Multimedia Tools and Applications (2023) 82:46611–46650
https://doi.org/10.1007/s11042-023-15372-3

FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning

Belal Abdullah Hezam Murshed[1,2] · Suresha[2] · Jemal Abawajy[3] · Mufeed Ahmed Naji Saif[4] · Hudhaifa Mohammed Abdulwahab[5] · Fahd A. Ghanem[6]

Received: 19 April 2022 / Revised: 18 February 2023 / Accepted: 15 April 2023 /
Published online: 2 May 2023

Abstract
The widespread use of Social Media Platforms (SMP) such as Twitter, Instagram, Facebook, etc. by individuals has recently led to a remarkable increase in Cyberbullying (CB). It is a challenging task to prevent CB in such platforms since bullies use sarcasm or passive-aggressiveness strategies. This article proposes a new CB detection model named FAEO-ECNN for detecting and classifying cyberbullying on social media platforms. The proposed approach integrates Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling and Extended Convolutional Neural Network (ECNN) to enhance the accuracy of CB detection process. Initially, pre-processing is performed in order to cleanse the dataset. Next, the features are extracted using multiple models. The unsupervised Fuzzy Adaptive Equilibrium Optimization (FAEO) is utilized for discovering the latent topics from the pre-processed input data, which automatically examines the text data and creates clusters of words. Finally, the cyberbullying classification makes use of the ECNN and Rain Optimization (RO) algorithm to detect CB from posts/texts. We evaluated the proposed FAEO-ECNN thoroughly with two short text datasets: Real-world CB Twitter (RW-CB-Twitter) and Cyberbullying Menedely (CB-MNDLY) datasets in comparison to State of The Art (SoTA) models like Long Short Term Memory (LSTM), Bi-directional LSTM (BLSTM), RNN, and CNN-LSTM. The proposed FAEO-ECNN model outperformed the SoTA models in detecting Cyberbullying on SMP. It has obtained 92.91% of accuracy, 92.28% of recall, 92.53% of precision, and 92.40% of F-Measure over CB-MNDLY dataset. Moreover, it has achieved 91.89% of accuracy, 91.32% of recall, 91.81% of precision, and 91.56% of F-Measure on RW-CB-Twitter dataset.

Keywords Social Media · Cyberbullying Detection · Fuzzy Adaptive Equilibrium Optimization · Short Text Topic Modelling · Deep Learning · CNN

✉ Belal Abdullah Hezam Murshed
belal.a.hezam@gmail.com

Extended author information available on the last page of the article

1 Introduction

Social media networks such as Instagram, Twitter, Facebook, and Flickr have attracted more users for online communication and socialization. Although these platforms help people to communicate with each other and share their ideas about any topic or event, such interaction and online communication on SMPs may ultimately end up with illegal and malicious activities such as CB. Typically, CB is a state of abusive psychological behaviour that has a serious impact on society. CB events have recently increased, primarily among youths, as they dedicate most of their time to browsing and involving in various interactions and discussions on different social media platforms. In particular, Twitter and Facebook, the popularity of these platforms and also their user's anonymity make them vulnerable to CB. As reported in [46], 14% of harassment events occur on Twitter and Facebook, and 37% of such events are involved by youths. Furthermore, CB may cause serious mental problems and lead to suicide, and such suicides may happen as a result of the depression and stress caused by CB events [69]. This highlights the importance of developing a method for detecting CB in social media short text messaging such as comments, posts, and tweets.

CB is increasingly becoming a common problem on Twitter, Facebook, and other SMPs, detecting CB events from tweets and posts on such platforms and also providing protective measures is desirable to overcome CB threats [22]. Bullying on SMPs is perceived to have a negative impact on users. It can be considered as a destructive and threatening act that can be the major cause of life-long problems to victims due to the easy access to SMPs by users through mobile devices, tablets, etc. Typically, monitoring and controlling CB manually on these platforms is a challenging task [2]. Additionally, mining the content of social media tweets, posts, and comments for detecting CB is a challenging task. It can also be more complex and challenging if the bully utilizes sarcasm or passive-aggressiveness strategies. Thus, intelligent and efficient CB detection models are needed to detect CB in social media short text posts [66]. This challenge is the motivation behind this research work, where it mainly focuses on addressing the aforesaid problem by developing an effective approach to detect CB in SMPs.

