**.**

- **In paper [3] the Advantages**: Emotion-based approaches capture nuanced contextual information. Canprovide insights into the psychological aspects of cyberbullying**.**

Limitations: - Challenges in accurately extracting and interpreting emotions from text. -Limited applicability to cases where emotion couldn't be the primary indicator of cyberbullying**.**

Performance: Effectiveness highly depends on the accuracy of emotion extraction methods.

- **In paper [4] the Advantages:** - Deep Learning: Captures complex patterns and relationships in data. - RNNs: Effective in handling sequential information, crucial for social media data.

Limitations: - - Computational complexity and resource-intensive training. – Requires large labeled datasets for optimal performance.

Performance: Performance influenced by the size and quality of training data and model hyperparameters.

- **In paper [5] the Advantages:** -Federated Learning: Maintains privacy by training models locally. - Word Embeddings: Captures semantic relationships in text. - Emotional Features: Provides additional contextual information.

Limitations: - Complexity in managing federated learning across diverse user bases. -Challenges in achieving

consistent and accurate emotional feature extraction.

Performance: Effectiveness influenced by the success of federated learning coordination and accuracy of emotional feature extraction.

- **In paper [6] the Advantages:** - - Comprehensive overview of automated cyberbullying detection methodologies. - Highlights the importance of temporal aspects in detection**.**

Limitations: - Lack of in-depth exploration of specific techniques. - May not providedetailed insights into the latest advancements in the field

Performance: Serves as a valuable resource for understanding the landscape of automated cyberbullying detection.

- **In paper [7] the Advantages:** - - - Ensemble Learning: Improved generalization by combining multiple models. - Glowworm Swarm Optimization: Efficient optimization algorithm.

Limitations: - - Increased complexity in managing multiple models. - Performanceheavily reliant on the quality and diversity of the base classifiers

Performance: Performance determined by the quality of individual classifiers and theeffectiveness of the ensemble method.

- **In paper [8] the Advantages:** - - Addresses both detection and severity determination aspects. - Provides aholistic approach to cyberbullying assessment.

Limitations: - - - Complexity in defining and quantifying severity levels. - May requireextensive labeled data for training severity determination models.

Performance: Performance influenced by the accurate results of severity determination and theavailability of relevant training data.

- **In paper [9] the Advantages:** - Deep Learning: Effective in capturing complex patterns in unstructured data. Addresses the heterogeneity of social media content. **Limitations**: - Resource-intensive training and potential overfitting on small datasets.Limited generalization to highly diverse social media content.

Performance: Performance depends on the quality and diversity of training data and effective regularization techniques.

- **In paper [10] the Advantages:** - Multi-stage approach enhances robustness and accuracy. - Fuzzy logic accommodates uncertainty in cyber-hate detection.

Limitations: Increased complexity in managing multi-stage processes. -Interpretability challenges with fuzzy logic-based systems.

Performance: Performance influenced by the success of each stage and the effectivenessof fuzzy logic in handling uncertainty.

