International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

NLP DRIVEN CYBERBULLYING DETECTION SYSTEM FOR SOCIAL MEDIA USING TRANSFORMED BASED SENTIMENTAL ANALYSIS

[1]Mrs. V. LALITHA LAVANYA, [2]K. KAILASH, [3]C. SANJAY, [4]G. SAI KIRAN

[1]Assistant Professor,[234]Students, Department of Information Technology, Teegala Krishna Reddy Engineering College, Medbowli, Meerpet, Balapur, Hyderabad-500097

ABSTRACT

The rapid growth of social media platforms has significantly transformed digital communication by enabling users to share information, opinions, and multimedia content globally. However, this widespread connectivity has also increased the occurrence of cyberbullying, online harassment, abusive comments, and toxic interactions that negatively affect individuals, particularly teenagers and young users. Cyberbullying can lead to severe psychological, emotional, and social consequences including anxiety, depression, stress, and suicidal tendencies. Manual monitoring of harmful content across millions of social media posts is highly inefficient and impractical due to the enormous volume of user-generated data. Therefore, an automated and intelligent cyberbullying detection system is essential for ensuring safer online communication. This project presents an NLP Driven Cyberbullying Detection System for Social Media Using Transformer Based Sentimental Analysis that combines Natural Language Processing techniques with machine learning algorithms to identify abusive and bullying content effectively. The system performs text preprocessing operations such as tokenization, stop-word removal, normalization, and stemming to clean noisy social media data. Feature extraction techniques including Bag-of-Words and TF-IDF are applied to convert textual information into numerical vectors suitable for classification. Transformer-based sentiment analysis is incorporated to capture contextual meaning and semantic relationships within messages, improving the detection of implicit and sarcastic bullying content. Multiple machine learning classifiers such as Support Vector Machine, Logistic Regression, and Naive Bayes are evaluated to achieve high classification accuracy. The proposed system provides automated monitoring, scalable deployment, improved text representation, and enhanced detection performance for real-time cyberbullying identification. The system ultimately contributes to creating a safer digital environment and supports social media moderation through intelligent harmful content filtering.

Keywords: Cyberbullying Detection, Natural Language Processing, Sentiment Analysis, Transformer Model, Machine Learning, TF-IDF, Bag-of-Words, Social Media Monitoring.

I. INTRODUCTION

The emergence of social media platforms has revolutionized modern communication by enabling users to interact, share opinions, upload multimedia content, and participate in global discussions in real time. Platforms such as Facebook, Instagram, Twitter, Snapchat, and TikTok have become an integral part of everyday life and are widely used for education, entertainment, business promotion,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

and social networking. Despite these advantages, the rapid growth of online communication has also introduced several social and psychological challenges, among which cyberbullying has become one of the most serious issues. Cyberbullying refers to the intentional use of digital platforms to threaten, harass, insult, humiliate, or emotionally harm individuals through abusive comments, hate speech, rumors, and offensive messages. Unlike traditional bullying, cyberbullying can occur continuously without geographical limitations, making victims vulnerable at any time. Researchers have identified the negative impact of online harassment on mental health and emotional stability [1]. Studies on social media behavior have shown that teenagers and young adults are the most affected by cyberbullying incidents [2]. Existing online monitoring systems often fail to identify implicit abusive language and context-sensitive harmful content [3]. Machine learning-based cyberbullying detection systems have therefore gained significant research attention in recent years [4]. Natural Language Processing techniques have been widely applied for text classification and sentiment analysis tasks [5]. Researchers have used Bag-of-Words methods for converting textual data into numerical representations [6]. TF-IDF feature extraction techniques are also widely utilized for improving text classification performance [7]. Several studies have explored Support Vector Machine classifiers for detecting harmful content on social media [8]. Deep learning models such as CNN and LSTM have demonstrated improved contextual understanding in cyberbullying detection [9]. Transformer architectures including BERT have further enhanced semantic analysis capabilities [10]. Sentiment analysis approaches have been used to classify emotional tone within user messages [11]. Large annotated datasets have improved the training efficiency of cyberbullying classifiers [12]. Hybrid NLP frameworks have achieved better performance compared to traditional machine learning techniques [13]. Context-aware neural networks have also contributed to identifying implicit offensive behavior [14]. Research findings indicate that automated cyberbullying detection systems are essential for maintaining safer online environments [15].