Regardless the challenges and difficulties posed by social media messages, the existing CB detection models on the Twitter platform have mainly relied on the classification of tweets and some models utilized topic modelling. The current tweet classification models are generally utilized to classify tweets into bullying and non-bullying. These models either adapted supervised machine learning (ML) such as [19, 74, 67, 48, 60] or Deep Learning (DL) as in [2, 73, 9, 27, 34, 51]. The supervised classifiers were found to have low classification performance since the labels of class are immutable and irrelevant to new events [28]. Furthermore, while it can be appropriate for a predetermined set of events, it is incapable of handling tweet's dynamic changes. On the other hand, Topic modelling has been used to capture the important topics or themes from a collection of data forming classes or patterns throughout the dataset. Despite their concept similarity, in the case of short text, the general unsupervised topic models are inefficient. Therefore, specialized unsupervised topic models were utilized for short text [76]. Typically, these models are effective in identifying and extracting trending topics from tweets. They can be helpful in extracting meaningful topics by using bidirectional processing. However, in order to obtain sufficient prior knowledge, these unsupervised models necessitate a substantial amount of training, which is not always sufficient [15]. Given these limitations, an effective tweet

classification model is desirable to fill the gap between classification and topic modelling, resulting in significantly improved adaptability.

Most CB research has concentrated on the social and psychological aspects of CB rather than tackling the problem utilizing technological solutions. Moreover, exploring the possible utilization of textual ML methods and providing solutions not only alleviate CB but also helps victims in the evidence documenting, which is desired for enforcement of the law. Moreover, existing research works require further enhancement to include a consistent and reliable method that accounts for both direct and indirect influences. However, these concerns motivated us to propose a model for latent CB detection in social media using clustering-based topic modelling and Extended Convolutional Neural Network (FAEO-ECNN), which aims at detecting CB in social media posts and classifying them into different classes such as insult, sexism, racism, aggression and non-bullying.

Nonetheless, progress in the identification of cyberbullying has been made through ML and DL techniques. This article proposes a new CB detection approach called FAEO-ECNN to detect CB from social media posts automatically. The proposed approach integrates the unsupervised Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling and Extended CNN with Rain Optimization (RO) algorithm to classify and detect cyberbullying from posts/texts in order to enhance the accuracy of detecting CB. The FAEO-ECNN outperformed SoTA approaches in detecting CB on the social media platforms based on different evaluation metrics. The main contributions of this study are highlighted as follow:

- Propose an approach named FAEO-ECNN for effective cyberbullying detection in social media.
- Present an unsupervised Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling to discover the latent CB topics from social media such data and generate clusters of words.
- Propose ECNN architecture by combining the Wavelet Pooling with CNN architecture to classify and detect the type of CB, reduce the dimensionality issue, and minimize the loss in CNN using meta-heuristic Rain Optimization (RO) algorithm.
- Introduce a way to cleanse and remove the noise in social media data to improve data quality issues related to social media posts. A wide variety of anomalies like acronyms, slang, Elongated, typos, spelling mistakes, concatenated words, and various abbreviations for the same words, etc., are corrected.
- A new dataset is gathered from Twitter using some CB key terms for evaluating the performance of the suggested FAEO-ECNN and other SoTA methods
- The proposed model is evaluated using two short text datasets: Cyberbullying Menedely (CB-MNDLY) and the Real-world Cyberbullying Twitter (RW-CB-Twitter) datasets, to prove the efficiency of the model in detecting and classifying short texts containing CB. The overall evaluation illustrated that the suggested FAEO-ECNN approach outperforms other SoTA approaches in precision, F1-measure, accuracy, and recall.

The article is structured as follows: Sec. 2 presents the recent works related to CB detection. The proposed methodology is presented in Sec. 3. Experimental analysis and performance evaluation are presented in Sec. 4. The discussion is given in Sec. 5. Finally, Sec. 6 concludes the study and provides the future scope.

