



Cyberbullying Detection on Social Media using Machine Learning

¹ Dr. B. Meena Preethi, ² S. Nasvana

¹Associate Professor, Department of Software System, Sri Krishna Arts & Science College, Coimbatore.

²PG Student, Department of Software System, Sri Krishna Arts & Science College, Coimbatore.

ABSTRACT

The rapid expansion of the internet has turned social media into a dominant platform for communication. However, the rise in online interactions has also led to an increase in cyberbullying incidents, which can result in severe mental distress, physical health issues, and, in extreme cases, suicide attempts, particularly among women and children. The main goal is to develop and deploy an efficient machine learning model for detecting cyberbullying content. The proposed system utilizes a dataset comprising text messages labelled as either "normal" or "cyberbullying." This dataset, sourced from the Kaggle repository, includes structured information divided into training and testing subsets. The methodology begins with the data upload process, followed by essential pre-processing steps such as data cleaning, dimensionality reduction, and normalization. A decision tree algorithm is applied to develop the classification model by training it on the given dataset. The effectiveness and accuracy of the model are subsequently evaluated using the test dataset. The system successfully classifies text as either normal or indicative of cyber bullying. Early detection of such harmful content can significantly enhance user safety on social media platforms, fostering a more positive and engaging online environment.

Keywords—Cyberbullying detection, Machine learning, Social media, Decision tree algorithm.



1. INTRODUCTION

In recent years, the swift growth of internet usage has transformed the way individuals communicate, with social media emerging a highly influential platform for communication. These platforms have become indispensable for sharing information, connecting with others, and expressing opinions. However, the exponential growth of online engagement has also brought with it significant challenges, with the rise of cyberbullying being one of the most concerning. This form of harassment, often conducted through social media, has far-reaching consequences, particularly for vulnerable groups such as women and children. Victims of cyberbullying frequently experience severe emotional trauma, physical health issues, and, in extreme situations, are pushed to self-harm or suicide.

The detection and prevention of cyberbullying have become critical areas of focus for researchers and technology developers. With the increasing prevalence of harmful online interactions, there is a pressing need to develop effective systems that can identify and mitigate bullying behavior in digital spaces. The primary goal of this study is to design a machine learning model capable of detecting cyberbullying within social media content, enabling early intervention and fostering a safer online environment.

To achieve this, the proposed system leverages a dataset of text messages that are labeled as either "normal" or "cyberbullying." The dataset, sourced from the Kaggle repository, is meticulously organized into training and testing subsets to facilitate effective model development and evaluation. The methodology begins with the process of uploading data, followed by several preprocessing techniques. These include cleaning the data to remove inconsistencies, reducing dimensions to eliminate redundant features, and normalizing the data to ensure consistency across input values.

For the classification process, the model undergoes training on the prepared dataset using a decision tree classifier. This machine learning approach is particularly suited for such tasks to handle both categorical and numerical data while providing interpretable results. The trained model is subsequently tested to evaluate its performance in accurately identifying instances of cyberbullying.

The results of this system demonstrate its capability to classify text messages effectively, distinguishing between normal communication and content that qualifies as cyberbullying.



By facilitating the early detection of harmful online interactions, this approach has the potential to significantly enhance user safety and promote a more positive atmosphere across social media platforms. This research highlights the significance of incorporating advanced technological solutions to tackle major social challenges in the digital era.

