IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 October 2023, accepted 22 October 2023, date of publication 2 November 2023, date of current version 9 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329347

RESEARCH ARTICLE

Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection

NAGWAN ABDEL SAMEE[1], UMAIR KHAN[2], SALABAT KHAN[3,4,5], MONA M. JAMJOOM[6], MUHAMMAD SHARIF[3], AND DO HYUEN KIM[4]

[1]Department of Information Technology, College of Computer and Information Science, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[2]Department of Computer Science, Air University, Islamabad, Aerospace and Aviation Campus, Kamra, Pakistan
[3]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Punjab 43600, Pakistan
[4]Department of Computer Engineering, Jeju National University, Jeju-si, Jeju Special Self-Governing Province 63243, Republic of Korea
[5]Big Data Research Center, Jeju National University, Jeju 63243, Republic of Korea
[6]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Salabat Khan (salabat.khan@jejunu.edu.pk)

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The work of Salabat Khan was supported in part by the National Research Foundation of Korea (NRF) through the Brain Pool Program under Grant 2022H1D3A2A02055024, and in part by the Creative Research Project under Grant RS-2023-00248526.

ABSTRACT Cyberbullying has emerged as a pervasive issue in the digital age, necessitating advanced techniques for effective detection and mitigation. This research explores the integration of word embeddings, emotional features, and federated learning to address the challenges of centralized data processing and user privacy concerns prevalent in previous methods. Word embeddings capture semantic relationships and contextual information, enabling a more nuanced understanding of text data, while emotional features derived from text extend the analysis to encompass the affective dimension, enhancing cyberbullying identification. Federated learning, a decentralized learning paradigm, offers a compelling solution to centralizing sensitive user data by enabling collaborative model training across distributed devices, preserving privacy while harnessing collective intelligence. In this study, we conduct an in-depth investigation into the fusion of word embeddings, emotional features, and federated learning, complemented by the utilization of BERT, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), and Long Short-Term Memory (LSTM) models. Hyperparameters and neural architecture are explored to find optimal configurations, leading to the generation of superior results. These techniques are applied in the context of cyberbullying detection, using publicly available multi-platform (social media) cyberbullying datasets. Through extensive experiments and evaluations, our proposed framework demonstrates superior performance and robustness compared to traditional methods. The results illustrate the enhanced ability to identify and combat cyberbullying incidents effectively, contributing to the creation of safer online environments. Particularly, the BERT model consistently outperforms other deep learning models (CNN, DNN, LSTM) in cyberbullying detection while preserving the privacy of local datasets for each social platform through our improved federated learning setup. We have provided Differential Privacy based security analysis for the proposed method to further strengthen the privacy and robustness of the system. By leveraging word embeddings, emotional features, and federated learning, this research opens new avenues in cyberbullying research, paving the way for proactive intervention and support mechanisms. The comprehensive approach presented herein highlights the substantial strengths and advantages of this integrated methodology, setting a foundation for future advancements in cyberbullying detection and mitigation.

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

INDEX TERMS Cyberbullying detection, federated learning, multi-platforms privacy preservation, decentralized edge intelligence, hyper-parameter optimization, neural architecture search.

I. INTRODUCTION

Text analysis, a vital field within natural language processing (NLP), plays a crucial role in extracting meaningful insights and valuable information from extensive amounts of textual data. As digital communication and the utilization of social media platforms continue to grow exponentially, the importance of text analysis in understanding human behavior, sentiment, and discourse patterns has significantly increased. This surge in demand for text analysis techniques is driven by the availability of massive textual data and the development of advanced machine learning algorithms. These techniques find applications in various domains, such as sentiment analysis, topic modeling, information retrieval, and more. Through the application of these methods, researchers and practitioners can uncover valuable insights, patterns, and trends embedded within textual data.

In today's digital society, cyberbullying has emerged as a prevalent issue, referring to the intentional and repetitive use of digital communication platforms to harass, intimidate, or harm individuals. Cyberbullying encompasses a wide range of harmful behaviors, including the dissemination of rumors, sharing explicit or defamatory content, sending abusive messages, and engaging in online hate speech [1]. The proliferation of digital platforms, such as social media, online forums, and messaging applications, has provided individuals with unprecedented means of communication and expression. However, it has also created breeding grounds for cyberbullying [2], enabling perpetrators to target their victims anonymously or under false identities, exacerbating the detrimental effects of their actions. The negative impact of cyberbullying on individuals, particularly their mental health, social interactions, and overall well-being, cannot be overstated [3]. Victims often experience heightened levels of stress, anxiety, depression, and decreased self-esteem [4], [5], [6]. The persistent nature of online harassment can lead to social isolation, strained relationships, and hindered academic or professional performance. In severe cases, cyberbullying has even been linked to self-harm and suicidal ideation among victims [7]. The purpose of this research paper is to contribute to the existing body of knowledge on cyberbullying by proposing a novel approach to its detection. By leveraging the power of federated learning and text analysis techniques, we aim to develop a robust and privacy-preserving framework that can effectively identify and combat instances of cyberbullying while upholding user privacy. Our research aligns with the overarching objective of creating safer online environments and promoting the well-being and mental health of individuals in the digital era. By combining federated learning, which allows for collaborative model training while preserving data privacy, with advanced text analysis techniques, we seek to empower institutions and individuals to collectively address the pressing issue of cyberbullying.

In previous works on cyberbullying detection, a major limitation lies in the centralization of sensitive user data for model training, posing significant privacy concerns and hindering the willingness of individuals to share their personal information. Moreover, traditional methods often struggle to encompass a diverse range of user behavior and language patterns, leading to suboptimal model performance. These shortcomings are effectively addressed through the implementation of federated learning in our research. By adopting a decentralized approach, federated learning [8] ensures that user data remains securely stored on individual devices, preserving the confidentiality of personal conversations and online activities. This privacy-preserving methodology encourages greater participation from users and institutions, leading to a more comprehensive and representative dataset. Additionally, the iterative nature of federated learning allows for continuous model refinement without compromising data privacy, enabling the global model to better understand cyberbullying patterns and linguistic cues from a diverse range of sources. As a result, our research leverages the strengths of federated learning to overcome the limitations of previous methods, offering a groundbreaking approach to cyberbullying detection that upholds the utmost level of data privacy and security. Our research paper follows a well-defined federated learning process sequence of key steps, meticulously designed to ensure both privacy and the development of effective cyberbullying detection models. This process commences with the initialization of a central model, which serves as the initial foundation for collaborative training endeavors. Subsequently, each participating user or device engages in model training utilizing their localized data. This decentralized training occurs exclusively on the user's device, guaranteeing that sensitive information remains securely stored and never leaves the confines of their respective environment. By adopting this decentralized approach, we affirm our commitment to maintaining the confidentiality of personal conversations, online activities, and private details throughout the federated learning process. Once the local model training reaches completion, the updated models from individual devices are aggregated using FedAvg to create a comprehensive global model. However, it is vital to emphasize that the aggregation process itself does not involve direct sharing or exposure of the individual models. Instead, only the model updates are securely transmitted to the central server. These transmitted updates are intelligently combined to effectively update the global model, ensuring the privacy and confidentiality of each individual user's data throughout the entire federated learning process. By meticulously following this privacy-preserving

federated learning approach, we empower institutions and users to collectively contribute to the development of robust cyberbullying detection models while upholding the utmost level of data privacy and security. The iterative nature of federated learning facilitates continuous refinement of the global model in our research on cyberbullying detection. By repeatedly conducting local model training and model aggregation, we enable the global model to enhance its understanding of cyberbullying patterns and linguistic cues while prioritizing user privacy. This approach empowers individuals to contribute to the detection process without compromising their personal information and communication data. Figure 1 shows the federated learning process.

In our experiments, we focused on optimizing the Federated Learning setup for cyberbullying detection by tuning the configuration and hyperparameters of the models used, including CNN, DNN, LSTM, and BERT. We conducted extensive tests to strike the right balance between efficiency and effectiveness in the training process. For CNN, DNN, and LSTM models, we experimented with various hyperparameters, such as learning rate, batch size, and the number of layers. By fine-tuning these hyperparameters, we aimed to achieve improved convergence and model performance. Additionally, we explored different architectures for these models to identify the most suitable configurations for cyberbullying detection. With BERT, being a pre-trained model, we focused on fine-tuning its parameters on our cyberbullying dataset. We optimized the learning rate and batch size specific to BERT to maximize its precision in identifying cyberbullying instances accurately. Our research paper contributes to cyberbullying detection by introducing eight novel emotional features from textual tweets. We empower collaborative detection through privacy-preserving federated learning, leverage BERT for improved precision, and optimize global model selection using a normal distribution-based approach. Evaluating with multiple clients validates the effectiveness of our framework in identifying cyberbullying instances. These findings provide practical guidance for configuring federated learning in the context of cyberbullying detection, optimizing the trade-off between efficiency and effectiveness. By applying these insights, researchers and practitioners can develop more efficient and accurate cyberbullying detection models while ensuring the utmost privacy and security for individual users participating in the federated learning process.

The contributions are as follows:

1) We introduce eight novel emotional features extracted from textual tweets, enhancing the identification of cyberbullying instances.
2) Empowered collaborative cyberbullying detection through privacy-preserving federated learning.
3) Utilized a normal distribution-based best client selection method for optimal global model selection.
4) Leveraged the powerful BERT model for improved precision in identifying cyberbullying instances.
5) Tested framework with multiple clients, evaluating local and global model performance.
6) We have provided a Differential Privacy-based security analysis for the proposed method to further strengthen the privacy and robustness of the system.

A. MOTIVATION

Text classification faces significant challenges, including data privacy and security concerns. Sharing sensitive text data from multiple sources can raise privacy issues, especially for personal or confidential information. Additionally, Data distribution across various platforms makes it challenging to consolidate and train models effectively. Moreover, the resource-intensive nature of deep learning models for text classification can be a limitation for smaller organizations or devices. Lastly, the dynamic and evolving nature of online text data requires models to adapt quickly, which may be difficult for traditional centralized approaches. The proposed method offers compelling solutions to these challenges. It addresses data privacy concerns by enabling model training on decentralized data sources without sharing raw data, preserving individual privacy. Data distribution across platforms is accommodated as Federated Learning operates in a decentralized manner, fostering collaboration among sources. This approach also resolves resource constraints by distributing the computational load, making it suitable for organizations with varying resources. Furthermore, the method adaptability ensures that models stay up-to-date with the evolving nature of online text data, making it a promising solution for text classification in dynamic and privacy-sensitive environments.