2 Related works

This section mainly reviews the existing models of cyberbullyin, classification and detection in SMP such as Twitter, Instagram, and Facebook. Essentially, ML was utilized with various Feature Selection (FS) methods for CB detection and classification, such as Alduailaj & Belghith [8] proposed a model to detect CB automatically of the Arabic text data. The suggested method determines the CB utilizing SVM classifier with the Bag of word (BoW) and TF-IDF model to extract features over Real world YouTube and Twitter datasets in order to test and train the classifier. Furthermore, the Farasa tool was included to cope text demerits and enhance bullying attack detection. Dalvi et al. [23] utilized Random Forests (RF) and SVM models with the TF-IDF model to extract features in order to detect CB in tweets. Although SVM attains high performance, complexity is increased when the class labels get increased. Purnamasari et al.[60] introduced a method for detecting CB from the Twitter dataset utilizing information gain (IG) as a FS technique and SVM as a classifier. The authors investigated the various IG selection threshold and various SVM parameters. Then, the findings of cyberbullying classification with IG feature selection achieve higher accuracy than utilizing overall features. Al-garadi et al. [41] utilized various ML classifiers methods like RF, SVM, and Naïve Bayes (NB) for detecting CB using variant extracted features from social media Twitter data like (network, tweet content, user, and activity). Balakrishnan et al. [11] developed a CB detection model to recognize and detect online bullying in the Twitter dataset using user personality feature extractions identified by Big Five and Dark Triad models. The RF classifier was utilized for CB detection. The major issue with the dataset was restricted to GamerGate, and most of the tweets were categorized into binary forms (normal or spam). Balakrishnan et al. [10] employed various ML algorithms such as J48, RF, and NB through the psychological behaviours of Twitter users to identify CB events from tweets and classify them into various classes such as a normal, bully, spammer, and aggressor. They ultimately found that the emotional feature has no effects on the rate of detection. Despite its competence, it is restricted to datasets with fewer class labels and a small size. An ensemble model was suggested by Alam et al. [7] for tweet classification employing the double and single ensemble based voting methods, such ensemble based methods utilize the mutual information bigrams and unigram TF-IDF for extracting features, while LR, Bagging ensemble classifiers, and decision tree were used for classification. The Bagging ensemble approach obtained the best precision over the Twitter dataset. Even though, these models minimized the classification execution and training time, they get affected when using sarcasm tweets and multi-meaning acronym words.

Huang et al. [33] developed a CB detection approach for social media data that combines textual and social media features. The information Gain (IG) approach was used to rank the features. Some standard classifiers models like J48, Bagging, and Naïve Bayes (NB) are used. The results illustrated that the social characteristics could help in improving the accuracy of CB detection. Also, Squicciarini et al. [65] introduced a CB detection approach for social media networking such as MySpace utilizing C4.5 as a decision tree classifier with personal, textual, and social media features. Talpur and O'Sullivan [67] proposed a feature-engineering-based technique for identifying cyberbullying posts on Twitter. The proposed model utilized the features contained in the content of the tweets to propose an ML classifier for categorizing the tweets into four classes: high, medium, low-level CB, and non-CB. This study focused only on the Twitter platform and did not investigate the other social media platforms. An automatic model was suggested by Van Hee et al.

[31] for CB detection on social media, including various kinds of CB, covering comments from bystanders, victims, and bullies. The suggested system was evaluated based on cyberbullying dataset of Dutch and English languages. In this study, the detection of fine-grain classification in CB classes such as curses, racism, threats, and hate has not yet been investigated. Besides, Chia et al. [19] used various ML and feature extraction-based methods for classifying sarcasm and irony from tweets containing CB. Several feature selection methods and classifiers were investigated in this method; Despite its efficiency in identifying the irony and sarcasm words from tweets, the rate of CB detection is still lower [57].