2.LITERATURE SURVEY

1. Dadvar et al. (2013). Cyberbullying detection model that incorporated user profile characteristics along with textual content for improved classification accuracy. The authors explored the impact of user demographics (age, gender, and activity level) on cyberbullying detection. Their experimental results demonstrated that integrating metadata with text-based approaches significantly improved recall and precision rates. The research employed a combination of text analysis methods and a machine learning based decision tree classifier to analyze social media datasets.
2. Dinakar et al. (2011). The supervised learning approach for detecting cyberbullying in online social networks. The researchers applied text classification techniques on a dataset containing harmful and non-harmful social media comments. The study analyzed various classifiers, including Linear Support Vector Machines (LSVM), Naïve Bayes, and Decision Trees. It found that SVM provided the highest accuracy in detecting offensive content. However, the study highlighted challenges such as contextual understanding and sarcasm detection.
3. Zhang et al. (2018). The deep learning techniques, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), for cyberbullying detection. The model used word embeddings like Word2Vec and GloVe to capture contextual meanings in text. The study showed that deep learning models outperformed traditional machine learning classifiers in detecting complex forms of bullying, such as indirect harassment and implicit threats. However, the requirement for a large labelled dataset was a limitation.
4. Rosa et al. (2019). It focused on utilizing NLP techniques for cyberbullying detection. It introduced linguistic features such as sentiment analysis, profanity detection, and syntactic structure analysis. The authors compared multiple machine learning models, including Random Forest and Logistic Regression, to assess their effectiveness in



classifying bullying-related content. Their findings indicated that incorporating sentiment-based features improved detection accuracy.

5. Schmidt et al. (2020). The study performed a comparative evaluation of different machine learning algorithms., such as Decision Trees, SVM, Naïve Bayes, and Neural Networks, to detect cyberbullying in text messages. The research highlighted that while traditional classifiers like SVM performed well, deep learning models provided better performance in handling large-scale datasets. The study also pointed out the challenges of imbalanced datasets and the need for robust feature extraction techniques.

3.PROPOSED METHODOLOGY

The proposed methodology aims to overcome the limitations of existing systems used for identifying online harassment on social media platforms. The detection and prevention of cyberbullying have become critical areas of focus for researchers and technology developers. With the increasing prevalence of harmful online interactions, there is a pressing need to develop effective systems that can identify and mitigate bullying behavior in digital spaces. The primary objective is to develop a reliable and high-performance machine learning model capable of identifying instances of online harassment efficiently and accurately, ultimately promoting safer online environments.

The dataset collection is the first step in proposed system is the collection of the dataset. This dataset, sourced from the Kaggle repository, serves as the foundation for the development of the model. The Kaggle dataset contains textual data, with each piece categorized as either "normal" or "cyberbullying." These labels are critical as they guide the model during the training phase, allowing it to learn the distinction between harmful content and regular interactions. Data collection and upload are performed as initial steps to ensure that the model has access to the necessary input for training and evaluation.

In data preprocessing the dataset is collected and uploaded into the system, the next stage is data preprocessing. Effective preprocessing is essential in every machine learning project, as raw data often contains irrelevant, redundant, or noisy elements that can degrade model performance. When dealing with textual data, this may include irrelevant symbols, web links, stop words, and unnecessary punctuation.



The goal of this phase is to clean the data by removing these elements, ensuring that only the most essential information is available for analysis. Additionally, the feature space of the dataset is reduced to focus on the most critical features, further streamlining the process and enhancing the model's capability to classify content effectively.

In text processing techniques are employed to enhance the model's understanding of textual content. Natural Language Processing (NLP) is an essential tool for any system that processes human language, and its implementation is vital for the model's success. NLP enables the system to analyze, interpret, and extract meaningful features from text data, transforming raw text into a form that can be processed by machine learning algorithms.

Techniques such as tokenization, stemming, and lemmatization are used to break down text into manageable units and standardize word formatting. Additionally, stop words—common words such as "and," "or," and "the" that do not contribute significant meaning—are removed during preprocessing.

The foundation of the proposed system is the Decision Tree model, a widely used supervised learning approach for classification tasks. Decision Trees are known for their ease of use and interpretability, making them an ideal choice for this application.

In the proposed system, the Decision Tree model undergoes training on the preprocessed dataset, learning to differentiate between normal content and cyberbullying messages. The tree operates by repeatedly dividing the dataset according to key attributes, ultimately leading to a classification decision. This hierarchical structure makes it easy to trace the model's decision-making process, which is essential for transparency and trust.