II. RELATED WORK

The rise of Internet 2.0 technology has significantly impacted society, with social media platforms like Twitter and Facebook playing a pivotal role in transforming various aspects of human life [9], [10], [11]. These platforms have become integrated into daily activities such as education, business, entertainment, and e-government. As projected by [10], by April 2023, there were 5.18 billion internet users worldwide, with 4.8 billion actively engaging in social media platforms. Among the multitude of social networks, Twitter stands out as a critical platform and an invaluable data source for researchers. With its real-time and public microblogging nature, Twitter often breaks news even before official sources. The platform's short message limit (currently 4000 characters) and unfiltered feed have contributed to its rapid growth, witnessing an average of 500 million daily tweets, particularly during events [10]. Undoubtedly, social media has become an integral part of daily life. However, it is essential to acknowledge that the usage of technology, including social media, by young individuals can expose them to various behavioral and psychological risks. One prominent risk is cyberbullying, a pervasive social attack that takes place on social media platforms. The implications of cyberbullying on mental health are substantial, including the development of

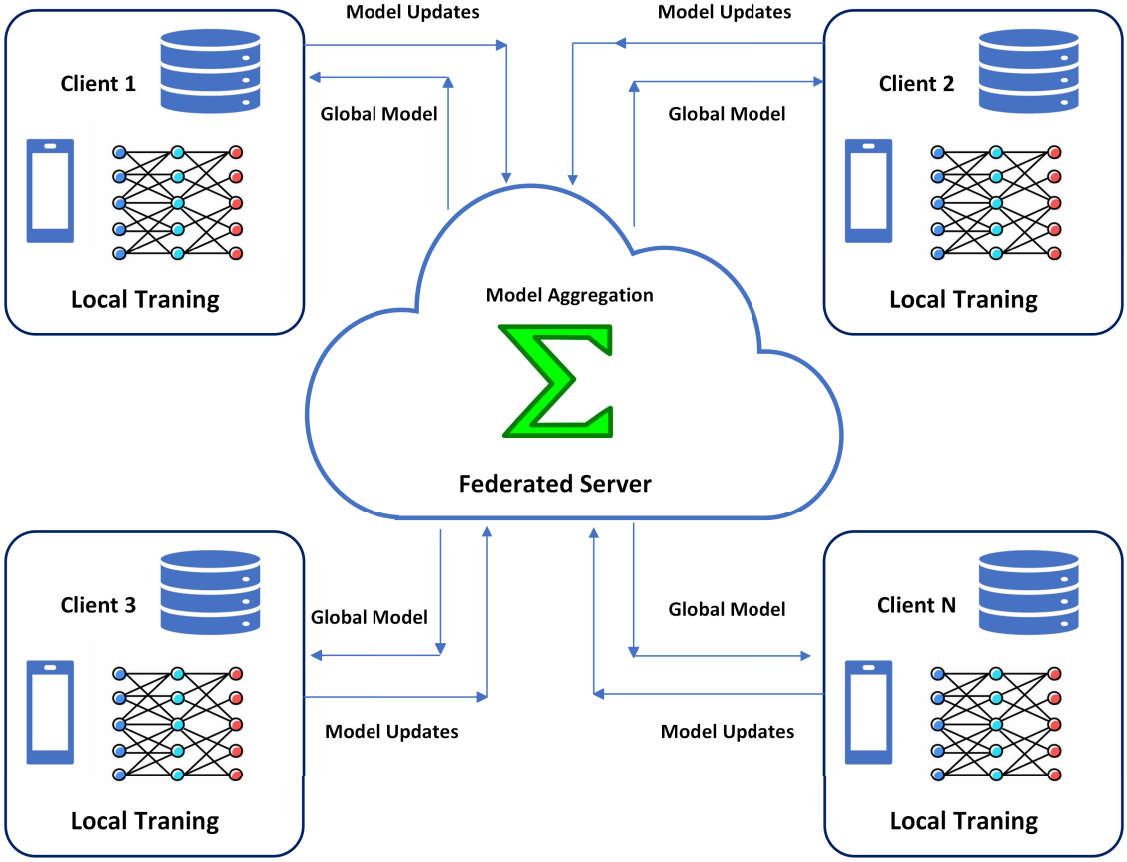

FIGURE 1. Cyberbullying detection in federated learning environment with 'N' clients.

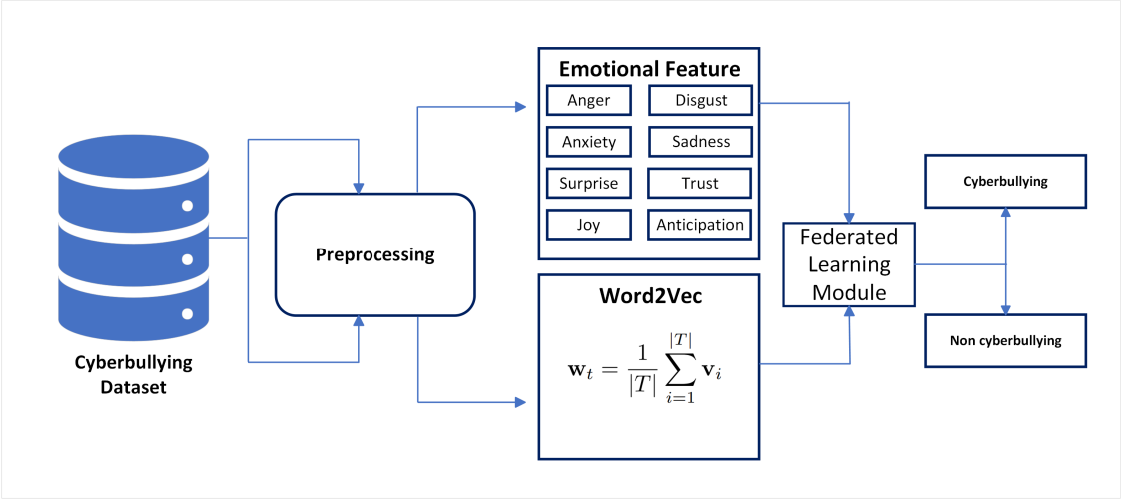

FIGURE 2. Emotional and Word2Vec feature extraction for cyberbullying detection framework.

depression, anxiety, self-harm, suicidal thoughts, attempted suicide, and social and emotional difficulties [7], [12], [13], [14]. Currently, there are global initiatives dedicated to preventing cyberbullying and enhancing internet safety, particularly for vulnerable groups such as children [15], [16]. The literature encompasses numerous intervention and prevention approaches rooted in psychology and education, although their implementation remains limited on a global scale. Complicating matters, victims of cyberbullying often hesitate to confide in parents [17], teachers [18], or other

adults [19]. Instead, they spend substantial amounts of time online [20], seek anonymous support [21], and express their need for information and assistance through online platforms [22]. Recognizing the significance of the internet as a medium, web-based approaches offer an effective means of delivering cyberbullying interventions, accessible at the convenience of the individual [23]. Notable initiatives include the University of Turku's Kiva program in Finland [24], the Anti-Harassment campaign in [25] France, and the anti-cyberbully initiative led by the Belgian government [26]. Throughout our pursuit to combat cyberbullying, our research has extensively explored various approaches, broadly categorized under three key headings: machine learning, deep learning, and transfer learning. These diverse methodologies have provided valuable insights and paved the way for our innovative contributions in enhancing cyberbullying detection.

A. MACHINE LEARNING FOR CYBERBULLYING DETECTION

A multitude of researchers have employed machine learning techniques in the field of cyberbullying detection. This consensus is further supported by a comprehensive analysis of the existing literature, which consistently highlights the effectiveness of Support Vector Machines (SVM) for this purpose [27], [28], [29], [30], [31], [32], [33], [34]. Furthermore, other researchers Raisi and Huang [35] took a different approach, and introduced a model designed to identify offensive comments on social networks, aiming to filter or notify the relevant parties. Their approach involved training the model using offensive word-containing comments extracted from Twitter and Ask.fm datasets. In a similar vein, other researchers [36], [37] developed communication systems employing intelligent agents to offer victims of cyberbullying emotional support and assistance.

B. DEEP LEARNING FOR CYBERBULLYING DETECTION

In recent years, the field of cyberbullying detection has witnessed a surge in interest in deep learning models. This increased attention can be attributed to the remarkable performance exhibited by deep learning approaches in addressing this issue. Researchers such as Zhao and Mao [38] extended the deep learning model smSDA to uncover hidden features within cyberbullying posts and learn robust and discriminative text representations. Zhang et al. [39] introduced the pronunciation-based convolutional neural network (PCNN), designed to handle spelling errors while maintaining accurate pronunciation, resulting in improved performance compared to other neural network models. Kumar and Sachdeva [40]developed the hybrid deep learning framework Bi-GAC, which combines bi-GRU self-attention encoding and capsule networks to capture semantic representations and spatial information in social media texts. Shriniket et al. [40] proposed the CNNSemi Trained GloVe model, which integrates semantic word embeddings with CNN for efficient cyberbullying detection with high prediction accuracy. Agrawal and Awekar [41] conducted extensive experiments comparing conventional machine learning models and deep neural network models, demonstrating the superior performance of deep learning models across multiple social platforms. Dadvar and Eckert [42] explored deep neural network-based models, including CNN, LSTM, BLSTM, and attention-based BLSTM, and showcased their superiority over conventional machine learning models using a YouTube dataset. These studies collectively highlight the advancements in deep learning approaches for cyberbullying detection, offering enhanced accuracy and performance compared to traditional machine learning methods.

C. TRANSFER LEARNING FOR CYBERBULLYING DETECTION

Adopting transfer learning with language models has shown promise in cyberbullying detection studies. Recent research has leveraged models like Bert and RoBerta to achieve improved accuracy. Paul and Saha [43]developed a cyberbullying detection model based on Bert with optimistic results. Jacobs et al. fine-tuned RoBerta for role classification, achieving the best performance [44]. Fatma et al. [45] demonstrated the superiority of fine-tuning Bert for cyberbullying detection over other deep-learning models. Verma et al. [46] fine-tuned the Hate-BERT model from HuggingFace's Transformer library, surpassing traditional models such as BiLSTM and SVM. Bhatia et al. [47] found that fine-tuning Bert with preprocessed data, including a slang-abusive corpus, significantly enhanced its performance. These studies showcase the potential of transfer learning with language models for effective cyberbullying detection.