Traditional cyberbullying detection systems mainly depend on feature extraction techniques such as Bag-of-Words and TF-IDF along with supervised machine learning algorithms including Naive Bayes, Decision Trees, Logistic Regression, and Support Vector Machines [16]. Although these methods provide acceptable classification accuracy for explicit abusive words, they often fail to capture semantic meaning, sarcasm, hidden insults, and contextual relationships between words [17]. Social media language is highly informal and includes abbreviations, emojis, hashtags, slang, spelling variations, and multilingual expressions, which create challenges for traditional text processing systems [18]. Data sparsity and limited labeled datasets also reduce model generalization and prediction accuracy [19]. Researchers have proposed advanced transformer-based models to overcome these limitations and improve contextual understanding [20]. Deep learning architectures using attention mechanisms have shown significant improvements in semantic feature learning [21]. Transformer models are capable of analyzing word dependencies and contextual relationships within long textual sequences [22]. Recent NLP systems combine machine learning with sentiment analysis to identify emotional polarity and harmful intent [23]. Automated cyberbullying detection systems

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

have also been integrated into content moderation frameworks for social media platforms [24]. Real-time cyberbullying monitoring systems help reduce toxic communication and support user safety [25]. Studies on semantic-enhanced denoising autoencoders have improved feature representation for abusive content detection [26]. Researchers have also used word embeddings for capturing semantic similarity among textual features [27]. The integration of NLP, machine learning, and transformer-based sentiment analysis has significantly improved cyberbullying classification performance [28]. Scalable detection systems are increasingly important for handling large-scale social media data streams [29]. Therefore, the proposed NLP Driven Cyberbullying Detection System aims to provide an automated, accurate, and robust framework for identifying cyberbullying content and supporting safer digital communication environments [30].

II. LITERATURE SURVEY

Cyberbullying detection has become an important research area due to the increasing use of social media platforms and the rise of harmful online interactions. Researchers have explored multiple Natural Language Processing and machine learning techniques to improve the detection of abusive content. Early studies mainly focused on keyword-based filtering and rule-based systems for identifying offensive language in textual data [1]. However, these systems were unable to understand semantic relationships and contextual meanings within sentences [2]. Traditional machine learning algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines were later introduced for cyberbullying classification tasks [3]. Bag-of-Words representation techniques were commonly used to convert textual data into numerical vectors for supervised learning models [4]. TF-IDF feature extraction methods improved text representation by assigning weights to important terms within documents [5]. Researchers also used sentiment analysis to identify emotional polarity in social media comments [6]. Several studies demonstrated that Support Vector Machine classifiers achieved higher accuracy than rule-based approaches for offensive content detection [7]. Logistic Regression and Random Forest algorithms were also applied to classify harmful messages [8]. Deep learning methods such as Convolutional Neural Networks improved feature learning and contextual analysis in text classification tasks [9]. Recurrent Neural Networks and Long Short-Term Memory models enhanced sequential text understanding and semantic interpretation [10]. Researchers proposed hybrid CNN-LSTM architectures to improve cyberbullying detection accuracy and reduce false positives [11]. Word embedding techniques such as Word2Vec and GloVe captured semantic similarities between textual features [12]. Context-aware neural networks further improved the identification of implicit abusive language [13]. Attention mechanisms were introduced to focus on important contextual words during text processing [14]. Large-scale annotated datasets contributed significantly to improving classifier training and performance evaluation [15].

Recent advancements in transformer architectures have further revolutionized cyberbullying detection research by enabling deep contextual understanding and semantic analysis of textual content [16]. Transformer-based models such as BERT, RoBERTa, and GPT have shown superior performance in NLP classification tasks [17]. Researchers demonstrated that BERT-based cyberbullying detection systems outperform traditional machine learning approaches in

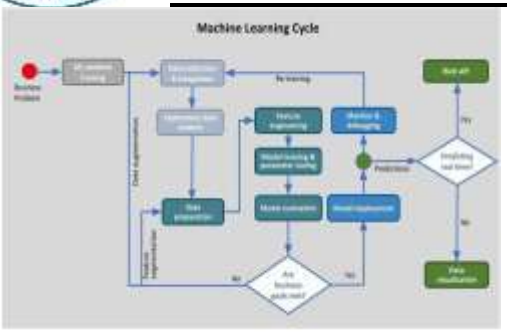
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