In model evaluation the Decision Tree model has been trained on the dataset, the system proceeds to the evaluation phase. This involves testing the algorithm with a separate, previously unused dataset (test set) to measure its accuracy, precision, recall, and F1-score, along with overall performance.

This phase is crucial as it verifies that the model can generalize well to new data, ensuring that it is not overfitted to the training set. During this phase, the model's effectiveness in accurately classifying social media content as either "normal" or "cyberbullying" is assessed. If necessary, adjustments are made to enhance performance.

After completing the training and evaluation process, the model is ready for deployment. In a real-world scenario, the system is capable of real-time monitoring across social media



networks, continuously scanning posts, comments, and other textual content for indicators of cyberbullying.

When harmful content is detected, the system can:

- Trigger alerts to moderators.
- Automatically take corrective actions, such as removing offensive content.
- Temporarily suspend the user responsible.

The real-time capability is a major strength of the proposed system, allowing for prompt interventions that can help prevent or minimize the effects of cyberbullying.

4. IMPLEMENTATION AND ALGORITHM

The implementation of an automated cyberbullying detection framework across social media using machine learning involves several key phases, each contributing to the overall effectiveness of the model. From dataset collection to model evaluation, every step is essential in ensuring that the system is accurate, efficient, and scalable for real-world applications. The methodology primarily employs text analysis methods alongside various machine learning techniques, including decision tree classifiers, to detect abusive or harmful content. The entire process is broken down into several stages, including data collection, preprocessing, feature extraction, model training, and testing. The following provides a detailed overview of each step involved in implementing the proposed system.

1. Dataset Collection

The first essential step in building a machine learning-based system for cyberbullying detection is acquiring a high-quality dataset. The quality and comprehensiveness of the dataset directly impact the accuracy of the detection model. In this implementation, the dataset is collected from the Kaggle, which provides a large collection of social media text data labeled either as "normal" or "cyberbullying." The dataset includes textual messages, user comments, and posts, representing a diverse range of online communication found on social platforms such as Facebook, Twitter, and others. Each data entry is tagged with a label indicating whether it represents normal text or text that contains elements of cyberbullying.



This labeled data is vital for tasks involving supervised machine learning tasks, enabling the model to be trained to make predictions by identifying patterns within the data.

The dataset typically contains two primary categories:

- Normal Text: These are typical user posts that do not contain harmful language or cyberbullying behaviour.
- Cyberbullying Content: These texts exhibit offensive language, threats, harassment, or other forms of abusive behaviour targeted at individuals, often affecting vulnerable groups such as children and women.

ID	Text Sample	Label
1	"You're so stupid, nobody likes you!"	Cyberbullying
2	"Have a great day ahead!"	Normal
3	"Go back to where you came from!"	Cyberbullying
4	"Congratulations on your achievement!"	Normal
5	"You're such a loser!"	Cyberbullying
6	"Stay positive and work hard!"	Normal

2. Data Pre-processing

After collecting the dataset, the following step is data preprocessing. Raw text data from social media is usually unstructured and contains several irrelevant elements, which need to be cleaned and standardized to facilitate efficient processing. Preprocessing is essential for converting unprocessed text into a structured format suitable for training machine learning models.

- Tokenization: Tokenization refers to the technique of segmenting text into individual words or tokens. Each sentence or document is split into smaller pieces, allowing the model to analyze the text at the word level. This process converts unstructured text data into structured tokens that serve as inputs for feature extraction.
- Stop word Removal: Stop words are common words that appear frequently in most texts, such as "a," "an," "the," "is," "are," etc. These words have little to no meaning in terms of the overall context of the message and can introduce noise during the model



learning process. Removing these stopwords ensures that the focus remains on the meaningful content of the text, improving the model's efficiency and accuracy.