Machine learning and deep learning techniques have shown promise in cyberbullying detection, but they face limitations. Traditional machine-learning approaches struggle with the centralized collection and analysis of distributed user data, hindering their ability to capture the diverse characteristics of cyberbullying across different platforms. Deep learning models require large amounts of labeled data, which can be challenging to obtain and share in privacy-sensitive federated settings. To overcome these limitations, federated learning offers a decentralized approach where models are trained collaboratively on local data, without the need to share sensitive information. By aggregating locally trained models, federated learning enables the collective intelligence of distributed devices while preserving privacy. This approach addresses the limitations of centralized data collection, labeled data availability, communication overhead, and biases in the dataset. Leveraging federated learning in cyberbullying detection allows for scalable, privacy-preserving models that can capture the nuances of cyberbullying across multiple platforms and diverse user populations.

D. FEDERATED LEANRING FOR CYBERBULLYING

In contrast to previous centralized methods for cyberbullying detection, Chehbouni et al. [48] introduces a privacy-preserving federated learning framework, leveraging

a BERT model for early detection of Sexual Predators. However, it's crucial to acknowledge a limitation regarding the potential biases that such a model may inherit, including racial and gender biases encountered during training. Additionally, the high cost of false accusations, especially when employing large pre-trained models, must be carefully considered within the context of online grooming detection. Nisha et al. [49] introduces 'FedBully,' a cyberbullying detection approach employing sentence encoders, including the BERT model, for feature extraction. In response to the rising concerns of hate speech and cyberbullying on social media, Ram et al. [50] introduce 'Schat,' an End-to-End messaging system equipped with Natural Language Processing (NLP) to combat offensive text. Recognizing the subjectivity of offensiveness, they leverage federated learning to develop a privacy-preserving approach, allowing model training without compromising user data privacy, and achieving strong performance in real-world offensive text detection.

III. FEDERATED LEARNING

In this section, we have discussed the significant role of Federated Learning in addressing privacy concerns and enhancing cyberbullying detection. Traditional machine learning approaches raise privacy issues due to centralized data collection, while Federated Learning offers a decentralized and privacy-preserving alternative. By distributing the learning process to local devices and edge nodes, Federated Learning ensures that sensitive user data remains secure and inaccessible to the central server. We will explore the key components of the Federated Learning framework, its advantages in capturing diverse cyberbullying patterns, and how it fosters user participation while maintaining data privacy. In traditional machine learning, data from various sources are collected and combined on a central server for model training. This approach raises privacy issues [51] as sensitive user data is exposed and vulnerable to breaches. Federated Learning, on the other hand, enables collaborative model training while keeping the data decentralized and preserving user privacy. Federated Learning is a decentralized machine learning paradigm [52], [53] that enables collaborative training of models across multiple data sources while preserving the privacy of individual data owners. Unlike conventional approaches that rely on aggregating data into a centralized server, Federated Learning distributes the learning process to local devices or edge nodes, ensuring that sensitive data remains on users' devices and only aggregated updates are shared with the central server [54]. This decentralized approach has gained attention for its ability to address privacy concerns, enabling the development of robust and accurate models while respecting user privacy. The FL task by the server is to minimize the global loss function denoted by:

$$\min_{\theta} \sum_{k=1}^{K} \frac{n_k}{n} \left(\mathbb{E}_{(x,y)\sim D_k} \left[\ell(f_\theta(x), y) \right] \right) \tag{1}$$

In the above equation, θ represents the model parameters to be optimized. The objective is to minimize the sum of the loss function, denoted as ℓ, across all clients. The expectation operator $\mathbb{E}$ calculates the expected loss over the data distribution Dk of each client. The fraction nk/n represents the proportion of samples contributed by client k to the total sample size n.

The Federated Learning framework consists of three main entities [55]: the central server, clients, and a global model. The central server coordinates the overall training process but does not have access to the raw data. Each client, such as a user's device or an edge node, possesses its own local data and contributes to the model training without sharing raw data. The global model is initialized on the central server and serves as a starting point for training. The training process in Federated Learning is divided into several rounds, with each round consisting of three main steps: client selection, local model training, and model aggregation.

After the local model training, the model updates from each client are aggregated on the central server to form a global model. Various aggregation techniques can be used, such as Federated Averaging [56], which takes the weighted average of the model updates from different clients. The aggregated global model is then sent back to the clients, serving as the new starting point for the next round. The decentralized nature of Federated Learning provides several advantages for addressing cyberbullying detection. By training models on diverse data sources, Federated Learning can capture the inherent heterogeneity of cyberbullying patterns across different platforms and user demographics. This enhances the model's ability to generalize and detect cyberbullying incidents effectively. Furthermore, Federated Learning ensures privacy preservation by keeping the user data localized. The raw data never leaves the client's device, reducing the risk of data breaches or unauthorized access. This privacy-centric approach encourages user participation and fosters a more secure and trustworthy environment for individuals to engage in cyberbullying prevention efforts.

IV. METHODOLOGY

A. DATASET

The dataset used in this research paper was obtained from the publicly available resource Kaggle, It contains data from social media platforms such as Kaggle [57], Twitter, Wikipedia Talk pages, and YouTube. The dataset consists of textual data along with corresponding labels that indicate whether the text belongs to the category of cyberbullying or non-cyberbullying. However, it is important to acknowledge that the dataset exhibits an imbalance between the two classes, with a proportion of 1,01,082 for non-cyberbullying instances and 14,782 cyberbullying instances. Imbalanced datasets can pose challenges in training classification models, as the model tends to be biased towards the majority class, resulting in suboptimal performance in detecting the minority class.

TABLE 1. Clients data distribution.

	Samples	Cyberbullying	Non-Cyberbullying
Client 1	8303	3805	4498
Client 2	8303	3666	4637
Client 3	8303	3651	4652
Client 4	8306	3660	4646

FIGURE 3. Word cloud representation of the dataset.

To address this class imbalance, the undersampling method was employed. Undersampling involves reducing the number of instances in the majority class (non-cyberbullying) to create a more balanced distribution between the classes, ensuring that both cyberbullying and non-cyberbullying instances are adequately represented in the dataset. This approach helps to mitigate the bias towards the majority class and promotes better model generalization by considering both classes equally during training. The resulting balanced dataset provided a solid foundation for training and evaluating the cyberbullying detection models in a fair and unbiased manner. Figure 3 represents the word cloud representation of the dataset used in our research. Word cloud uncover the prominent words and phrases within our dataset. Table 1 represents the data distribution for clients after balancing the dataset. The data distribution, as depicted in the Figure 4, illustrates our initial step of balancing the dataset, followed by subsequent segmentation into specific client groups. The dataset is evenly partitioned into local datasets for the given number of clients. For each client, 80% of the local dataset is used for training the model and 20% is used as testing set. For a fair comparison of all the deep learning models, the dataset is partitioned once among the clients randomly and the same set of training and testing sets are used to evaluate all the models (meaning that the local instance numbers are same for all the models for a fair comparison).

B. PREPROCESSING

Cyberbullying detection involves text preprocessing that helps to standardize and enhance the quality of the text data, enabling more accurate and effective analysis. In our research paper on cyberbullying detection, we employ various text preprocessing steps as shown in Figure 5 to enable effective analysis of online abusive behavior. Several preprocessing steps are employed for our research, including tokenization to break the text into smaller units, case normalization to eliminate case sensitivity, removal of punctuation marks to reduce noise, stopword removal to filter out common words, lemmatization and stemming to reduce word variations, handling of noisy text such as HTML tags and special characters. These preprocessing steps collectively improve the performance of the classification model and contribute to a more comprehensive understanding of cyberbullying behavior. For Equation given below, Preprocess() is implemented as a function. Algorithm 1 takes the dataset X as input and for each sample in the dataset, it performs all the pre-processing steps and returns the pre-processed data for further usage.

Algorithm 1 Proposed algorithm for text Preprocessing

Input: A dataset of text samples
Output: Pre-ProcessedDataset
procedurePreprocesssample
  for each sample in dataset **do**
    sample ← Lowercase(*sample*)
    tokens ← Tokenize(*sample*)
    tokens ← RemoveHTMLtags(*tokens*)
    tokens ← RemoveURL(*tokens*)
    tokens ← RemoveNumericalValues(*tokens*)
    tokens ← LemmatizeAndStem(*tokens*)
    tokens ← RemovePunctuation(*tokens*)
    tokens ← RemoveStopwords(*tokens*)
    tokens ← RemoveEmoji(*tokens*)
    tokens ← JoinTokens(*tokens*)
    Append *tokens* to $Pre - ProcessedDataset$
  end for
  return $Pre - ProcessedDataset$
end procedure

$$X_{\text{preprocessed}} = \text{Preprocess}(X) \tag{2}$$

where $X_{\text{preprocessed}}$ is the preprocessed text data. Preprocess function encompasses the detailed text preprocessing steps, including lowercasing, tokenization, punctuation removal, stopwords removal, stemming/lemmatization, numerical value removal, and token joining, as previously described.