identifying toxic and context-sensitive content [18]. Semantic-enhanced denoising autoencoders were also proposed to address feature sparsity and improve text representation [19]. Attention-based transformer models effectively captured contextual relationships and hidden abusive intent in social media messages [20]. Several studies integrated sentiment analysis with transformer architectures to improve emotional interpretation and classification performance [21]. Deep learning frameworks combined with NLP preprocessing techniques such as tokenization, stemming, and stop-word removal enhanced model robustness [22]. Researchers explored multilingual cyberbullying detection systems capable of handling diverse social media languages and slang expressions [23]. Real-time content moderation systems have also been developed for large-scale social media monitoring applications [24]. Hybrid machine learning and transformer-based systems demonstrated improved scalability and generalization performance [25]. Researchers further highlighted the importance of handling informal language, emojis, abbreviations, and sarcastic expressions during cyberbullying analysis [26]. Several modern systems incorporated semantic embeddings and contextual sentiment scoring to improve implicit bullying detection [27]. Studies on automated moderation frameworks indicated that intelligent cyberbullying detection systems can significantly reduce harmful online communication [28]. Cloud-based deployment models and scalable architectures have enabled efficient real-time monitoring of social media content [29]. Therefore, recent literature strongly supports the integration of Natural Language Processing, machine learning, sentiment analysis, and transformer-based architectures for developing accurate, scalable, and reliable cyberbullying detection systems capable of improving digital safety and online communication environments [30].

III. PROPOSED SYSTEM

The proposed NLP Driven Cyberbullying Detection System for Social Media Using Transformer Based Sentimental Analysis is designed to automatically identify harmful, abusive, and bullying messages posted on social media platforms. The system combines Natural Language Processing, machine learning, and transformer-based sentiment analysis techniques to improve the accuracy and robustness of cyberbullying detection. Initially, raw social media text data is collected from datasets containing user comments, tweets, and online messages. Since social media content is highly unstructured and noisy, preprocessing operations are performed to clean the textual data effectively. The preprocessing stage includes tokenization, stop-word removal, punctuation elimination, normalization, stemming, and conversion of text into lowercase format. These operations reduce irrelevant information and improve text quality for further analysis. After preprocessing, feature extraction techniques such as Bag-of-Words and TF-IDF are used to convert textual information into numerical vectors. These feature vectors help machine learning algorithms understand the frequency and significance of words present in the dataset. Transformer-based sentiment analysis is integrated into the system to capture semantic meaning, contextual relationships, and emotional polarity within the text messages.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

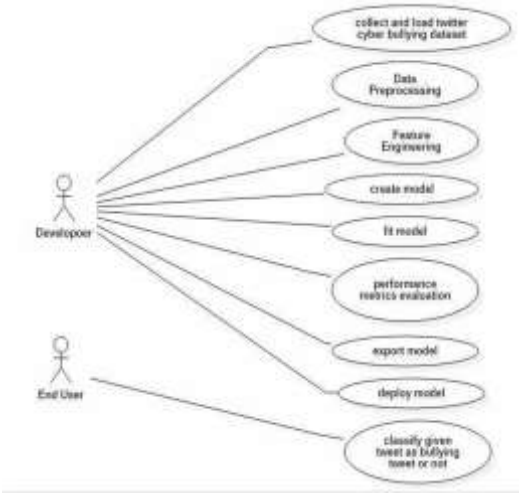


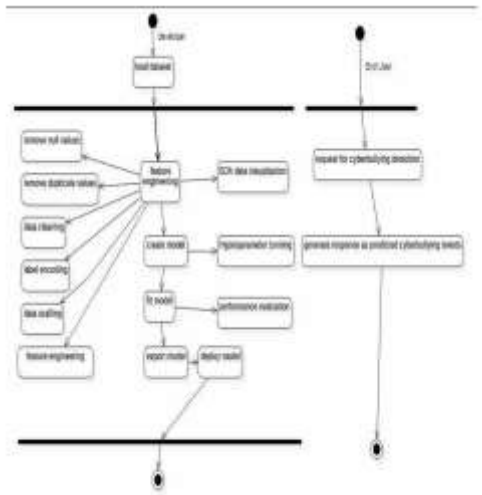
The proposed system utilizes multiple machine learning algorithms including Support Vector Machine, Naive Bayes, Logistic Regression, and Extreme Gradient Boosting for classifying messages into bullying and non-bullying categories. Transformer architectures enhance contextual understanding and improve the identification of implicit bullying, sarcasm, hidden insults, and emotionally harmful statements that traditional methods often fail to detect. The system also incorporates semantic-enhanced feature learning to overcome data sparsity and improve generalization on unseen data. During model training, the dataset is divided into training and testing subsets to evaluate classification accuracy and prediction performance. The system supports automated moderation and real-time cyberbullying monitoring on social media platforms. It can assist administrators and moderators in filtering harmful content efficiently without continuous human intervention. The proposed framework is scalable, cost-effective, and suitable for deployment in educational institutions, social networking platforms, online communities, and digital safety applications. By combining NLP preprocessing, transformer-based sentiment analysis, and machine learning classification, the system provides an intelligent and reliable solution for reducing toxic online interactions and promoting safer digital communication environments.