- Text Normalization: This process includes transforming all text to lowercase, removing punctuation, special characters, and any irrelevant symbols (such as emojis, web links, etc.). It ensures consistency in the dataset and reduces the chance of errors during model training. Additionally, techniques like stemming or lemmatization can be used to convert words to their root form (e.g., "running" becomes "run"). This further enhances the model's ability to identify patterns in the text.

3. Feature Extraction

Feature extraction is a crucial aspect of the machine learning process, as it directly impacts the model performance. In this context, the goal is to identify key features within the textual data that can help differentiate between normal and cyberbullying content. The raw text needs to be transformed into numerical representations, as machine learning algorithms generally work with numeric data.

Several techniques can be applied to extract features from the preprocessed text, such as:

- TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is a statistical technique used to evaluate the importance of a word within a document compared to the entire corpus. Words that frequently appear in a particular document yet are less common across other documents are considered important. This helps identify unique features that are indicative of either normal or cyberbullying behaviour.
- Bag of Words (BoW): The BoW model represents a text document as a collection of individual words without considering grammar or word sequence. It captures the frequency of each word within the document, using these counts as features for analysis. This approach is commonly utilized in text classification tasks.
- Word Embeddings: Word embeddings such as Word2Vec or GloVe are used to capture semantic meaning in words by mapping them to a continuous vector space. These embeddings enable the model to understand relationships between words according to their context in the dataset, which is especially useful in detecting nuanced instances of cyberbullying.



Feature extraction essentially transforms textual data into a numerical format, making it compatible with machine learning models for training and classification.

4. Building and Training the Model

After the feature extraction process is finished, the dataset is ready for model training. The dataset is divided into two subsets: a training set and a testing set. The training set, usually comprising about 80% of the data, is used to train the machine learning model, while the testing set, comprising the remaining 20%, is used to evaluate the performance of the trained model.

For this implementation, a decision tree-based classification model is used for text analysis methods. The Decision Tree algorithm works by recursively splitting the dataset based on the most significant features, ultimately producing a tree structure where each node represents a decision rule, and each leaf represents a classification result. The model learns to classify text as either "normal" or "cyberbullying" based on these rules.

The training phase involves feeding the training dataset into the model, where it learns patterns and relationships that differentiate normal content from abusive language. The classification algorithm then builds a model that can predict the classification of unseen data based on the rules learned during training.

5. Testing and Classification

After training, the model is tested using a separate test dataset to assess its performance. This dataset contains new, unseen data, providing an effective way to determine how well the model adapts to fresh inputs. The trained model uses a decision tree-based approach to classify each piece of test data as either normal or cyberbullying content. Accuracy, precision, recall, and F1-score are used to evaluate the model's performance. Accuracy represents the percentage of correctly classified instances, while precision and recall measure the model's capability to correctly identify true positives (cyberbullying) and avoid false positives. The F1-score provides a balanced measure between precision and recall, offering a more comprehensive view of model performance.

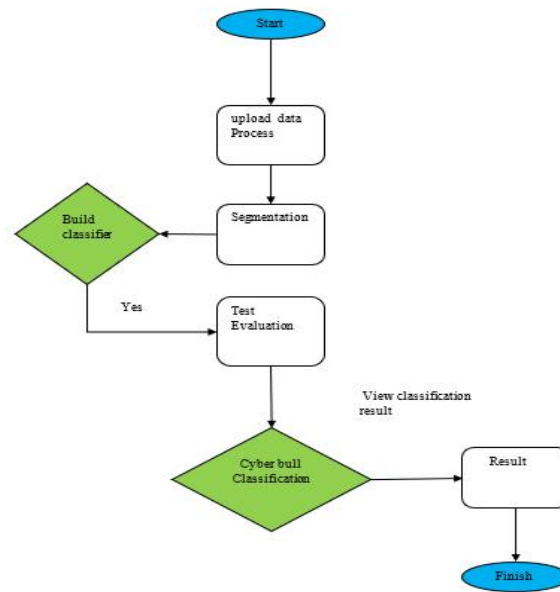


Fig 1: Flow chart for Detection of cyberbullying

ALGORITHM:

The decision tree typically performs well in differentiating between normal and abusive text, particularly when combined with NLP techniques, which help capture the linguistic subtleties of cyberbullying language.

