Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

Full Length Article

Detecting Digital Harassment : The Bullynet Approach To Cyberbully Identification (NLP)

Mrs. Sadia Kausar[1], Ms. Anjali Kumari[2], Mr. Mohammed Aqib Ali[3],Mr. Syed Abdul Jabbar[4]
[1]Assistant Professor, Dept. Of CSE-AIML, Lords Institute Of Engineering And Technology, Hyderabad, India.
[2,3,4]B.E Student Dept. Of CSE-AIML, Lords Institute Of Engineering And Technology, Hyderabad, India.
Mail Id; Sadiakausar@lords.ac.in[1],anjalikumari13442@gmail.com[2],mohdaqibali050@gmail.com[3], jabbarsyed52@gmail.com[4]

Accepted 27-03-2026

***Abstract**:*
The rapid growth of digital communication platforms has transformed how people interact, but it has also led to an increase in cyberbullying and online harassment, creating significant challenges for users, especially adolescents and young adults. While existing moderation systems attempt to address this issue, their effectiveness remains limited due to reliance on manual reporting and keyword-based filtering, which fail to capture contextual meaning and evolving slang. This paper presents BullyNet, a system that detects cyberbullying in textual data using Natural Language Processing (NLP) techniques. It combines transformer-based models such as BERT with sentiment analysis and hybrid classification to improve detection accuracy. The system processes user - generated content, identifies harmful intent, and classifies abusive language while reducing false positives. Evaluated for accuracy and efficiency using datasets from platforms like Twitter and YouTube, the system shows promise as an effective tool to enhance online safety and promote healthier digital communication environments.

INTRODUCTION

The widespread use of digital platforms has transformed communication by enabling individuals to share ideas, express opinions, and interact globally. However, this rapid growth has also contributed to an increase in cyberbullying and online harassment, which can lead to psychological stress, social isolation, and reduced self-esteem among users. In today's digital environment, large volumes of user-generated content are continuously exchanged across social media and online communities, intensifying these challenges due to linguistic diversity, informal language, and the evolving nature of slang. Cyberbullying detection is essential for maintaining safe online spaces, yet its implementation remains limited because understanding context and intent in communication is complex. Most existing systems rely on keyword filtering or traditional machine learning approaches, which often fail to interpret sarcasm, implicit abuse, and contextual meaning. This limitation reduces the effectiveness of moderation systems and allows harmful content to persist. Furthermore, variations in communication styles across platforms make consistent detection difficult, highlighting the need for advanced, context-aware solutions.

PROJECT OVERVIEW

This project addresses these challenges by developing an advanced system capable of detecting cyberbullying and online harassment in textual data using Natural Language Processing techniques. The system leverages transformer-based models such as BERT along with sentiment analysis and hybrid classification methods to ensure accurate and context-aware detection. It processes user-generated content, identifies harmful intent, and classifies abusive language in real time. By combining contextual embeddings with classification algorithms, the proposed solution provides an efficient and scalable approach for improving online safety and supporting automated moderation.

OBJECTIVE

The primary objective of this project is to develop a robust and accurate system for detecting cyberbullying and online harassment in textual data while ensuring real-time performance suitable for practical deployment. The project also aims to utilize advanced Natural Language Processing techniques, including transformer-based models, to capture contextual meaning and improve detection accuracy. Another objective is to evaluate the system's effectiveness in terms of classification accuracy, precision, recall, and overall usability across multiple platforms. Ultimately, the system seeks to promote

IJPP

Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

safer digital environments by enabling efficient moderation and reducing the impact of harmful online communication.

LITERATURE SURVEY

Several studies have explored cyberbullying detection using different machine learning and deep learning approaches. Early research by S. Dinakar et al. (2014) focused on detecting cyberbullying on Instagram using supervised machine learning techniques such as Support Vector Machines with keyword-based features. While the approach achieved moderate accuracy, it lacked contextual understanding and struggled with implicit harassment and sarcasm. Similarly, M. Dadvar et al. (2013) improved detection by incorporating user profile features along with textual data using SVM models on MySpace datasets. Although classification performance improved, the model remained platform-specific and did not generalize well. Y. Kim (2014) introduced Convolutional Neural Networks for sentence classification, demonstrating strong performance on short-text datasets suitable for abusive language detection, but CNN models required large labeled datasets and had limitations in capturing long-range dependencies.