In order to improve the representation of data for cyberbullying detection, researchers have explored various techniques such as TF-IDF, BOW, n-grams, skip-grams, and

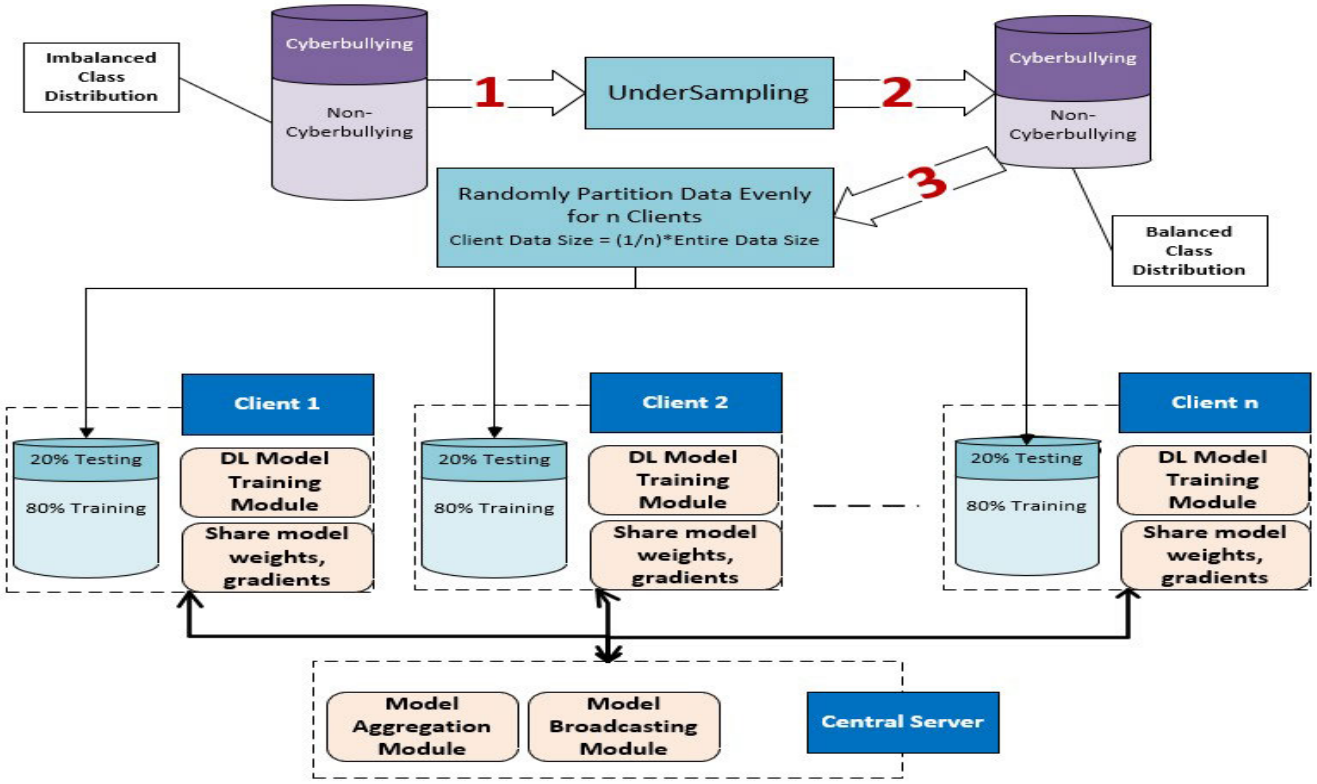

FIGURE 4. Federated learning data distribution among clients.

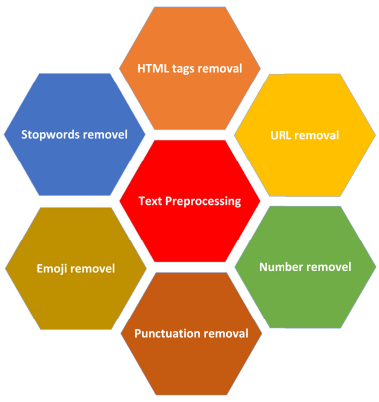

FIGURE 5. Text preprocessing steps.

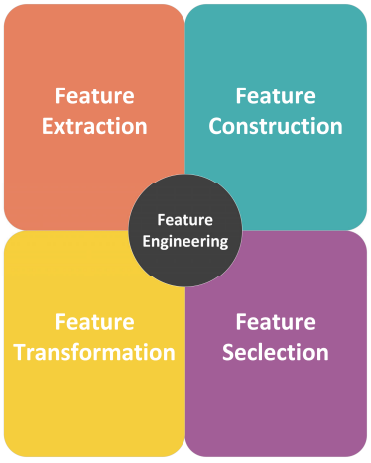

FIGURE 6. Feature engineering steps.

word2vec [58], [59], [60], [61], [62], [63], [64]. Additionally, feature engineering has been employed to incorporate relevant contextual information and linguistic patterns. This section focuses on the use of feature engineering techniques, specifically word2vec [65] for embedding feature extraction and the incorporation of emotional features in cyberbullying detection. Figure 6 represents the features engineering methods.

1) WORD EMBEDDINGS

Word embeddings have proven to be effective in capturing semantic relationships and contextual information in textual data [66]. Among the widely used word embedding models, word2vec has gained significant attention due to its ability to capture word similarities and semantic meanings [65]. In the context of our research, word2vec is leveraged to create embedding representations that capture the underlying meaning and context of the text. By representing words as dense vectors in a continuous vector space, word2vec facilitates the measurement of word similarity and enables the model to generalize based on the learned representations.

word2vec's ability to capture semantic meaning, contextual understanding, dimensionality reduction, robustness to vocabulary size, and capture of syntactic and semantic regularities make it a preferred choice for embedding feature extraction in cyberbullying detection. we have extracted 300 features using word2vec. Incorporating word2vec as a feature engineering technique enhances the representation of textual data for our research. While n-grams, TF-IDF, and other techniques have their own merits, word2vec offers distinct advantages that enhance the model's ability to understand the underlying meaning, linguistic patterns, and contextual cues associated with cyberbullying behaviour.

$$\mathbf{w}_t = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbf{v}_i \tag{3}$$

The above equation stated that each text sample T is represented as a sequence of words, denoted by w_t. The feature vector w_t is computed by averaging the word vectors $\mathbf{v}_i$ of all the cyberbullying and non-cyberbullying words in the text sample. The resulting feature vector represents the numerical representation of the text for further analysis and classification. Figure 2 shows the block diagram of Word2Vec features and emotional features.

2) EMOTIONAL FEATURES

A team of researchers [67], [68] from the National Research Council Canada (NRC) employed word selection as a method to detect specific positive and negative emotions from text. They meticulously curated comprehensive dictionaries for each emotional category, encompassing all relevant words and expressions. Leveraging the NRC dictionaries, we extracted eight emotional attributes: anticipation, joy, surprise, trust, anxiety, sadness, anger, and disgust from the used dataset. Among these, anticipation, joy, surprise, and trust are classified as Positive and remaining as Negative emotions. Cyberbullying often involves the expression of negative emotions and aggressive behaviour. Recognizing the importance of emotional context in detecting cyberbullying instances, we have incorporated emotional features in our previous work [66]. These features aim to capture the emotional tone or sentiment expressed in the text, enabling the model to identify instances that contain offensive or abusive content. In this study, we extracted eight emotional features from our dataset for each record. These features include anger, fear, sadness, disgust, joy, surprise, trust, and anticipation. By quantifying the emotional content of the text, these features provide valuable information for distinguishing between cyberbullying and non-cyberbullying instances.

$$F_{x_i} = \frac{(W_{mt} * W_n)}{100} \tag{4}$$

The emotional score, F_{x_i}, represents the intensity of a specific emotional feature. Each emotional feature, such as anticipation, joy, surprise, trust, anxiety, sadness, anger, and disgust, contributes to the overall emotional score. The emotional score is affected by factors such as the number of matching words in the dictionary for a given feature (W_{mt}), as well as the total number of words in a sentence (W_n). For a given preprocessed tweet/ post text, equation 4 calculates the matching word in the text and corresponding dictionary related to an emotion and assigns the emotion feature score considering a total number of words in the text. For each of the given tweets/posts, equation 4 is evaluated eight times to assign emotional scores for the eight emotional features. Therefore, when evaluating the emotional score of a sentence or its paraphrase, all the emotional features, their respective matching words, and the sentence's word count are taken into consideration. These emotional features have not been previously employed in the context of cyberbullying detection. This pioneering approach brings a fresh perspective to the field and lays the foundation for a more nuanced understanding of online behavior.

Feature engineering techniques, such as word2vec for embedding feature extraction and the inclusion of emotional features, contribute to a more comprehensive representation of our data for cyberbullying detection. These techniques enable the model to capture semantic relationships, contextual information, and emotional cues, enhancing its ability to accurately identify instances of cyberbullying. The fusion of these features empowers the model to leverage both linguistic patterns and emotional context in detecting and addressing cyberbullying incidents effectively.

V. FEDERATED LEARNING FRAMEWORK FOR CYBERBULLYING

The Federated Learning framework enables collaborative model training across multiple clients while preserving the privacy of individual client data. In the context of cyberbullying detection, We utilized Federated Learning to build an effective model using a Deep Neural Network (DNN) [69] as the initial global model, which is then distributed to all participating clients.

The central server coordinates the training process, while the clients possess their own local data containing text samples relevant to cyberbullying. Through iterative rounds of client participation, the global model is updated, allowing for continuous improvement in cyberbullying detection accuracy. Figure 7 shows the complete working of the federated learning process for Cyberbullying Detection.

A. ARCHITECTURE AND MECHANISMS OF FEDERATED LEARNING FOR CYBERBULLYING

1) INITIALIZATION

Our Federated Learning process begins with the initialization of the global model. In our work, we choose a Deep Neural Network (DNN) model as the initial global model due to its capability to capture complex patterns in text data and its high performance in cyberbullying detection [70]. The central server is responsible for initializing the DNN model using an initialization function, denoted as InitializeDNNModel(), and distributing it to all participating clients.