Additionally, some Deep Learning (DL)-based models were proposed in the literature for detecting CB contained in posts/tweets from social media. Lu [45] suggested a new model, namely Character-level CNN with Shortcuts(Char-CNNS), for detecting CB from social media text. The characters were regarded as the smallest learning unit, which was used to enable the proposed model to cope with the wide variety of anomalies in a real-world corpus, such as spelling errors, etc. To address the issue of class imbalance, a focal loss function is used, while shortcut was utilized to stitch various features levels in order for learning hybrid bullying signals. N. Yuvaraj et al. [74] utilized Deep Reinforcement Learning (DRL) and Artificial Neural Network (ANN) for classifying CB tweets. This model led to high computational complexity. A new model was Introduced by Chen et al. [16] for detecting verbal aggression from Twitter data based on Convolutional Neural Networks (CNN), utilizing 2-D TF-IDF matrix of features instead of Glove and Word2vec to enhance sentiment analysis performance. The CNN model was found to be superior to other existing SVM and LR classifiers. Agrawal and Awekar [3] utilized Deep Neural Network(DNN) methods such as BLSTM with attention, BLSTM, LSTM, and CNN methods to analyze and detect CB across different kinds of SMPs on various topics and apply transfer learning to detect CB. The DL-based models outperformed conventional ML models in detecting cyberbullying tasks in social media. Dadvar and Eckert [21] reproduced the study of [3] for detecting CB events in SMP based on DNN models. The author's utilized a new SM dataset (Youtube dataset) in their works to test the transferability and adaptability of their approaches to Youtube dataset and analyze the performance of Deep NN with traditional machine learning models. For tweet classification, Natarajan Yuvaraj et al. [73], utilized multi-feature-based AI along with a deep decision tree classifier. Where classifier was established by integrating the deep neural network's hidden layers with the decision tree classifier. In addition, three feature selection models were utilized: IG, Pearson Correlation, and Chi-Square. However, this approach obtains low accuracy when handling high-dimensional data. For detecting CB hate speech in Arabic tweets, Al-Hassan and Al-Dossari [9] used SVM classifier in comparison to four deep learning models, namely GRU, CNN + GRU, LTSM, and CNN + LTSM for identifying CB in Arabic Twitter texts. However, the CNN + GRU and CNN + LSTM are more complex and less efficient when dealing with large-scale datasets. Chen and Cheng-Te [14] developed a Deep Neural Network model, named HENIN (HEterogenous Neural Interaction Networks) to identify CB in social media platforms. It includes three parts: (1) comment encoder, (2) post-comment co-attention technique, and (3) post-post and session-session interaction extractors. It can be observed that the learning of graph-based post-post and session-session interactions makes up the majority of the contribution to overall performance. To discover CB in tweets, a classification model using Bi-directional Gated Recurrent Unit (Bi-GRU) was suggested by Fang et al. [27] in order to learn the underlying relationship among (terms) words in combination with the self-attention mechanism to enhance the classification of CB tweets. However, the attention network's context-independent behaviour limits its learning ability

to learn all relationships between tweets. Zhao et al. [79] proposed a novel representation learning model called Semantic-Enhanced Marginalized Denoising Autoencoder (smSDA) to detect CB. This approach is capable of generating robust and discriminative representations, which are subsequently fed into SVM model.

Iwendi et al. [34] introduced a CB detection approach based on RNN and Bi-LSTM. This approach demonstrated high performance when using RNN and Bi-LSTM showed a significant efficiency. Also, CNN performs well in some cases. Akhter et al. [5] used a variety of DL models, including CNN, CLSTM [80], LSTM, BLSTM, and others, to detect abusive Urdu expressions in a social media text. Other researchers, such as [12, 6, 62, 32], and [75], used CNNs to improve CB detection. Agarwal et al. [2] employed RNN with Class Weighting and Under-Sampling, enabling the RNN model to outperform the RNN approach. This suggests that fine-tuning the parameters can improve recurrent neural network efficiency. An Attention-based RNN method was designed by Edo-Osagie et al. [24] for classifying short texts, which showed high classification accuracy. But, this method's location filtering is limited. Pitsilis et al. [59] utilized recurrent neural networks and word frequency vectors to detect hate speech. Using the RNN model, Khodabakhsh et al. [36] predicted personal life future events from tweets based on the recurrent neural network. This model is not effective in classifying highly class imbalanced data. A hybrid DL model called (Bi-GAC) was developed by Kumar and Sachdeva [38] for CB classification in social media. This model combines the benefits of the capsule network (CapsNet) and BiGRU with a self-attention encoder. Fine-grain classification in CB has not yet been investigated in this study, nor has it been expanded to multi-class cyberbullying classification.