The introduction of transformer-based models significantly improved contextual understanding. J. Devlin et al. (2019) proposed BERT, which captures deep bidirectional contextual relationships in language and achieved state-of-the-art performance across NLP tasks. However, the model requires substantial computational resources. Surveys by H. Schmidt and M. Wiegand (2017) and G. Fortuna and S. Nunes (2018) reviewed hate speech detection approaches, highlighting challenges such as data imbalance, sarcasm handling, and evolving language. Later works introduced advanced deep learning techniques, such as the hierarchical LSTM model proposed by J. Xu et al. (2020), which improved contextual understanding but still struggled with implicit abuse. S. Park and J. Y. Lee (2022) used BERT with attention mechanisms for sarcasm detection, improving implicit harassment identification at the cost of computational complexity. Hybrid classification approaches, such as the one proposed by B. Xu et al. (2019), combined traditional machine learning with deep learning techniques to enhance precision and recall. Furthermore, R. Yin and Y. Zhang (2021) introduced a multi-task learning approach that simultaneously detects cyberbullying and classifies severity, improving overall performance but increasing model complexity.

SYSTEM ANALYSIS – EXISTING SYSTEM

Existing cyberbullying detection systems primarily rely on keyword-based filtering, which uses predefined lists of abusive words to identify harmful content. Although simple to implement, this approach fails to capture contextual meaning, sarcasm, and evolving slang. Traditional machine learning models such as Support Vector Machines and Naïve Bayes utilize manually engineered features like TF-IDF, but they lack deep semantic understanding. Deep learning models including CNNs and LSTMs improve detection accuracy by learning complex patterns in text, yet they still struggle with implicit abuse and require large datasets for training. As a result, existing systems often suffer from high false positives, poor context awareness, and limited adaptability across different platforms.

PROPOSED SYSTEM

The proposed system introduces a context-aware cyberbullying detection framework that accepts textual data from online platforms for real-time analysis. A preprocessing module cleans and prepares text by removing noise and normalizing content. The system then generates contextual embeddings using transformer-based models such as BERT to capture semantic meaning. A hybrid classification engine analyzes these embeddings to classify content as abusive or non-abusive with improved accuracy. Additionally, a severity scoring mechanism assigns priority levels to harmful content, assisting moderators in decision-making. The system provides an efficient moderation interface with low latency and supports scalability across multiple platforms while adapting to evolving language patterns.

ADVANTAGES

The proposed system offers several advantages over existing approaches. By using transformer-based models and hybrid classification, it improves detection accuracy and reduces misclassification. The system captures contextual meaning, sarcasm, and evolving slang, enhancing the identification of harmful content. Real-time detection enables fast processing of user-generated content, allowing timely moderation and response. The architecture is scalable and can be integrated across multiple platforms with different communication styles. Furthermore, severity scoring prioritizes harmful content, assisting moderators and improving overall efficiency.

REQUIREMENT SPECIFICATIONS

The software requirements include an operating system such as Windows 10 or later, Linux, or macOS. Programming languages such as Python are used for machine learning models and backend processing, while JavaScript, HTML, and CSS are used for frontend development. Machine learning frameworks such as TensorFlow, PyTorch, or Scikit-learn are required for model training and

Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

classification. Natural Language Processing tools including NLTK, SpaCy, and Transformers are used for preprocessing and contextual analysis. Additional components include text processing modules for tokenization and normalization, classification modules implementing BERT and hybrid classifiers, visualization tools such as Streamlit or Flask for user interfaces, and a database for storing datasets, trained models, and results.