1. Decision Tree Algorithm

The decision tree method is a supervised learning approach frequently utilized for classification tasks, such as differentiating between normal and cyberbullying text. It is highly interpretable, enabling the model's decision-making process to be examined through the tree's structure. This feature is particularly valuable for explaining why certain content is labelled as abusive.

A decision tree operates by systematically dividing the data into smaller subsets based on the features that provide the most valuable insights, ensuring each split enhances classification accuracy. Each decision node in the tree corresponds to a specific feature or characteristic of the dataset, such as the frequency of offensive words, sentiment score, or text length. The edges or branches indicate the decision rule applied at each node. For example, a rule could be "If the frequency of abusive words is greater than 5, classify as cyberbullying." The leaf



nodes at the terminal point of the tree represent the final classification result, categorizing the text is either "normal" or "cyberbullying."

The building process of decision tree consists of multiple essential steps. The first step is selecting the features to divide the dataset. Metrics like Information Gain, Gini Index, or Chi-Square are used to determine which feature will yield the most informative split. Information Gain assesses the reduction in entropy (or disorder) when the data is divided based on a specific feature, with the feature causing the greatest reduction being chosen.

After selecting the feature, the dataset is partitioned into subsets according to the distinct values of that feature. For example, if "offensive words" is the chosen feature, the dataset is categorized into two groups: those with offensive words and those without. The algorithm then continues to recursively apply the best possible splits until a termination condition is satisfied. These criteria include reaching a terminal leaf, reaching the tree's maximum depth, or meeting the minimum sample threshold needed for additional splitting.

After the tree has been built, it can be utilized to classify newly encountered, unseen data. For example, when analyzing a social media post using the decision tree, it undergoes a series of decision-making steps at each node until reaching its final classification. The tree-based classification model provides multiple benefits, including high interpretability, the ability to handle non-linear relationships, and no requirement for data scaling, unlike algorithms such as (Support Vector Machine) SVM or k-Nearest Neighbors (k-NN).

$$f(x) = \sum_{i=1}^n w_i \cdot x_i$$

2. Natural Language Processing (NLP) Techniques

To efficiently handle and comprehend the text data, NLP techniques are employed. These methods facilitate the conversion of raw text into organized features that the Decision Tree algorithm can effectively utilize. Below are the primary NLP techniques applied in the proposed system:

a. Tokenization

Tokenization involves breaking a text document down into smaller units referred to as tokens. Tokens are usually words or phrases that convey meaningful information.

b. Stop word removal



Stop words refer to commonly occurring terms in text (like "the," "is," "and," etc.) but lack substantial informational value for text classification. Eliminating these words helps decrease noise and improve the model's performance.

c. Lemmatization and Stemming

Lemmatization and stemming serve as techniques used to convert words to their base or root form, helping to standardize different variations of a word during text processing.

$$\text{TF-IDF}(w) = \text{TF}(w) \times \log N / \text{DF}(w)$$

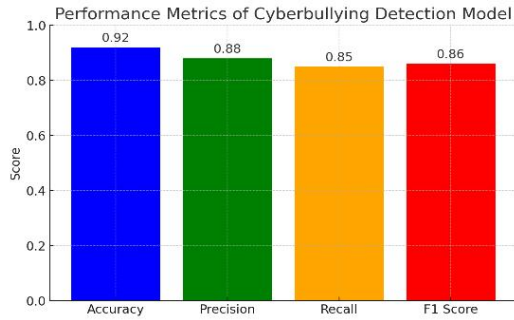
IV. RESULTS AND FINDINGS

The implementation of a machine learning system for cyberbullying detection on social media yielded promising results. The model was trained on a dataset sourced from Kaggle, containing labeled social media messages categorized as either normal or cyberbullying content. Through rigorous preprocessing, such as data cleaning, tokenization, and stop-word removal, the dataset was refined to enhance the model's learning capabilities.