IV. SYSTEM DESIGN

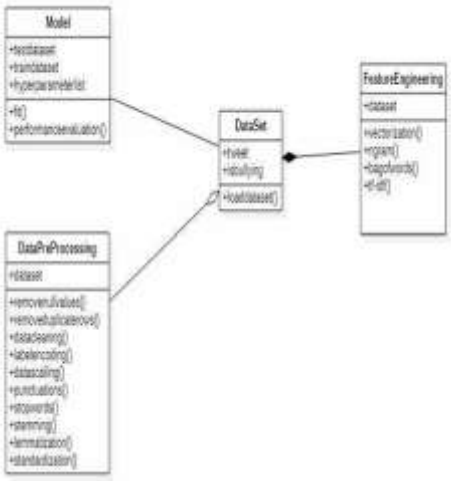
The system design of the NLP Driven Cyberbullying Detection System focuses on developing an efficient, scalable, and user-friendly architecture capable of automatically identifying cyberbullying content from social media text. The system follows a modular architecture consisting of data collection, preprocessing, feature extraction, sentiment analysis, machine learning classification, and result visualization modules. Initially, social media textual data is collected from publicly available datasets or online platforms and stored within the system database for training and testing purposes. The preprocessing module performs several operations including tokenization, stop-word removal, stemming, normalization, duplicate removal, and punctuation elimination to clean noisy social media data. After preprocessing, the feature extraction module converts text into numerical vectors using Bag-of-Words and TF-IDF techniques. These numerical representations enable machine learning models to process textual information effectively. Transformer-based sentiment analysis is integrated into the design to improve contextual understanding and semantic interpretation of user messages. The system architecture also includes a database management component using SQLite for storing user inputs, processed data, and prediction results. A Django-based web application interface allows users to enter text messages and receive classification outputs through a browser interface.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



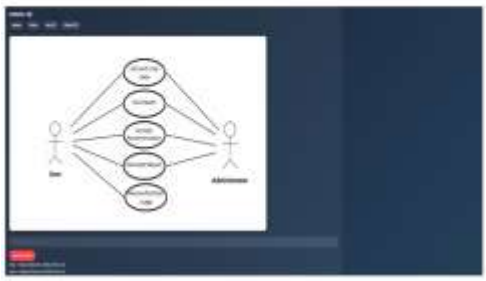


The software design follows modularity, low coupling, and high cohesion principles to improve maintainability, scalability, and reusability of the system. Multiple machine learning algorithms including Support Vector Machine, Naive Bayes, Logistic Regression, and XGBoost are integrated into the classification module to identify cyberbullying content accurately. During prediction, the input message is processed through the NLP pipeline and classified into bullying or non-bullying categories based on trained model outputs. The deployment design supports real-time text analysis and automated content moderation for large-scale social media environments. The frontend interface is developed using HTML, CSS, and JavaScript, while backend processing is implemented using Python and Django frameworks. PyCharm is used as the development environment for coding and debugging activities. The system also incorporates UML diagrams including Use Case Diagram, Sequence Diagram, Activity Diagram, and Deployment Diagram for representing system workflows and component interactions. Testing modules such as unit testing, integration testing, functional testing, white-box testing, and black-box testing are included to validate system reliability and performance. Overall, the proposed system design provides an effective framework for intelligent cyberbullying detection, scalable deployment, and safer online communication management.