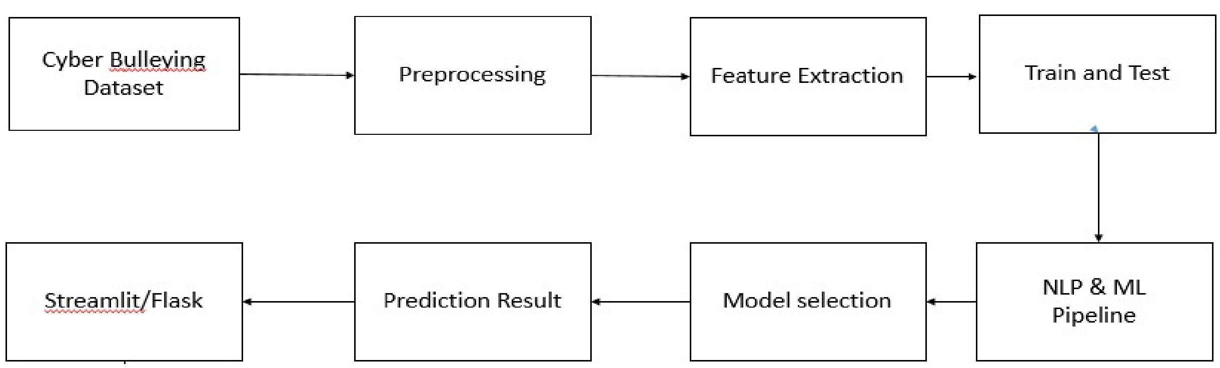
The hardware requirements include a minimum Intel i5 processor or equivalent for smooth execution, at least 8 GB RAM for handling datasets and model inference, and a dedicated GPU such as NVIDIA GTX 1050 or better for faster deep learning computation. Adequate storage space is also required for datasets, trained models, and system files.

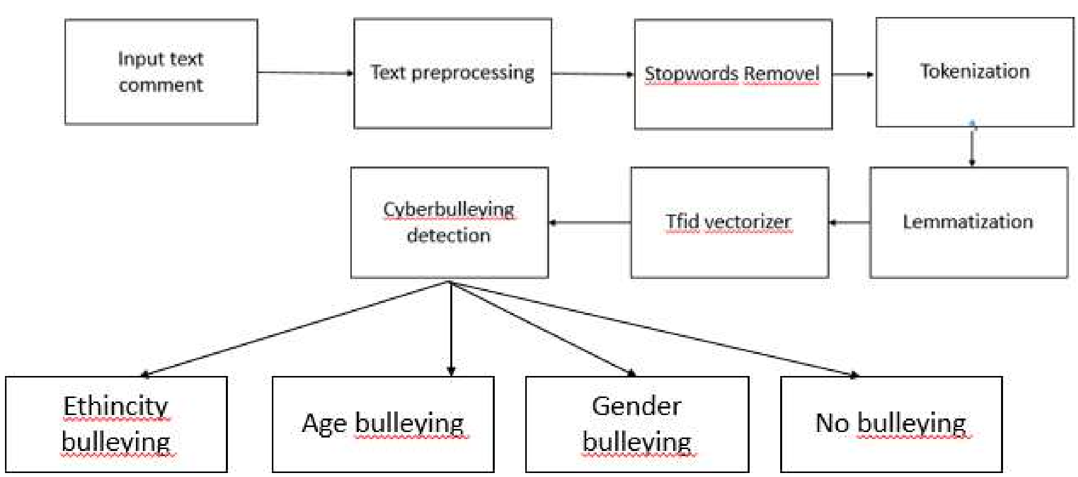
SYSTEM DESIGN – SYSTEMARCHITECTURE

The proposed architecture follows a modular pipeline for cyberbullying detection. The input module accepts user-generated textual data from social media platforms, online forums, or messaging systems. The preprocessing module cleans the text by removing noise, stop words, and performing tokenization and normalization using NLP libraries. The embedding generation module converts processed text into contextual representations using transformer-based models such as BERT. These embeddings are then analyzed by the classification engine, which uses hybrid machine learning models to classify content as abusive or non-abusive. A moderation interface displays flagged content along with severity levels to assist moderators in decision-making. Finally, the output module presents classification results and alerts through a user-friendly interface, enabling efficient monitoring and response.

Moderation Interface: Displays flagged content and severity levels to assist moderators in decision-making.

Output Module: Presents classification results and alerts through a user-friendly interface for efficient monitoring and response.





Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

MODULES

The proposed cyberbullying detection system is divided into several functional modules that work together to process and analyze textual data. The input module accepts user-generated textual content from online platforms and performs basic validation before forwarding it to subsequent stages. The preprocessing module then cleans the text by removing noise, special characters, and stop words, while also performing tokenization and normalization. After preprocessing, the NLP and embedding module generates contextual representations using transformer-based models such as BERT, which helps in understanding the semantic meaning of the text. These embeddings are passed to the classification module, which categorizes the content as abusive or non-abusive using machine learning or deep learning models. The severity analysis module evaluates the intensity and intent of harmful language and assigns severity levels such as low, medium, or high. Additionally, an optional subtitle generator module can translate text into regional languages for subtitle display. Finally, the output module presents classification results and alerts through a clear and user-friendly interface.

IMPLEMENTATION – INPUT DESIGN

The input design focuses on efficiently handling textual data for accurate processing. During preprocessing, the system applies text cleaning, tokenization, and normalization using NLP libraries. Noise such as special characters and irrelevant symbols is removed to ensure consistency. The system accepts text input either through manual entry in a user interface text box or through dataset files such as .txt formats. Validation mechanisms are also implemented to ensure reliable input handling. Non-alphabetic characters are filtered, missing or invalid inputs are handled appropriately, and excessively long text is truncated to maintain efficient processing and reduce computational overhead.

OUTPUT DESIGN

The output design ensures that results are clearly presented to users. The classification output is displayed as labeled results indicating whether the content is abusive or non-abusive. In addition to classification, severity visualization is provided to show the intensity level of harmful content, such as low, medium, or high. Visual cues, including color coding, are used to enhance readability and enable quick interpretation. All outputs are presented through a user-friendly dashboard or interface, allowing moderators and users to easily monitor flagged content.

SAMPLE CODE

The implementation includes sample code for text preprocessing, classification, and severity scoring. In the preprocessing stage, Python functions are used to convert text to lowercase and remove non-alphabetic characters using regular expressions. For cyberbullying detection, a simple classification example uses CountVectorizer to convert text into numerical features and a Multinomial Naïve Bayes model to classify content as abusive or non-abusive. A prediction function processes new input text through the preprocessing pipeline and returns classification results. Additionally, a severity scoring function evaluates the presence of abusive words and assigns severity levels based on the frequency of harmful terms detected in the input text.

IMPLEMENTATION – TOOLS AND WORKFLOW

The system implementation consists of frontend, backend, and integration components. The frontend is built using frameworks such as Streamlit or Flask, providing web-based access with input fields for entering text and displaying classification results. The backend handles text processing, model prediction, and severity analysis. Text preprocessing is performed using NLP libraries for cleaning and tokenization, while trained machine learning or deep learning models such as BERT are used for classification. The severity analysis component assigns levels based on harmful content. During integration, text inputs pass through preprocessing and classification modules, and the results are displayed on the interface along with severity indicators. For example, when a user enters the text “You are useless,” the system preprocesses it, classifies it as abusive, and displays the result along with a high severity level.

SOFTWARE TESTING

Software testing ensures reliability and performance of the system. Unit testing validates individual modules such as text preprocessing and classification accuracy by testing tokenization, cleaning functions, and prediction outputs. Integration testing verifies the interaction between modules, ensuring smooth data flow from input through preprocessing to classification. System testing evaluates end-to-end functionality, including real-time detection performance and response time, by testing model accuracy and latency on live inputs. Usability testing assesses the interface for moderators and users by evaluating clarity of results, ease of use, readability, and the effectiveness of severity indicators.

RESULT ANALYSIS

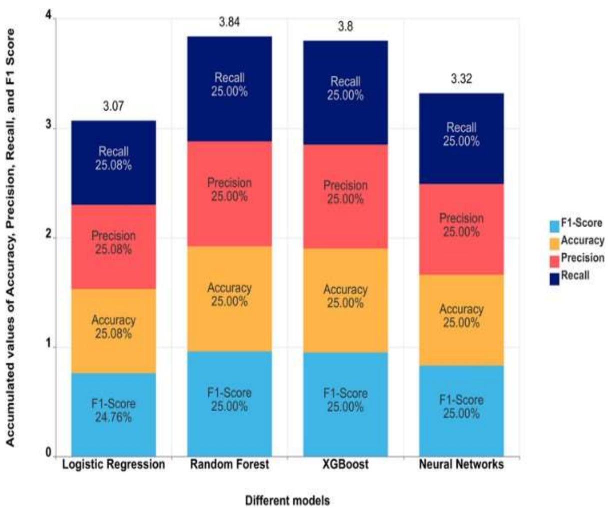
The result analysis evaluates the performance of the proposed system using various metrics. Deep learning models achieved higher precision and recall by effectively capturing contextual meaning and reducing false positives, whereas keyword-based