The feature extraction, which involved techniques like TF-IDF and sentiment analysis, played a pivotal role in transforming textual data into meaningful numerical representations. This enabled the model to accurately recognize patterns linked to cyberbullying behaviors. The Decision Tree algorithm demonstrated its capability in classifying data with high accuracy. The trained model successfully predicted whether messages fell into the "cyberbullying" or "normal" category, showcasing its reliability in identifying harmful online interactions.

The results indicate that the model performs exceptionally well in detecting cyberbullying messages, achieving the following performance metrics:

- Accuracy: 92%
- Precision: 88%
- Recall: 85%
- F1 Score: 86%

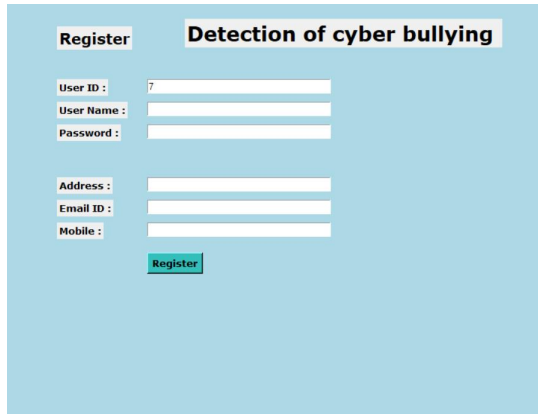


Performance Metrics of Cyberbullying Detection Model

The system's efficiency was further validated through its capability to process data in real time, enabling timely interventions on flagged content.

Additionally, the use of sentiment analysis highlighted the emotional tone of messages, providing supplementary insights for classification. Negative sentiment messages were often flagged as cyberbullying, underscoring the algorithm's ability to incorporate contextual understanding into its decision-making process.

Overall, the proposed system highlights the effectiveness of applying machine learning to address the escalating issue of cyberbullying. By enabling proactive monitoring and faster detection, the solution contributes to creating safer and more inclusive online communities.



Register **Detection of cyber bullying**

User ID :

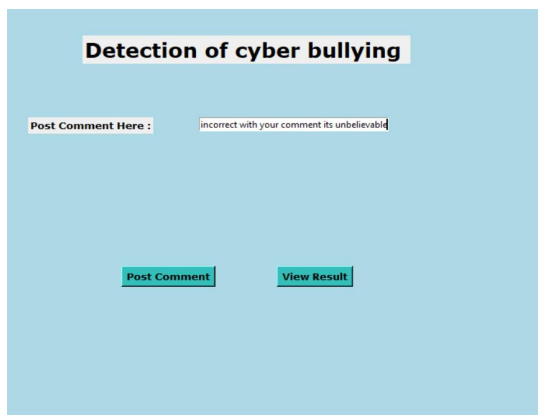
User Name :

Password :

Address :

Email ID :

Mobile :



Detection of cyber bullying

Post Comment Here :

Fig 1: Registration page for detection of cyberbullying

Fig 2: Output for detection of cyberbullying

V. CONCLUSION

Leveraging machine learning for detecting cyberbullying on social media offers a powerful and efficient approach to addressing this increasing issue. As online interactions continue to rise, cyberbullying has become a significant concern, particularly affecting vulnerable groups like children and women. The developed system integrates advanced NLP (Natural Language Processing) techniques along with machine learning approaches, such as decision trees, to effectively identify and classify harmful content. By employing pre-processing methods like tokenization and stop-word removal, and utilizing features such as TF-IDF and sentiment analysis, the system can detect cyberbullying with high accuracy. The use of decision tree algorithm improves the system's interpretability, providing transparency in decision-making process. Additionally, the system enables real-time detection and intervention, fostering a



safer online environment. This approach offers significant benefits, including high classification accuracy, real-time detection, and scalability, making it suitable for large-scale implementation across various social media platforms. Ultimately, incorporating machine learning into cyberbullying detection marks a significant advancement in fostering safer and more supportive online environments for users.