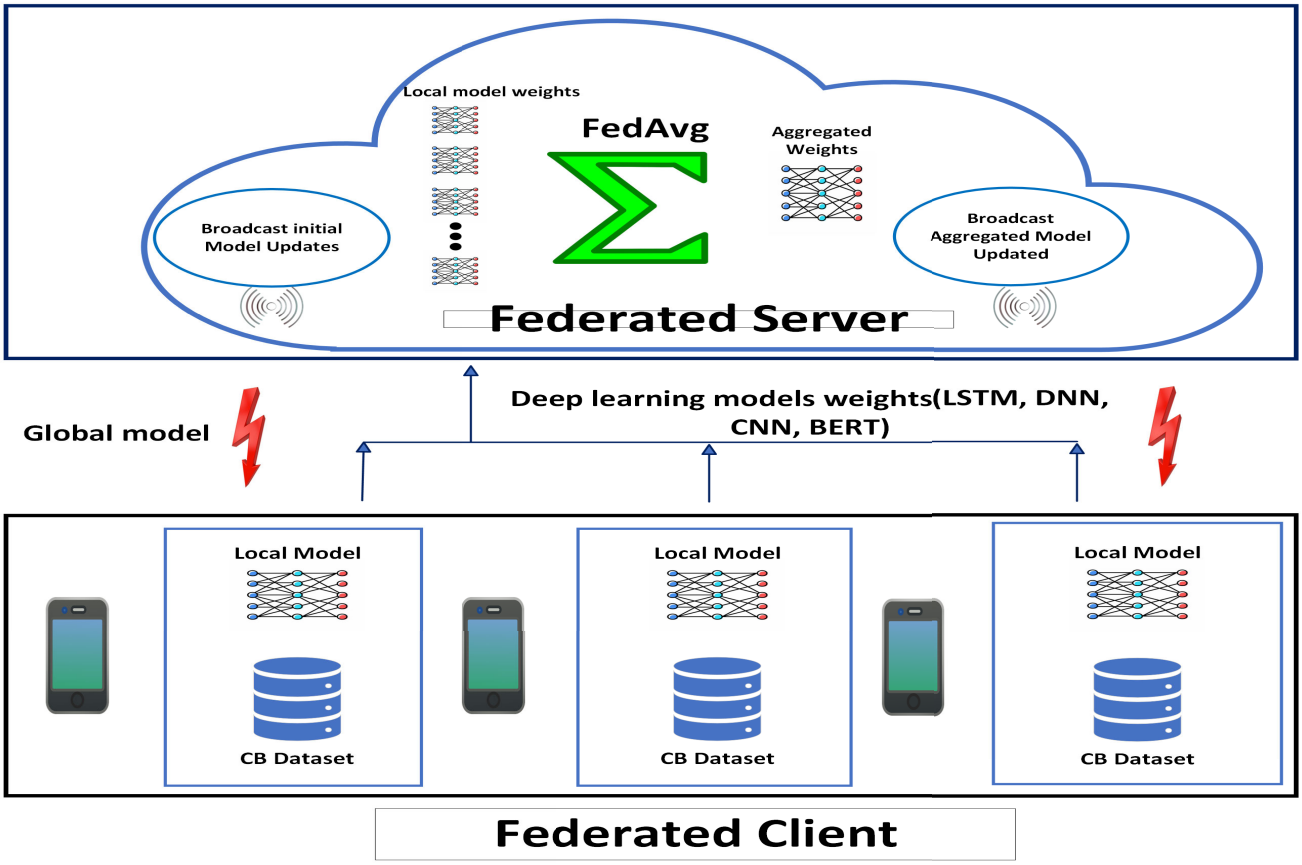

FIGURE 7. Proposed federated learning framework with detailed steps.

Mathematically, the global model initialization can be represented by the equation:

$$\theta^0_{\text{global}} = \text{InitializeDNNModel}() \tag{5}$$

Here, θ^0_{global} represents the initial model parameters of the global DNN model at the start of the federated learning process. The function InitializeDNNModel() sets the initial weights and biases of the DNN model based on the chosen architecture.

By using the DNN model as the initial global model, we can effectively capture intricate patterns in text data, which is essential for accurate cyberbullying detection. Throughout the federated learning process, the global model undergoes refinement through collaborative local model updates performed by individual clients using their respective local data, leveraging the power of federated learning while maintaining privacy and data security.

2) CLIENT SELECTION

In each training round, a subset of clients is selected to participate based on their performance in previous rounds. The set of available clients in round t is denoted as $\mathcal{C}_t$, and the subset of clients selected to participate in round t based on their performance metrics is represented by $\mathcal{S}_t$. Our proposed selection process is performed using a function SelectClients that takes into account the performance metrics, such as accuracy or F1-Score, and returns the subset of clients $\mathcal{S}_t$ chosen for participation.

Mathematically, the selection process can be represented by the equation:

$$\mathcal{S}_t = \text{SelectClients}(\mathcal{C}_t, \text{PerformanceMetrics}) \tag{6}$$

This equation signifies that the subset of clients $\mathcal{S}_t$ for round t is determined by applying the selection function SelectClients to the set of available clients $\mathcal{C}_t$ and considering the performance metrics. By selecting clients with higher performance metrics, such as accuracy or F1-Score, the intention is to involve the most competent clients in the model training process, leading to improved global performance.

3) LOCAL MODEL TRAINING

After client selection, each client trains its own local Deep learning model broadcasted by the server using its local cyberbullying data. The local model training process involves feeding the local data into the deep learning model and optimizing the model's weights through backpropagation and gradient descent. This step is performed securely on the client's device without sharing raw cyberbullying data,

as only the model updates are communicated to the central server.

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla f_k(\theta_k^t) \tag{7}$$

The equation represents the update rule for the local model parameters in federated learning. At each iteration t, the local model parameters for client k are updated using the gradient descent optimization algorithm. The update is performed by subtracting the product of the learning rate η and the gradient $\nabla f_k(\theta_k^t)$ of the local objective function $f_k(\theta_k^t)$ with respect to the current model parameters θ_k^t from the previous model parameters θ_k^t. This update process enables each client to refine its model based on its local data and the locally computed gradient, contributing to the overall learning process in federated learning.

4) MODEL AGGREGATION

Once the clients (representing social platforms) have completed their local model training, the central server performs model aggregation using the standard Federated Averaging (FedAvg) [69] algorithm. FedAvg computes the weighted average of the model updates received from clients selected for participation.

$$\theta_{\text{avg}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \frac{n_k}{n} \theta_k \tag{8}$$

The FedAvg algorithm in federated learning is used to aggregate model parameters across K clients. It computes the average of the client parameters, denoted by θ_k, weighted by the number of samples n_k contributed by each client, and scaled by the total sample size n. The resulting averaged parameter, θ_{avg}, represents the updated global model.

5) MODEL BROADCASTING

After the model aggregation step, the new global DNN model is broadcast back to the participating clients. This ensures that all clients have access to the updated global model for the next round of training. Mathematically, the model broadcasting can be represented by the equation:

$$\theta_k^{t+1} = \theta_{\text{global}}^t \tag{9}$$

In the above equation, θ_k^{t+1} represents the updated model parameters for client k in the next round of training $(t+1)$, and θ_{global}^t denotes the global model parameters obtained from the model aggregation step in the current round of training (t). All federated clients initially receive the same deep learning model, as created by the central server. However, as training progresses over multiple rounds, the deep learning model on each client diverges differently due to the fact that each client trains its model exclusively on its local dataset. These divergences are primarily due to differences in the local data distributions, resulting in distinct weight values for each client's model. After a specific round of training, all these client models are shared back with the central server for aggregation. This aggregation process combines the knowledge from all clients' models, creating a global model with updated weight values. This mechanism characterizes one round of federated learning in our system. Importantly, each federated client possesses its own unique test dataset, which is not shared with other clients. Thus, when evaluating a trained local model, it is assessed solely on its respective local test data. There is no requirement for combining the output of federated clients to classify data instances. The broadcasting operation ensures that every client receives an identical updated global model. This uniform distribution of the updated model enables all clients to utilize the latest model parameters for their subsequent training iterations. This uniformity is crucial for maintaining consistency and accuracy across all clients' training processes.

6) ITERATIVE TRAINING

The Federated Learning process involves multiple rounds of collaboration between the central server and participating clients. In our research, we perform 20 server rounds, where all clients perform local model training for 200 epochs in each round. Mathematically, this can be represented as:

Server Rounds: t = 1, 2, …, 20

Local Model Training: for each client k in round t

for $i = 1, 2, \ldots, 200$ (epochs)

Update local model $\theta_k^{t,i} = \theta_k^{t,i-1} - \eta \nabla f_k(\theta_k^{t,i-1})$

During the local model training, each client utilizes its individual data to update its model parameters. The notation $\theta_k^{t,i}$ represents the model parameters of client k after i epochs in round t, η is the learning rate, and $\nabla f_k(\theta_k^{t,i-1})$ denotes the gradient of the loss function f_k with respect to the model parameters $\theta_k^{t,i-1}$.

After the completion of local training, clients send their model updates to the central server. The server selectively aggregates the models based on performance metrics using the FedAvg algorithm. This can be mathematically represented as:

Model Aggregation:

$$\begin{aligned} \theta_{\text{global}}^t &= \text{FedAvg}(\{\theta_k^{t,5}\}) \\ &= \frac{1}{K} \sum_{k=1}^{K} \left(\theta_k^{t,5} \cdot \frac{n_k}{n} \right) \end{aligned}$$

Here, θ_{global}^t represents the aggregated model parameters at the central server in round t, K is the total number of participating clients, n_k denotes the number of samples available from client k, and n represents the total sample size. The FedAvg algorithm calculates the weighted average of the client models based on their relative contributions to the training process.

Finally, the aggregated model, which represents the collective knowledge of the high-performing clients, is broadcast back to all participants. This iterative process enables the global model to gradually improve its accuracy and performance for cyberbullying detection.

VI. EXPERIMENTAL SETUP

The experiments were conducted using Google Colab, a cloud-based platform equipped with GPU acceleration, to facilitate efficient computation. The Python programming language, specifically the latest version, was utilized for implementing the experimental procedures. TensorFlow, a widely adopted deep learning framework, served as the primary tool for model development and training. The experimental setup for federated learning involved exploring various configurations by varying the number of clients, aiming to assess their impact on the overall performance. In our research paper, we devised a robust Federated Learning configuration for cyberbullying detection, combining the power of BERT with 20 global aggregation rounds and 200 epochs for local model training. This optimized setup exhibited exceptional performance, outperforming all other deep learning models in precision for cyberbullying identification.

During the Federated Learning process, four clients were involved, and each client's local model was trained over 200 epochs with carefully tuned hyperparameters. Table 2 shows the best selected hyperparameters with the structure of the deep learning model used in the experiments. For global model aggregation, we strategically employed 20 rounds, effectively consolidating insights from each client without compromising individual data privacy. The integration of BERT, with its contextual understanding of text, proved to be a game-changer. Through fine-tuning on our cyberbullying dataset, BERT demonstrated superior precision in identifying cyberbullying instances compared to other deep learning models, including CNN, DNN, and LSTM. Our comprehensive analysis of the results revealed the undeniable success of this configuration. The fusion of Federated Learning's privacy-preserving paradigm with the precision-boosting capabilities of BERT led to groundbreaking performance improvements in cyberbullying detection. By harnessing the collective intelligence of distributed clients while respecting data privacy, our research opens new avenues for effective cyberbullying detection and demonstrates the immense potential of Federated Learning with advanced models like BERT.

In evaluating the performance of the federated learning model for cyberbullying detection, a comprehensive set of metrics was employed. The evaluation metrics encompassed accuracy, precision, and recall, alongside the F1-Score. These metrics provided a holistic understanding of the model's effectiveness in correctly identifying instances of cyberbullying across the different deep-learning architectures. A comprehensive discussion of the detailed results obtained from these experimental configurations can be found in the dedicated ''Results'' section, where the performance of the federated learning models in cyberbullying detection will be thoroughly examined and analyzed.