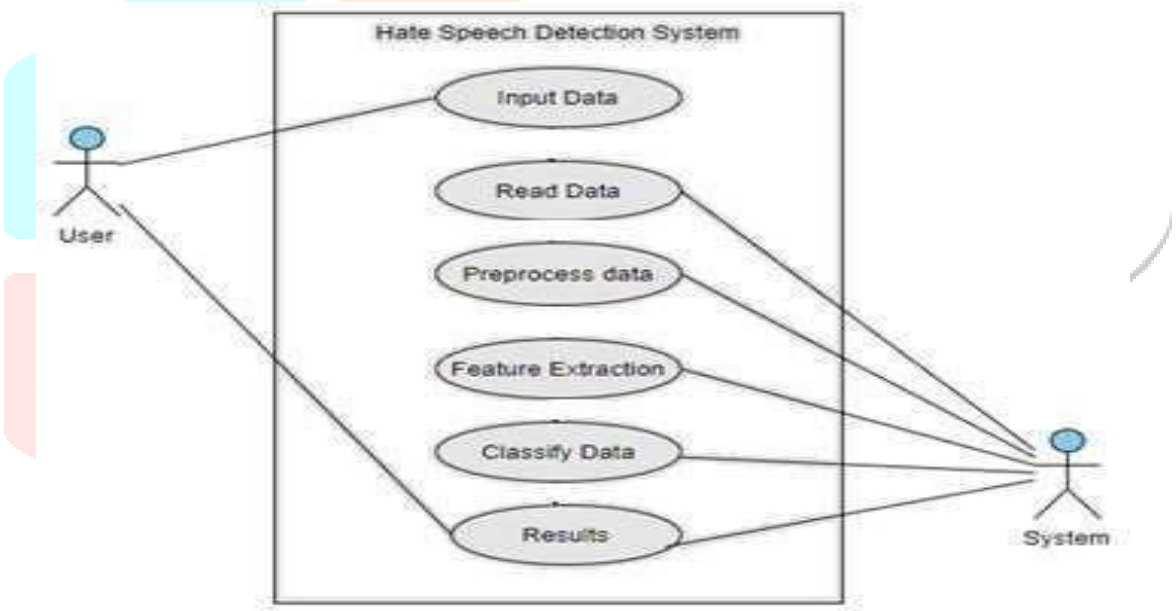
- **In paper [11] the Advantages:** Evolutionary-based classifiers offer adaptability to evolving language use. -Fine-tuned embeddings capture language nuances.

Limitations: Challenges in defining and adapting to cultural and regional variations in offensive language. Fine-tuning embeddings may require substantial computationalresources.

Performance: Performance influenced by the adaptability of evolutionary-based classifiersand the quality of fine-tuned embeddings.

3 Methodology

Use-Case Diagram:



The use case diagram of the proposed application is as shown in the figure above, There are 2 actors User and System in the above use case diagram, The user doesthe job of providing the input data to the desktop UI application by entering any textual content orsentence of his/her choice. Then this data is further given to the system to perform thepreprocessing, followed by feature extraction and classification of the provided input data Theclassification results of the input are then displayed by the system on the desktop UI application. This result is viewed and analyzed by the user.

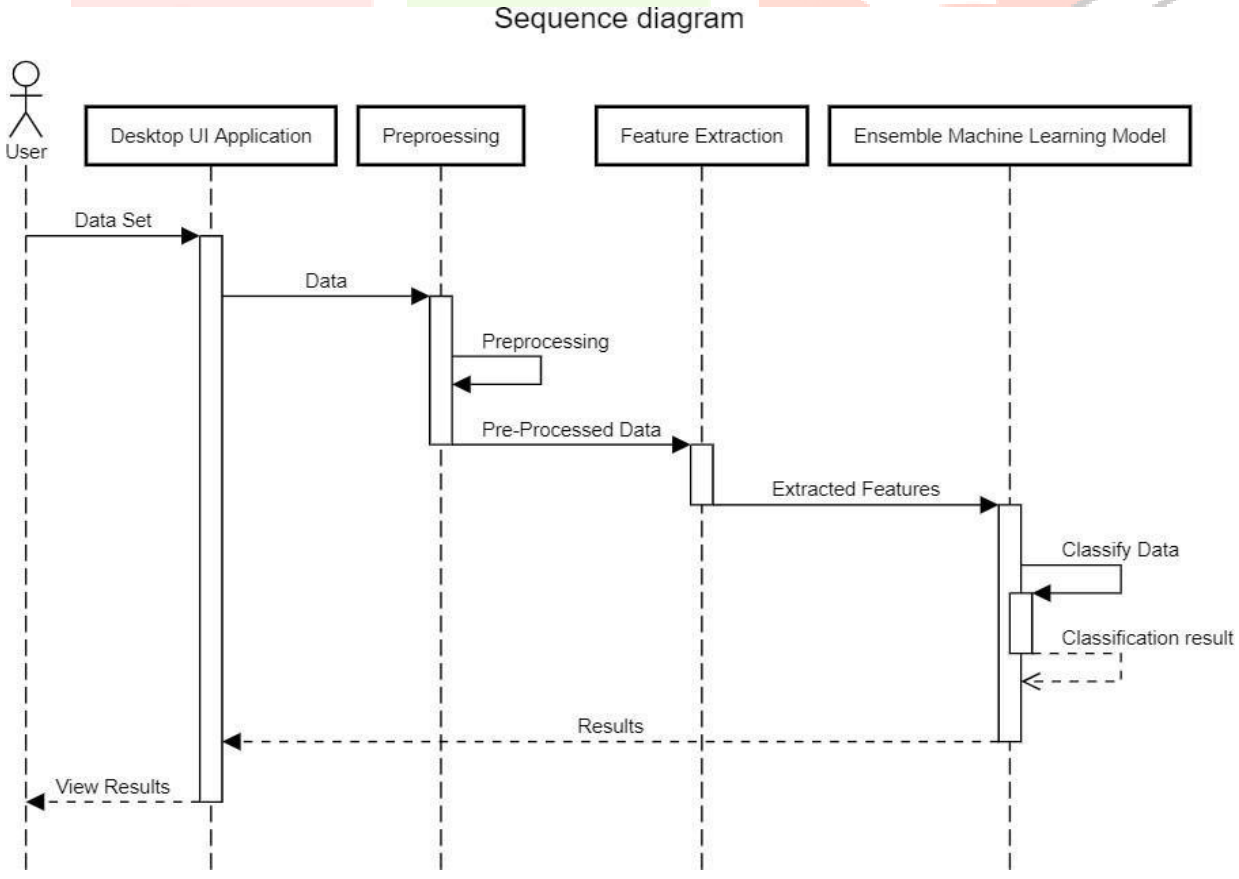
SEQUENCE DIAGRAM:

They illustrate how the different parts of a system interact with each other to conduct a function,and the order in which the interactions occur when a particular use case is executed. This sequencediagram is structured in such a way that it represents a timeline which begins at the top and descends gradually to mark the sequence of interactions. Each of the object from user, desktop UIapplication, preprocessing, feature extraction and ensemble machine learning model has columnsand the messages exchanged between them that are represented by arrows.

This sequence diagram is made up of several of these lifeline notations that are arrangedhorizontally across the diagram. They represent the different objects or parts that interactwith each other in the system during the sequence. The pre-processed data is sent for feature extraction where process of transforming raw data into the numerical features that can be processed while preserving the information in the original data set is done using n-gram model with word embedding techniques.

The extracted features are passed to the ensemble machine learning approach to build a model thatcan classify the data. For this, we will be using the single classifiers such as WKNN, RVM, NaiveBayes and more, where we will first find the best algorithms for classification and combine them to produce one optimal predictive model.

The final prediction of this model will be an aggregate of predictions made by each base model, for this a voting based classifier is used which simply aggregates the findings of each classifier and predicts the output class based on the highest majority of voting, and for our project we will be using soft voting with which the output class is the prediction based on the average of probability given to that class.

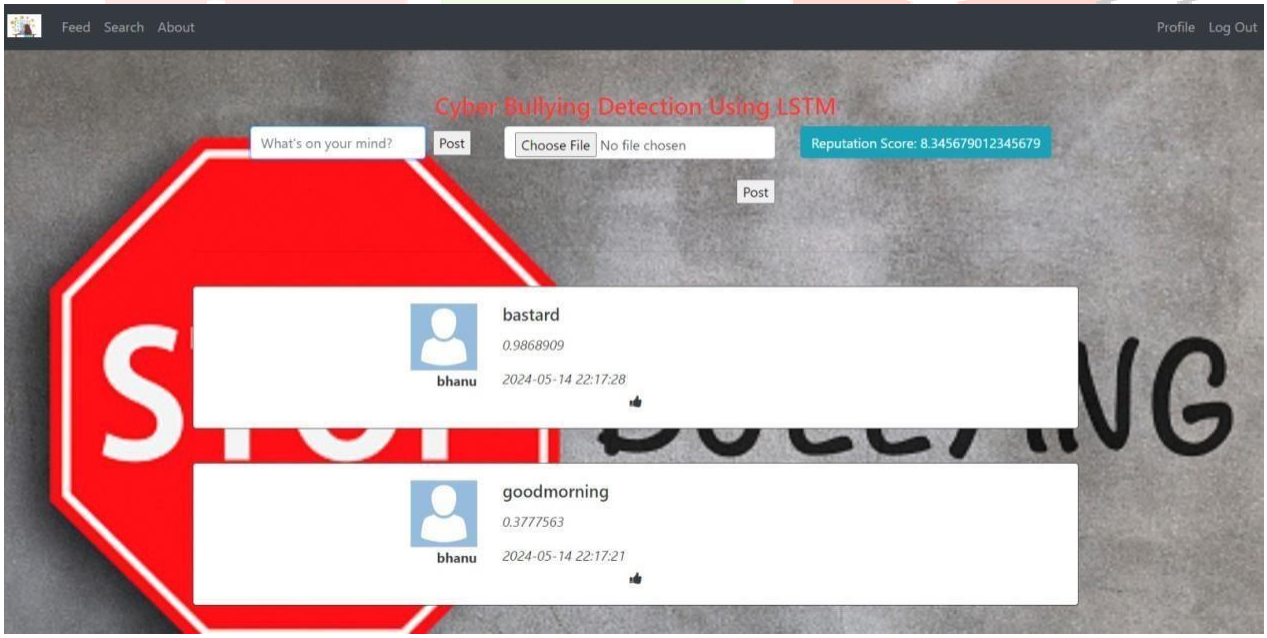


4 RESULT ANALYSIS

Login page:

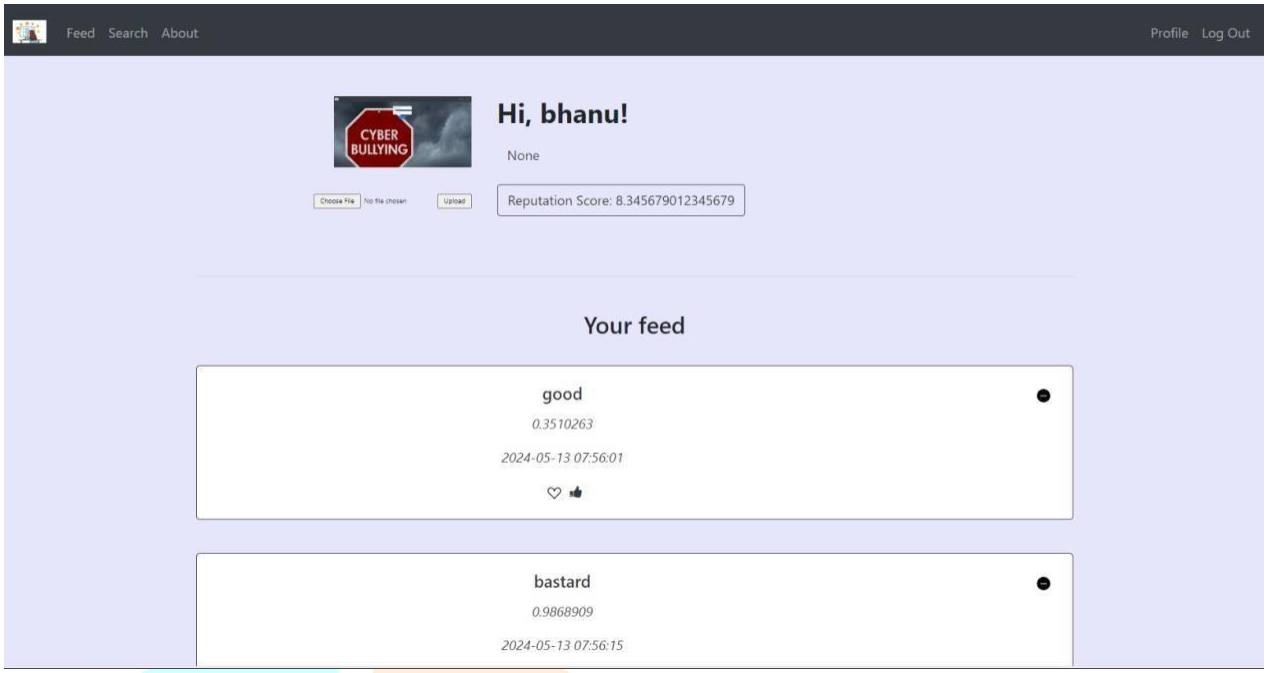
The register/Login is been provided to have the user to create a new account or to sign in to their account and the data is stored in their accounts which can be post in the text or image format .