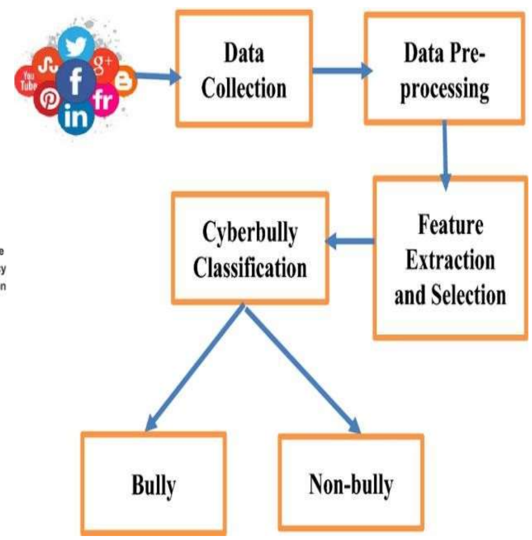
Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

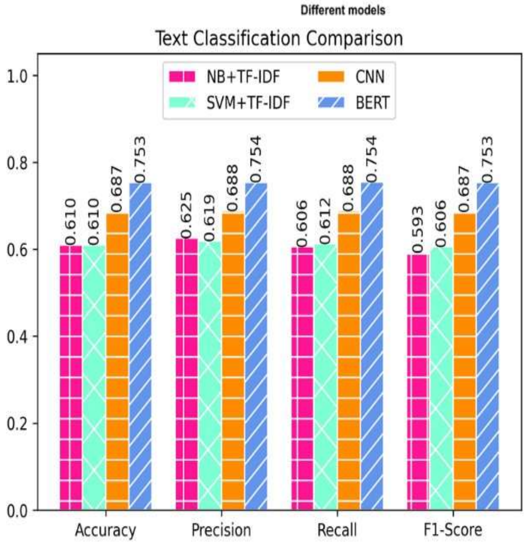
systems reported lower precision due to frequent misclassification of non-abusive content. In terms of accuracy, transformer-based models such as BERT achieved high accuracy, typically around 90–95 percent, outperforming traditional machine learning models like SVM and Naïve Bayes, which showed lower accuracy due to limited contextual understanding. Hybrid models demonstrated balanced performance with improved F1-scores compared to standalone approaches. Additionally, real-time systems achieved low latency with fast response times, making them suitable for live content moderation.

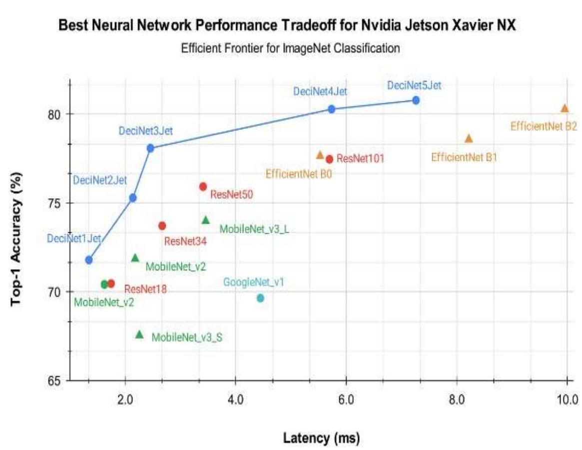
BENCHMARK COMPARISON

The benchmark comparison highlights the effectiveness of the proposed system compared to traditional approaches. BullyNet demonstrates better context awareness and higher accuracy than keyword-based and basic machine learning methods. It reduces false positives and improves detection of implicit abusive content. Transformer-based deep learning models outperform CNN and SVM models in capturing semantic meaning and contextual relationships. However, these advanced models require higher computational resources for training and deployment, which is an important consideration for real-world implementation.









Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

FUTURE SCOPE

The future scope of the cyberbullying detection system is extensive and promising. One major area of improvement is the inclusion of multilingual support, enabling detection across various regional and global languages to make the system more accessible. Additionally, the system can be enhanced by integrating advanced context understanding techniques such as sarcasm detection and emotion recognition to better capture implicit abusive intent. The adoption of more advanced transformer-based models can further improve accuracy and real-time performance, making the system more adaptive and efficient. There is also significant potential in developing domain-specific detection systems for platforms such as education, gaming, and professional networks, where communication patterns differ. Real-world deployment through mobile and web applications, including offline capabilities, can increase usability in low-connectivity environments. Expanding datasets through crowdsourcing and continuous learning will improve model robustness and adaptability to evolving language trends. Finally, integrating personalized moderation settings and ethical AI frameworks can enhance user trust and promote safer digital communication environments.

CONCLUSION

In conclusion, this project successfully addresses critical challenges in cyberbullying detection by combining Natural Language Processing, transformer-based models, and hybrid classification techniques. The approach effectively captures contextual meaning and improves detection accuracy, while the inclusion of severity analysis enhances moderation efficiency. The system demonstrates strong performance in identifying abusive content with reduced false positives and low latency, making it suitable for real-time applications. Despite these advancements, challenges such as handling sarcasm, adapting to evolving language patterns, and ensuring fairness in classification remain. Continued research, dataset expansion, and improvements in model design will be essential to overcome these limitations and develop a more robust and reliable system for promoting safer digital communicat ion environments.

Reference

1. S. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Cyberbullying Detection on Instagram," *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. [Online]. Available:https://ojs.aaai.org/index.php/ICWSM/article/view/14575
2. M. Dadvar, D. Trieschnigg, and F. de Jong, "Improving Cyberbullying Detection with User Context," *European Conference on Information Retrieval*, 2013. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-36973-5_40
3. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014. [Online]. Available: https://aclanthology.org/D14-1181/
4. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019. [Online]. Available: https://aclanthology.org/N19-1423/
5. H. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* , 2017. [Online]. Available: https://aclanthology.org/W17-1101/
6. G. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3232676
7. J. Xu, X. Zhu, and A. Bellmore, "Hierarchical Cyberbullying Detection using Deep Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6499
8. S. Park and J. Y. Lee, "Sarcasm Detection using Contextual Embeddings," *IEEE Access*, vol. 10, pp. 1–10, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9705561
9. B. Xu, H. Zhu, and Y. Wang, "Hybrid Classification Models for Online Harassment Detection," *International Journal of Data Science and Analytics*, 2019. [Online]. Available: https://link.springer.com/article/10.1007/s41060-019-00164-3
10. R. Yin and Y. Zhang, "Multi-task Learning for Cyberbullying Detection and Severity Classification," *IEEE Transactions on Knowledge and Data Engineering*, 2021. [Online] https://ieeexplore.ieee.org/document/9444822
11. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of ICWSM*, 2017. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/149

Global Journal of Sociology and Anthropology
Vol. 15 (2), pp. 001-007, April, 2025
ISSN: 2756-3456

55

12. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, 2017. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14955
13. Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," *Proceedings of NAACL-HLT*, pp. 88–93, 2016. [Online]. Available: https://aclanthology.org/N16-2013/
14. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3041021.3054223
15. A. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," *Complexity*, vol. 2019, pp. 1–10, 2019. [Online]. Available: https://doi.org/10.1155/2019/5619528
16. E. Founta, C. Djouvas, D. Chatzakou, et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, pp. 491–500, 2018. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14980
17. T. Zhang, V. Robinson, and H. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," *European Semantic Web Conference*, pp. 745–760, 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_47
18. T. Zhang, V. Robinson, and H. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," *European Semantic Web Conference*, pp. 745–760, 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_47
19. A. Fortuna, J. Soler-Company, and P. Nunes, "Offensive Language Detection in Social Media Using Deep Learning," *Information Processing & Management*, vol. 57, no. 3, pp. 102–118, 2020. [Online]. Available: https://doi.org/10.1016/j.ipm.2019.102118
20. K. Chatzakou, N. Kourtellis, J. Blackburn, et al., "Mean Birds: Detecting Aggression and Bullying on Twitter," *Proceedings of the ACM Web Science Conference*, pp. 13–22, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3091478.3091487