VII. RESULTS

The results of our comprehensive experimentation provide valuable insights into the effectiveness of various deep learning models, including DNN, LSTM, and BERT, for cyberbullying detection. Our experiments were conducted on a dataset comprising English text from multiple platforms, where each tweet was already pre-labeled to indicate the presence of cyberbullying or not. To begin, we utilized the powerful computational resources provided by Google Colab, leveraging GPU support to accelerate the training process. The dataset was meticulously pre-processed, ensuring its suitability for training and evaluation purposes. We explored different model architectures to understand their impact on cyberbullying detection (in addition to varying the model architectures). Specifically, we implemented DNN models with different layer configurations, LSTM models with varying numbers of hidden units, and BERT models with distinct pre-trained configurations, including BERT-Base, BERT-Large, and BERT-Multilingual.

A. LOCAL VS. GLOBAL MODEL PERFORMANCE

In this section, two types of experiments are conducted and compared, namely: 1) Local model learning, and 2) Global model learning using the federated framework. For local model learning, the clients do not interact with each other and there is no server for cooperative learning (A deep learning model is only trained on local data and tested on test data). For global model learning, the central server performs aggregation (in each round of federated learning) of the local models learned by the clients and that is how the federated clients cooperatively learn the underlying pattern in the dataset without sharing their local data. Among the models evaluated, global BERT model consistently demonstrated exceptional performance in cyberbullying detection, achieving an impressive accuracy rate of 92 %. Considering its superior performance, we established the global BERT model as our benchmark model for further comparison. Next, we present the results obtained from other models in comparison to our benchmark BERT model. The DNN models exhibited satisfactory performance, achieving an average accuracy of 86 %. However, their performance was slightly lower than that of BERT. Similarly, the LSTM models yielded a respectable accuracy rate of 88 %, while CNN gave 86 % accuracy, showcasing their potential for cyberbullying detection but falling short of BERT's performance. By leveraging the capabilities of federated learning, we were able to maintain data privacy while still achieving remarkable accuracy rates across all models. The detailed results of all the models are presented in Table 3 to Table 6. The distributed learning approach proved to be effective in addressing the challenges associated with cyberbullying detection, allowing for accurate detection without compromising user privacy.

Figure 8 shows a graph that provides a visual representation of the training and validation accuracy of our benchmark model averaged across all four clients in the last round. It tracks how well the model is performing during the training process. The x-axis represents the number of training epochs or iterations, while the y-axis represents the accuracy

TABLE 2. Parameters of the used deep learning models.

Hyper Parameters	DNN	CNN	LSTM
LSTM units	-	-	2
Hidden neurons	-	-	1024,512
Dense layers	4(1024,256,128,1)	1 (3)	4 (1024,512,256,2)
Max- pooling	-	4	-
Act func on hidden Layer	ReLU	ReLU	ReLU
Act func on output Layer	Softmax	Softmax	Softmax
Epochs	200	200	200
Batch_size	128	128	128
Optimizer	Adam	Adam	Adam

TABLE 3. Local and global performance for CNN.

CNN	Local Precision	Global Precision	Local Recall	Global Recall	Local F1-Score	Global F1-Score	Local Accuracy	Global Accuracy
Client 1	82.21 ± 0.24	85.55 ± 0.14	83.31 ± 0.29	85.16 ± .28	82.41 ± 0.41	85.71 ± 0.83	83.23 ± 0.31	85.41 ± 0.52
Client 2	83.61 ± 0.29	85.96 ± 0.14	84.35 ± 0.24	84.32 ± .29	81.44 ± 0.27	85.82 ± 0.26	82.54 ± 0.34	86.25 ± 0.21
Client 3	79.32 ± 0.14	85.63 ± 0.44	80.22 ± 0.24	84.73 ± .21	82.42 ± 0.26	86.14 ± 0.21	81.29 ± 0.21	86.72 ± 0.28
Client 4	86.33 ± 0.31	87.37 ± 0.12	85.37 ± 0.24	88.75 ± 26	86.34 ± 0.33	86.17 ± 0.31	84.21 ± 0.27	86.18 ± 0.12
Average	82.87	86.28	83.31	85.74	83.15	85.96	82.82	86.14

TABLE 4. Local and global performance for DNN.

DNN	Local Precision	Global Precision	Local Recall	Global Recall	Local F1-Score	Global F1-Score	Local Accuracy	Global Accuracy
Client 1	84.11 ± 0.21	86.17 ± 0.13	86.32 ± 0.19	84.56 ± 0.18	82.31 ± 0.33	85.11 ± 0.43	84.43 ± 0.21	88.12 ± 0.12
Client 2	86.51 ± 0.19	87.56 ± 0.14	84.34 ± 0.21	87.52 ± 0.29	87.14 ± 0.37	87.01 ± 0.28	86.34 ± 0.34	86.01 ± 0.24
Client 3	84.12 ± 0.44	86.25 ± 0.14	85.21 ± 0.26	86.46 ± 0.41	86.32 ± 0.22	88.11 ± 0.51	86.39 ± 0.24	85.12 ± 0.21
Client 4	84.44 ± 0.11	85.14 ± 0.23	86.21 ± 0.32	84.42 ± 0.36	84.01 ± 0.33	85.22 ± 0.41	85.22 ± 0.37	86.10 ± 0.32
Average	84.80	86.28	85.52	85.74	84.98	86.11	85.60	86.34

TABLE 5. Local and global performance for LSTM.

LSTM	Local Precision	Global Precision	Local Recall	Global Recall	Local F1-Score	Global F1-Score	Local Accuracy	Global Accuracy
Client 1	88.15 ± 0.39	88.18 ± 0.31	84.17 ± 0.41	86.28 ± 0.31	87.27 ± 0.61	87.22 ± 0.32	87.27 ± 0.45	87.95 ± 0.51
Client 2	86.27 ± 0.46	87.48 ± 0.39	85.27 ± 0.32	89.84 ± 0.51	85.27 ± 0.34	88.64 ± 0.38	85.19 ± 0.35	88.89 ± 0.27
Client 3	84.27 ± 0.32	87.19 ± 0.41	82.27 ± 0.33	89.97 ± 0.43	83.17 ± 0.41	88.57 ± 0.32	84.17 ± 0.54	84.37 ± 0.23
Client 4	88.27 ± 0.33	85.01 ± 0.51	86.27 ± 0.33	88.58 ± 0.38	84.17 ± 0.32	86.06 ± 0.44	86.26 ± 0.44	88.47 ± 0.33
Average	86.74	86.96	84.50	88.66	84.97	87.76	85.72	88.17

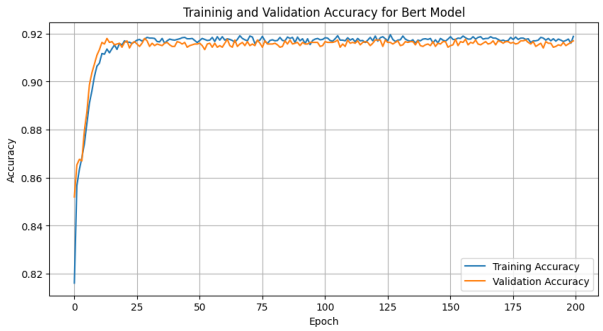

FIGURE 8. Fedbert model accuracy graph for training and validation.

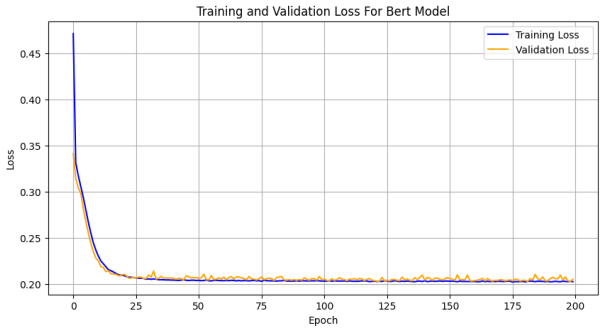

FIGURE 9. FedBERT model loss graph for training and validation.

percentage. Figure 9 illustrates the training and validation loss of our benchmark model averaged across all four clients in the last round. Like the first graph, it has the number of training epochs on the x-axis and the loss on the y-axis. The graph demonstrates that the training loss steadily decreases, indicating that the model is improving its fit to the training data. The validation loss follows a similar pattern, decreasing initially and then stabilizing. The lowest point on this curve corresponds to the model's best performance on the validation dataset.

B. SCALABILITY ANALYSIS OF BEST MODEL

In order to analyze the scalability performance of the benchmark model (FedBERT), the dataset is partitioned evenly among a varying number of clients in the range

TABLE 6. Local and global performance for BERT.

BERT	Local Precision	Global Precision	Local Recall	Global Recall	Local F1-Score	Global F1-Score	Local Accuracy	Global Accuracy
Client 1	88.27 ± 0.33	88.71 ± 0.31	86.27 ± 0.33	88.17 ± 0.32	87.17 ± 0.23	87.17 ± 0.42	86.27 ± 0.43	91.17 ± 0.13
Client 2	90.87 ± 0.08	91.75 ± 0.34	91.47 ± 0.13	92 .20 ± 0.43	89.27 ± 0.33	91.16 ± 0.42	90.17 ± 0.32	92.81 ± 0.41
Client 3	87.31 ± 0.21	89.88 ± 0.44	89.31 ± 0.43	86.11 ± 0.21	86.21 ± 0.23	86.10 ± 0.44	84.13 ± 0.33	91.84 ± 0.13
Client 4	91.17 ± 0.31	90.67 ± 0.13	91.37 ± 0.34	91.10 ± 0.44	90.28 ± 0.18	90.18 ± 0.22	91.27 ± 0.33	92.78 ± 0.34
Average	89.41	90.25	89.61	89.15	88.23	88.65	87.96	92.15

TABLE 7. Comparison of deep learning models using proposed federated learning framework.

Models	F1-Score	Accuracy
FedBERT	**88.65**	**92.15**
FedLSTM	87.76	88.17
FedCNN	85.96	86.14
FedDNN	86.11	86.34

of 4-30 for our best-performing model FedBERT. Table 8 presents a comprehensive scalability analysis of the proposed cyberbullying detection system across varying numbers of clients for the best-performing global model. As the number of clients increases from four to thirty, the system maintains a consistently high level of accuracy, hovering around 90%. This indicates its robustness in handling diverse online communities. In terms of precision, recall, and F1-score, the system also demonstrates remarkable stability, with precision scores ranging from 87 to 90, recall scores from 86 to 89, and F1-scores from 85 to 89. These findings underscore the system's ability to perform effectively at scale, making it a promising solution for addressing cyberbullying across varying numbers of clients and characteristics.