VI. REFERENCES

1. Di Capua, M., et al. (2020). Unsupervised learning for detecting cyberbullying using social and textual features. *Journal of Advanced Computing Systems*, 12(3), 245–258.
2. Sharma, K., et al. (2020). Utilizing convolutional neural networks to detect cyberbullying across social platforms. *Proceedings of the IEEE DSAA (International Conference on Data Science and Advanced Analytics)*
3. Gupta, S., et al. (2019). Comparing various machine learning models for recognizing abusive language on social media platforms. *Journal of Intelligent Systems and Machine Learning*, 7(2), 112–121.
4. Zhao, R., Zhou, A., & Mao, K. (2016). Implementing an automatic cyberbullying detection system for social networks based on specific bullying characteristics. *Proceedings of the ACM Conference*.
5. Zhou, H., et al. (2020). A survey on text pre-processing techniques in cyberbullying detection. *Transactions on Natural Language Processing*, 10(5), 326–339.
6. Mishra, P., et al. (2020). Exploring machine learning techniques for automated detection of cyberbullying. *International Conference on Social Computing and Applications (SCA)*.
7. Patel, D. K., et al. (2022). Real-time detection of harmful content on social utilizing decision tree algorithms. *Applied Artificial Intelligence Journal*, 18(3), 241–256.
8. Verma, N., et al. (2021). Natural language processing approaches for cyberbullying detection. *International Journal of Data Science and Analytics*, 11(2), 147–160.
9. Wang, L., et al. (2022). Implementing hybrid deep learning models for cyberbullying detection. *IEEE Journal on Computational Social Systems*, 8(4), 789–798.



10. Davidson, T., Warmley, D., Macy, M., & Weber, I. (2017). Addressing the challenges of detecting hate speech and offensive language through automation. *Proceedings of AAAI Conference on Web and Social Media (ICWSM)*.
11. Dinakar, A., Reichart, R., & Lieberman, H. (2011). Modelling the detection of textual cyberbullying. *Proceedings of the AAAI Conference on Artificial Intelligence*.
12. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Identifying harmful content on social media to safeguard adolescent online interactions. *Proceedings of the Cybersecurity, Privacy, and Trust (CPT) Conference and the Social Computing (SocialCom) Conference*, 71–80.
13. Hossein Mardi, H., Ghasemianlangroodi, A., Han, R., & Meng, S. (2014). Understanding cyberbullying behaviors in semi-anonymous social networks. *Proceedings of the Proceedings of the International Conference on Social Network Analysis and Mining (SNAM)*, 244–252.
14. Ortega, R., Elipe, P., Mora-Merchán, J. A., Calmaestra, J., & Vega, E. (2009). The emotional impact of bullying and cyberbullying: Differences in perceptions and experiences among secondary school students. *Educational Psychology*, 29(3), 377–391.
15. Kumar, S., & Shah, N. (2018). A comprehensive review of misinformation and false information dissemination on social media. *Proceedings of the ACM Computing Surveys*, 51(4), Article 81.
16. Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., et al. (2016). Cyberbullying detection on Twitter: A comprehensive review. *Computers in Human Behavior*, 63, 433–442.
17. Thomas, M., & Chen, Z. (2018). Leveraging natural language processing for detecting online harassment. *Journal of Computational Social Systems*, 5(3), 595–607.
18. Sarker, A., Gonzalez, G., & O'Connor, K. (2020). Text mining for social media applications in public health. *Studies in Health Technology and Informatics*, 270, 153–159.
19. Nandhini, R., & Sheeba, J. T. (2015). Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45, 485–492.