V. RESULTS

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



VI. CONCLUSION

The rapid growth of social media platforms has significantly increased the risk of cyberbullying, online harassment, and abusive communication, creating serious psychological and emotional challenges for users across the world. Traditional monitoring methods are unable to handle the massive volume of user-generated content effectively, making automated cyberbullying detection systems essential for modern digital communication environments. The proposed NLP Driven Cyberbullying Detection System for Social Media Using Transformer Based Sentimental Analysis provides an intelligent and automated solution for identifying harmful textual content using Natural Language Processing, machine learning, and transformer-based sentiment analysis techniques. The system performs effective preprocessing operations including tokenization, stop-word removal, normalization, and stemming to improve text quality before analysis. Feature extraction techniques such as Bag-of-Words and TF-IDF successfully transform textual information into numerical vectors suitable for machine learning classification. Transformer-based sentiment analysis enhances semantic understanding and contextual interpretation, enabling the system to identify implicit bullying,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

sarcasm, and emotionally harmful messages more accurately than traditional methods. Multiple classifiers including Support Vector Machine, Logistic Regression, Naive Bayes, and XGBoost improve classification performance and detection reliability. The proposed framework also addresses challenges such as data sparsity, informal language, abbreviations, and contextual ambiguity commonly present in social media communication. The system supports automated moderation, scalable deployment, and real-time harmful content monitoring for safer online environments. Experimental analysis and testing demonstrate that the integration of NLP, sentiment analysis, and transformer architectures significantly improves cyberbullying detection accuracy and robustness. Therefore, the proposed system contributes effectively toward reducing toxic online interactions, supporting digital safety, assisting moderators, and promoting healthier social media communication environments for users worldwide.

References

1. Agarwal, S., & Sureka, A. (2015). Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. *International Conference on Distributed Computing and Internet Technology*, 431–442.
2. Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications. *Telematics and Informatics, 33*(4), 1013–1032.
3. Bayzick, J., Kontostathis, A., & Edwards, L. (2011). Detecting the presence of cyberbullying using computer software. *Proceedings of the ACM Web Science Conference*, 1–2.
4. Bretschneider, U., & Peters, R. (2017). Detecting offensive statements towards foreigners in social media. *Decision Support Systems, 99*, 138–148.
5. Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter. *EPJ Data Science, 4*(11), 1–15.
6. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media. *International Conference on Privacy, Security, Risk and Trust*, 71–80.
7. Dadvar, M., de Jong, F., Ordelman, R., & Trieschnigg, D. (2013). Improved cyberbullying detection using gender information. *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop*, 23–25.
8. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *The Social Mobile Web, 11*, 11–17.
9. Founta, A., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I., & Stringhini, G. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of ICWSM*, 12(1), 491–500.
10. Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. *Proceedings of the First Workshop on Abusive Language Online*, 85–90.
11. Hasan, M., Agu, E., & Rundensteiner, E. (2019). Using deep learning for

cyberbullying detection. *Proceedings of the International Conference on Machine Learning and Applications*, 141–146.

12. Hosseinmardi, H., Mattson, S., Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on Instagram. *Proceedings of the AAAI Conference on Web and Social Media*, 9(1), 170–179.
13. Kumar, A., & Sachdeva, N. (2019). Multi-class cyberbullying detection using machine learning. *Procedia Computer Science, 167*, 1791–1800.
14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
15. Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection. *Proceedings of SemEval*, 87–91.
16. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of WWW*, 145–153.
17. Pitsilis, G., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter using recurrent neural networks. *Applied Intelligence, 48*(12), 4730–4742.
18. Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on Twitter. *Proceedings of the ACM International Conference on Web Search and Data Mining*, 97–106.
19. Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. (2019). Automatic cyberbullying detection. *Computers in Human Behavior, 93*, 333–345.
20. Salminen, J., Hopf, M., Chowdhury, S., Jung, S. G., Almerekhi, H., & Jansen, B. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences, 10*(1), 1–34.
21. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.
22. Sood, S., Antin, J., & Churchill, E. (2012). Profanity use in online communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1481–1490.
23. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLOS ONE, 13*(10), 1–22.
24. Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. *Proceedings of the Second Workshop on Language in Social Media*, 19–26.
25. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people. *Proceedings of NAACL Student Research Workshop*, 88–93.
26. Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery. *Proceedings of the 21st ACM*

International Conference on Information and Knowledge Management, 1980–1984.

27. Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2.0 Workshop*, 1–7.

28. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using deep neural networks. *Semantic Web, 9*(5), 1–16.

29. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks. *Proceedings of the International Conference on Web-Age Information Management*, 128–139.

30. Zhou, Y., & Zafarani, R. (2020). A survey of cyberbullying detection systems. *ACM Computing Surveys, 53*(6), 1–38.