C. STATISTICALLY SIGNIFICANCE TEST RESULTS

Table 9 presents the comparison of significance differences between the two competing algorithms using nonparametric two-tailed Wilcoxon signed rank test [71] based on F1-Score and accuracy values (given in Table 7) at the 0.05 significance level. The null hypothesis is that the mean F1-Scores or accuracies of two competing algorithms are equal. The null hypothesis is rejected when there is a statistically significant difference between the performance of two competing algorithms, while the alternative hypothesis suggests otherwise. FedBERT emerged as the standout performer in our analysis, consistently outperforming all other competing algorithms. The p-values obtained for the comparisons between FedBERT and other algorithms were consistently highly significant ($p < 0.05$), leading us to reject the null hypothesis. This rejection indicates a clear and statistically significant superiority of FedBERT in terms of both F1-Scores and accuracy. In addition to the FedBERT comparisons, we observed that all other pairs of competing algorithms also exhibited significant differences in their performance. The null hypothesis was consistently rejected for these pairs as well. This underscores the importance of careful algorithm selection, as the choice of algorithm can significantly impact the quality of results in tasks related to F1-Scores and accuracy. Notably, when comparing FedCNN and FedDNN, we found that the null hypothesis was accepted for both F1-Scores and accuracy, indicating that there was no statistically significant difference in the performance of these two algorithms. Our comprehensive statistical analysis using the Wilcoxon signed rank test has provided strong evidence to support that FedBERT consistently demonstrated its statistical superiority over other competing algorithms, highlighting its potential as a top-performing choice for cyberbullying detection in federated learning environment related to F1-Scores and accuracy.

D. SECURITY ANALYSIS

In this section, we perform the security analysis of the proposed Federated Learning-based cyberbullying detection method. Although sharing model weights and gradients instead of clients' local data provides security to some extent, we need to quantify the level of privacy preservation for a more thorough security analysis. Differential Privacy (DP) [72], a well-known method, can be used to quantify the bounds on privacy leakage in the proposed method. First, we discuss the intuition behind DP using its standard definition and then provide details of how it is integrated into the proposed method. Let's assume we have two datasets, D and D', with only one record difference. M represents the mechanism that acts or queries over these datasets. We can guarantee the privacy of M using a privacy budget ϵ, denoted as (ϵ-differentially private), if the probability P of every outcome S never differs by more than e^{ϵ} between D and D', (i.e., with and without one record). Suppose $P(M(D) \in S)$ is the probability of $M(D)$ belonging to set S. Then, we aim for $\log\left(\frac{P(M(D)\in S)}{P(M(D')\in S)}\right)$ to be a small value (ideally zero) to achieve better privacy preservation.

The following equation represents the DP method:

$$\log\left(\frac{P(M(D) \in S)}{P(M(D') \in S)}\right) \leq \epsilon \tag{10}$$

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S) \tag{11}$$

For (ϵ, δ)-differentially private guarantee with δ representing the failure probability:

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S) + \delta \tag{12}$$

TABLE 8. FedBERT with different number of clients for scalability analysis.

Number of Clients	Precision	Recall	F1-Score	Accuracy
4	90	89	89	92
10	89	89	89	91
20	89	87	87	91
30	87	86	85	90

TABLE 9. Statistical significance comparison of algorithms using two tailed Wilcoxon signed rank test.

Algorithm 1	Algorithm 2	F1-Scores		Accuracy	
		P-value	Null Hypothesis	P-value	Null Hypothesis
FedBERT	FedLSTM	**0.00015**	Rejected	**0.00024**	Rejected
FedBERT	FedCNN	**0.00976**	Rejected	**0.00488**	Rejected
FedBERT	FedDNN	**0.00390**	Rejected	**0.00097**	Rejected
FedLSTM	FedCNN	**0.00195**	Rejected	**0.00244**	Rejected
FedLSTM	FedDNN	**0.02343**	Rejected	**0.01171**	Rejected
FedCNN	FedDNN	0.93750	Accepted	0.97656	Accepted

TABLE 10. Differential Privacy based security analysis for FedBERT setup.

Gradient Clipping	Noise Multiplier	Privacy Budget (ϵ)	Averaged Test Accuracy
1	4	3.10	91.75
1.5	4	3.33	88.30
2	4	3.67	89.60
1	5	2.34	92.25
1	2	9.20	88.90

The limits on the privacy loss can be imposed by adding noise in the original data. For this purpose, we have used DP-SGD [73] (differentially private stochastic gradient descent) to train the proposed models on the clients before sharing them. In particular the norms of the gradients are clipped (to bound the gradients) and added with Gaussian noise to make the weights sharing privacy preserved. The privacy gaurantee is computed using DP [74] accountant with $\delta = 0.0001$. For various values of gradient clipping norm and Gaussian noise standard deviation, experiments are conducted for the setup of four clients using FedBERT method and the results are summarized in Table 10. In the first three rows of Table, we can observe variations in gradient clipping and noise multiplier values while keeping the privacy budget relatively constant. As gradient clipping increases from 1 to 2, the averaged test accuracy remains reasonably high, with values above 88%. This suggests that moderate gradient clipping does not significantly impact accuracy. The choice of noise multiplier (4) seems to be effective in these cases. A noise multiplier of 5 and a lower privacy budget (2.34) yields the highest averaged test accuracy of 92.25%. This demonstrates that, with the right combination of gradient clipping, noise multiplier, and privacy budget, we can achieve very high accuracy while maintaining reasonable privacy guarantees. It is evident that we can achieve both strong privacy protection and high accuracy with the right combination of hyperparameters, demonstrating the effectiveness of proposed method.

In conclusion, our extensive experimentation demonstrated that BERT outperforms other deep learning models, such as DNN and LSTM, in cyberbullying detection. The results emphasize the significance of leveraging pre-trained language models like BERT, which exhibit strong contextual understanding and contribute to accurate detection. Furthermore, our exploration of various federated learning setups highlighted their potential to preserve data privacy while achieving robust performance. These findings contribute to the advancement of cyberbullying detection techniques and provide valuable insights for future research in this domain. The performance of different deep learning models for cyberbullying detection is summarized in Table 7. Among these models, the BERT (Bidirectional Encoder Representations from Transformers) model consistently demonstrates strong performance across clients. It achieves high precision, recall, and accuracy values at both the local and global levels, indicating its effectiveness in accurately identifying and classifying instances of cyberbullying. The results highlight BERT's superior performance and its potential

for enhancing online safety by combating cyberbullying incidents.

VIII. CONCLUSION

The research conducted in this study made significant contributions to the field of cyberbullying detection using federated learning. By introducing eight novel emotional features extracted from textual tweets, the study enhanced the identification of cyberbullying instances by providing a deeper understanding of the emotional context within messages. The incorporation of privacy-preserving federated learning empowered collaborative cyberbullying detection, ensuring data privacy while promoting cooperation among diverse entities for a more scalable and effective approach. Additionally, the study utilized a performance-based best client selection method for global model aggregation, leading to a more robust and accurate global model. Extensive experimentation demonstrated that the powerful BERT model outperforms other models like CNN, DNN, and LSTM in identifying cyberbullying instances, especially when configured with 200 local model epochs and 20 global aggregation rounds. Extensive set of experiments are performed to highlight the scalibility performance of the proposed method. Differential Privacy based security analysis is provided to quantify the level of privacy preservation offered by the proposed method.

In conclusion, this research has made significant strides in the field of cyberbullying detection using federated learning. It not only advances the current understanding of cyberbullying detection methodologies but also lays a strong foundation for future research in this area. By combining innovative features, privacy-preserving techniques, and powerful models, the study contributes to the development of more accurate and efficient cyberbullying detection systems. Ultimately, the research aims to create a safer and more respectful online environment, positively impacting the ongoing efforts to combat cyberbullying and foster a healthier digital society.

ACKNOWLEDGMENT

The authors would like to express their grateful to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Dr. Salabat is working for National Research Foundation of Korea (NRF) under the Brain Pool Program (Grant No. 2022H1D3A2A02055024) and Creative Research project (ID: RS-2023-00248526). Any correspondence related to this article should be addressed to Salabat Khan.

(Nagwan Abdel Samee and Umair Khan contributed equally to this work.)

REFERENCES

[1] J. Chun, J. Lee, J. Kim, and S. Lee, "An international systematic review of cyberbullying measurements," *Comput. Hum. Behav.*, vol. 113, Dec. 2020, Art. no. 106485.

[2] G. W. Giumetti and R. M. Kowalski, "Cyberbullying via social media and well-being," *Current Opinion Psychol.*, vol. 45, Jun. 2022, Art. no. 101314.

[3] W. Cassidy, C. Faucher, and M. Jackson, "Adversity in university: Cyberbullying and its impacts on students, faculty and administrators," *Int. J. Environ. Res. Public Health*, vol. 14, no. 8, p. 888, Aug. 2017.

[4] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: Review of an old problem gone viral," *J. Adolescent Health*, vol. 57, no. 1, pp. 10–18, Jul. 2015.

[5] G. M. Abaido, "Cyberbullying on social media platforms among university students in the united Arab Emirates," *Int. J. Adolescence Youth*, vol. 25, no. 1, pp. 407–420, Dec. 2020.

[6] E. R. Kutok, S. Dunsiger, J. V. Patena, N. R. Nugent, A. Riese, R. K. Rosen, and M. L. Ranney, "A cyberbullying media-based prevention intervention for adolescents on Instagram: Pilot randomized controlled trial," *JMIR Mental Health*, vol. 8, no. 9, Sep. 2021, Art. no. e26029.

[7] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, "Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102145.

[8] Q. W. Khan, A. N. Khan, A. Rizwan, R. Ahmad, S. Khan, and D.-H. Kim, "Decentralized machine learning training: A survey on synchronization, consolidation, and topologies," *IEEE Access*, vol. 11, pp. 68031–68050, 2023.

[9] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, "The history of social media and its impact on business," *J. Appl. Manage. entrepreneurship*, vol. 16, no. 3, p. 79, 2011.

[10] S. Bauman, *Cyberbullying: What Counselors Need to Know*. Hoboken, NJ, USA: Wiley, 2014.