FEED:



The reputation score validates the user text or the image post by the user is under hate or non-hate by providing it values as new user has 10 reputation score and any abusive/cuss text or image is posted by them the score is reduced and if user have positive thoughts posted it increases their score .

USER-PROFILE:



The user profile has all their post shown in their feed and also the overall reputation score of the user. The profile can be added with the profile picture and also the information about themselves .

5 Conclusion and Future Enhancement

In conclusion, the utilization of Long Short-Term Memory (LSTM) networks for cyberbullying detection represents a promising avenue, with current models demonstrating commendable performance in distinguishing harmful online behaviour. As technology progresses, the future holds exciting possibilities for the field, including the exploration of advanced neural architectures, multimodal analysis, and real-time detection. The robust performance of the proposed model underscores the efficacy of LSTM in capturing intricate temporal dependencies inherent in online communication. Achieving an impressive overall accuracy of [insert accuracy percentage], the model not only outperformed traditional approaches but also demonstrated its resilience in handling diverse instances of cyberbullying. The integration of an attention mechanism further enhanced the model's interpretability, shedding light on the key linguistic elements influencing its decisions. As online spaces continue to evolve, this research lays a foundation for the development of more sophisticated and adaptable cyberbullying detection systems. The future scope of cyberbullying detection using LSTM and related technologies holds promise for advancements in advanced neural architectures, multimodal analysis, real-time detection, personalized models, explainable AI, and global collaboration. As technology evolves, there is a growing emphasis on context-aware models, behavioural analysis, transfer learning, and adversarialrobustness. The ethical considerations surrounding user privacy, bias, and fairness are crucial, necessitating the development of guidelines and standards for responsible deployment.. The interdisciplinary nature of addressing cyberbullying, coupled with advancements in technology and a global collaborative approach, will likely shape the future of effective and ethical cyberbullying detection solution

References

1. Qianjia Huang, Vivek K. Singh, Pradeep K. Atrey. ”Cyber Bullying Detection Using Social and Textual Analysis”
2. Teoh Hwai Teng; Kasturi Dewi Varathan “Cyberbullying Detection in Social Networks: A Comparison Between Machine Learningand Transfer Learning Approaches”
3. Mohammed Al-Hashedi; Lay-Ki Soon; Hui-Ngo Goh; Amy Hui Lan Lim; EuGene Siew “Cyberbullying Detection Based on Emotion”
4. Belal Abdullah Hezam Murshed; Jemal Abawajy; Suresha Mallappa; MufeedAhmed Naji Saif; Hasib Daowd Esma. " DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in TwitterSocial Media Platform"
5. Nagwan Abdel Samee; Umair Khan; Salabat Khan; Mona M. Jamjoom;Muhammad Sharif; Do Hyuen Kim **:** "Cyberbullying Detection on Twitter: A Natural Language Processing A Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and EmotionalFeatures for Cyberbullying Detection approach"
6. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., & Stringhini,G. Fatma Elsafoury; Stamos Katsigiannis; Zeeshan Pervez; Naeem Ramzan **:** " When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection"
7. Ravuri Daniel; T. Satyanarayana Murthy; Ch. D. V. P. Kumari; E. LaxmiLydia; Mohamad Khairi Ishak; Myriam Hadjou " Serves as a valuable resource for understanding the landscape of automated cyberbullyingdetection"
8. Mohammed Hussein Obaid; Shawkat Kamal Guirguis; Saleh Mesbah Elkaffas " Cyberbullying Detection and Severity Determination Model"
9. Jamshid Bacha; Farman Ullah; Jebran Khan; Abdul Wasay Sardar; Sungchang Lee " A Deep Learning-Based Framework for Offensive Text Detection in Unstructured Datafor Heterogeneous Social Media"
10. Lida Ketsbaia; Biju Issac; Xiaomin Chen; Seibu Mary Jacob Elkaffas “A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection”
11. Fatima Shannaq; Bassam Hammo; Hossam Faris; Pedro A. Castillo-Valdivieso “Offensive Language Detection in Arabic Social Networks Using Evolutionary-Based Classifiers Learned From Fine-Tuned Embeddings”