[11] S. Khan, A. Rizwan, A. N. Khan, M. Ali, R. Ahmed, and D. H. Kim, "A multi-perspective revisit to the optimization methods of neural architecture search and hyper-parameter optimization for non-federated and federated learning environments," *Comput. Electr. Eng.*, vol. 110, Sep. 2023, Art. no. 108867.

[12] K. Miller, "Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress," *Southern California Interdiscipl. Law J.*, vol. 26, p. 379, Jul. 2016.

[13] M. Price and J. Dalgleish, "Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people," *Youth Stud. Aust.*, vol. 29, no. 2, pp. 51–59, 2010.

[14] P. K. Smith, "Cyberbullying and cyber aggression," *Handbook of School Violence and School Safety: International Research and Practice*, vol. 2. 2012, pp. 93–103.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[16] Y. Goldberg and O. Levy, "Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.

[17] J. Li, G. Huang, C. Fan, Z. Sun, and H. Zhu, "Key word extraction for short text via word2vec, doc2vec, and textrank," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 27, no. 3, pp. 1794–1805, May 2019.

[18] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[19] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[20] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70701–70718, 2019.

[21] M. Maalouf, "Logistic regression in data analysis: An overview," *Int. J. Data Anal. Techn. Strategies*, vol. 3, no. 3, pp. 281–299, Jul. 2011.

[22] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.

[23] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 2354–2358.

[24] C. Mc Guckin and L. Corcoran, *Cyberbullying: Where are we now? A Cross-National Understanding*. Wuhan, China: MDPI, 2017.

[25] T. Vaillancourt, R. Faris, and F. Mishna, "Cyberbullying in children and youth: Implications for health and clinical practice," *Can. J. Psychiatry*, vol. 62, no. 6, pp. 368–373, Jun. 2017.
[26] A. Görzig and K. Ólafsson, "What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries," *J. Children Media*, vol. 7, no. 1, pp. 9–27, Feb. 2013.
[27] J. Eronen, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Exploring the potential of feature density in estimating machine learning classifier performance with application to cyberbullying detection," 2022, *arXiv:2206.01949*.
[28] J. Bhagya and P. Deepthi, "Cyberbullying detection on social media using SVM," in *Inventive Systems and Control: Proceedings of ICISC 2021*. Cham, Switzerland: Springer, 2021, pp. 17–27.
[29] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Proc. Comput. Sci.*, vol. 181, pp. 605–611, Jan. 2021.
[30] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, Jan. 2016, pp. 1–6.
[31] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Commun. Inf. Sci. Manage. Eng.*, vol. 3, no. 5, p. 238, 2013.
[32] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, Sep. 2012.
[33] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, 2011, pp. 11–17.
[34] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," in *Proc. Content Anal. Web*, vol. 2, Apr. 2009, pp. 1–7.
[35] J. Qiu, M. Moh, and T.-S. Moh, "Multi-modal detection of cyberbullying on Twitter," in *Proc. ACM Southeast Conf.*, Apr. 2022, pp. 9–16.
[36] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Syst.*, vol. 28, pp. 2043–2052, Feb. 2021.
[37] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Towards cyberbullying-free social media in smart cities: A unified multi-modal approach," *Soft Comput.*, vol. 24, no. 15, pp. 11059–11070, Aug. 2020.
[38] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 328–339, Jul. 2017.
[39] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 740–745.
[40] K. Shriniket, P. Vidyarthi, S. Udyavara, R. Manohar, and N. Shruthi, "A time optimised model for cyberbullying detection," *Int. Res. J. Modernization Eng., Technol. Sci.*, vol. 4, no. 7, pp. 808–815, 2022.
[41] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2018, pp. 141–153.
[42] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; a reproducibility study," 2018, *arXiv:1812.08046*.
[43] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification," *Multimedia Syst.*, vol. 28, no. 6, pp. 1897–1904, 2022.
[44] P. Yi and A. Zubiaga, "Cyberbullying detection across social media platforms via platform-aware adversarial encoding," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16, pp. 1430–1434, 2022.
[45] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the timeline meets the pipeline: A survey on automated cyberbullying detection," *IEEE Access*, vol. 9, pp. 103541–103563, 2021.
[46] K. Verma, T. Milosevic, K. Cortis, and B. Davis, "Benchmarking language models for cyberbullying identification and classification from social-media texts," in *Proc. 1st Workshop Lang. Technol. Resour. Fair, Inclusive, Safe Soc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 26–31.
[47] B. Bhatia, A. Verma, Anjum, and R. Katarya, "Analysing cyberbullying using natural language processing by understanding jargon in social media," in *Sustainable Advanced Computing*. Singapore: Springer, 2022, pp. 397–406.
[48] K. Chehbouni, G. Caporossi, R. Rabbany, M. De Cock, and G. Farnadi, "Early detection of sexual predators with federated learning," in *Proc. Workshop Federated Learn., Recent Adv. New Challenges (Conjunct NeurIPS)*, 2022, pp. 1–14.
[49] N. P. Shetty, B. Muniyal, A. Priyanshu, and V. R. Das, "FedBully: A cross-device federated approach for privacy enabled cyber bullying detection using sentence encoders," *J. Cyber Secur. Mobility*, vol. 12, no. 4, pp. 465–496, 2023.
[50] A. S. Ram, A. Cn, and K. Nandagopan, "End-to-end messaging system enhancement using federated learning for cyberbullying detection," Tech. Rep., 2022.
[51] D. Apple, "Learning with privacy at scale," *Apple Mach. Learn. J.*, vol. 1, no. 8, 2017.
[52] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
[53] A. Jana and C. Biemann, "An investigation towards differentially private sequence tagging in a federated framework," in *Proc. 3rd Workshop Privacy Natural Lang. Process.*, 2021, pp. 30–35.
[54] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2019.
[55] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Federated learning of deep networks using model averaging," 2016, *arXiv:1602.05629*.
[56] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[57] S. Shahane, "Cyberbullying dataset," in *Cyberbullying*. San Francisco, CA, USA: Kaggle, 2000.
[58] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic monitoring and prevention of cyberbullying," *Int. J. Comput. Appl.*, vol. 144, no. 8, pp. 17–19, Jun. 2016.
[59] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," in *Advances in Information Retrieval*. Moscow, Russia: Springer, 2013, pp. 693–696.
[60] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Proc. Comput. Sci.*, vol. 176, pp. 612–621, 2020.
[61] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proc. 3rd Int. Workshop Socially-Aware Multimedia*, Nov. 2014, pp. 3–6.
[62] M. Dadvar, F. M. de Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 12th Dutch-Belgian Inf. Retr. Workshop (DIR)*. Ghent, Belgium: Universiteit Gent, 2012, pp. 23–25.
[63] B. S. Nandhini and J. I. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm," in *Proc. Int. Conf. Adv. Res. Comput. Sci. Eng. Technol. (ICARCSET)*, Mar. 2015, pp. 1–5.
[64] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[65] S. S. Birunda and R. K. Devi, "A review on word embedding techniques for text classification," in *Proc. Innov. Data Commun. Technol. Appl. (ICIDCA)*, 2021, pp. 267–281.
[66] U. Khan, S. Khan, A. Rizwan, G. Atteia, M. M. Jamjoom, and N. A. Samee, "Aggression detection in social media from textual data using deep learning models," *Appl. Sci.*, vol. 12, no. 10, p. 5083, 2022.
[67] M. S. I. Malik and A. Hussain, "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Comput. Hum. Behav.*, vol. 73, pp. 290–302, Aug. 2017.
[68] R. Plutchik. (1994). *The Psychology and Biology of Emotion*. [Online]. Available: https://www.worldcat.org
[69] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020, *arXiv:2003.01200*.
[70] A. Rizwan, R. Ahmad, A. N. Khan, R. Xu, and D. H. Kim, "Intelligent digital twin for federated learning in AIoT networks," *Internet Things*, vol. 22, Jul. 2023, Art. no. 100698.
[71] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics: Methodology and Distribution*. Cham, Switzerland: Springer, 1992, pp. 196–202.
[72] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*. New York, NY, USA: Springer, Mar. 2006, pp. 265–284.
[73] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, pp. 308–318, 2016.
[74] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Jul. 2017, pp. 263–275.

NAGWAN ABDEL SAMEE received the B.S. degree in computer engineering from Ain Shams University, Egypt, in 2000, and the M.S. degree in computer engineering and the Ph.D. degree in systems and biomedical engineering from Cairo University, Egypt, in 2008 and 2012, respectively. Since 2013, she has been an Assistant Professor with the Information Technology Department, CCIS, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include data science, machine learning, bioinformatics, and parallel computing. Her awards and honors include the Takafull Prize (Innovation Project Track), the Princess Nourah Award in Innovation, the Mastery Award in Predictive Analytics (IBM), the Mastery Award in Big Data (IBM), and the Mastery Award in Cloud Computing (IBM).

UMAIR KHAN received the M.S. degree in computer science from Comsats University Islamabad. He is currently a Lecturer with the Department of Computer Science, Air University, Islamabad, Aerospace and Aviation Campus, Kamra, Pakistan. His research interests include data mining, deep learning models, federated learning, and decentralized edge AI.

SALABAT KHAN received the Ph.D. degree from NUCES FAST, Islamabad, in 2015. He is currently a Brain-Pool Overseas Researcher with the Big Data Research Center, Jeju National University, South Korea. He is also an Associate Professor with COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include evolutionary computation, medical image processing, machine learning, federated learning, decentralized cooperative learning, big data analytics, recommender systems, and the IoT. He was a reviewer of several prestigious international journals and conferences.

MONA M. JAMJOOM received the Ph.D. degree in computer science from King Saud University. She is currently an Associate Professor with the Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include artificial intelligence, machine learning, deep learning, medical imaging, and data science. She has published several research articles in the above areas.

MUHAMMAD SHARIF received the Ph.D. degree from NUCES FAST, Islamabad, in 2018. He is currently an Assistant Professor with COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include medical image processing, deep learning, and big data analytics.

DO HYUEN KIM received the B.S. degree in electronics engineering and the M.S. and Ph.D. degrees in information telecommunication from Kyungpook National University, South Korea, in 1988, 1990, and 2000, respectively. From 1990 to 1995, he was with the Agency for Defense Development (ADD). Since 2004, he has been with Jeju National University, South Korea, where he is currently a Professor with the Department of Computer Engineering. From 2008 to 2009, he was a Visiting Researcher with the Queensland University of Technology, Australia. His research interests include sensor networks, M2M/IoT, energy optimization and prediction, intelligent service, and mobile computing.