Aind et al.[4] suggested a CB detection framework using the Reinforcement Learning model, namely Q-Bully. The proposed Q-Bully was utilized to discover cyberbullying on carious social media platforms automatically utilizing NLP with Reinforcement Learning models. This framework leads the detection process by leveraging human-like behavioural patterns utilizing integrating various NLP approaches and Reinforcement Learning. **A** new model for classifying CB posts (tweets) named transformer network-based word embedding is presented by Pericherla and Ilavarasan [58]. This approach uses the RoBERTa optimization model to construct word embedding and also uses the Light Gradient Boosting Machine for classifying posts such as tweets. This model circumvents the context-independent constraints associated with conventional word embedding approaches. Yet, this approach led to high training time in comparison with CNN model. A fine-tuning model was proposed by Tripathy et al. [68] to identify cyberbullying using ALBER. Paul and Saha [55] suggested an approach for CB detection, called CyberBERT, utilizing the BERT method. Eronen et al. [26] proposed a CB detection model utilizing linguistically supported data cleansing and the Feature Density (FD) method. In this study, authors analyzed the efficiency of FD utilizing linguistically-backed data cleansing and preprocessing techniques like Named Entity Recognition (NER), Parts of Speech (POS), stop words filtering and others for evaluating the performance of classification and dataset complexity. On the other hand, some works such as [37, 38, 40, 39, 71 and [56] presented a multi-model to discover CB in three various modalities of social media data called info-graphic, visual, and textual. Kumari et al. [40] presented a unified multi-model for CB detection in smart cities via social media. The work's major intention was to excavate social media posts to detect bullying comments, including images and text. The suggested model used to remove the separate learning modules requirement by the unified representation of both image and text. With this unified representation, a single-layer CNN is utilized to identify bullying comments. From the analysis of results, the text represented as an image was better for information encoding. Moreover, the single-layer CNN provided better outcomes

with two-dimensional representation. The drawback was that video, posts, and audio data were not considered for detection, and only image and text data were considered for detecting cyberbullying. Wang et al. [71] developed a multi-modal cyberbullying detection (MMCD), which extracts multiple information such as video, image, time, and comments from social networks. Totally 3 modules were utilized for modality features extraction and to fuse the multiple data types. BiLSTM (Bidirectional LSTM) with attention was used to extract post characteristics. Next hierarchical Attention (HA) networks were applied at the comment and word levels. Also, to encode information such as images and video, MLP (Multilayer perceptron) was used. Dual real social media datasets such as Vine and Instagram were used for experimentation. The limitations of this work do not detect different bullying categories such as (harassment, sexual, etc.). Roy and Mali [63] designed a model to detect and prevent image-based CB problems on SMPs. Initially, The deep learning-based Two dimensional CNN was utilized to develop the model. Then, the authors used transfer learning models for this study. The drawback of the suggested model it did not investigate only with text for cyberbullying detection, it focused on the combination of images and texts. Kumari et al. [40] suggested a DL-based approach in a bilingual for classifying various types of cyber aggression on social media comments.

Wang et al. [71] developed a subspace clustering approach called Fast Adaptive K-Means (FAKM) for processing high-dimensional data. The FAKM approach uses adaptive LF (loss function), which offers flexible calculation of cluster indicators suitable for processing diverse dataset distributions. This FAKM was utilized for performing feature selection and clustering, which can be suited for real-world applications. The limitation with FAKM was suitable for processing single-view clustering. Yan et al. [72] introduce an adaptive multi-view based subspace clustering for integrating heterogeneous amounts of data within the lower-dimensional feature space. Based on the cluster structure compactness, weights have been evaluated for distinct views in addition to multi-view data. The limitation noticed with multi-view clustering was the difficulty of discovering the information sharing properly.

To this extent, some significant issues and limitations were observed from the above-detailed literature review on CB classification and detection. Initially, the DL classifiers outperform ML models in terms of classification performance due to their superior accuracy when trained on large datasets. The major issue with the dataset or corpus was restricted to a certain community such as GamerGate for some of SoTA studies such as [11]. Another issue is that most of the SoTA studies of cyberbullying detection categorize the posts only as binary classification (bully and non-bully) or (normal and spam). Secondly, detecting fine-grain multi-class cyberbullying classification such as sexism, racism, insult, harassment, suicide, and non-bullying has not yet been investigated. Data representativeness and scarcity are the issues related to fairness constraints on cyberbullying detection. Thus, designing an accurate detection model that tolerates overfitting issues and enhances the accuracy of CB detection is a key challenge. Therefore, developing a model that characterizes accurate cyberbullying detection is necessary. In this article, the new FAEO-ECNN model is proposed to enhance the accuracy and efficiency of cyberbullying detection to overcome the aforementioned limitations and problems of existing DL and ML models.

3 Proposed methodology

This section presents the proposed model for CB detection from social media data. It integrated Fuzzy AEO clustering and ECNN deep learning model. The fuzzy AEO with a meta-heuristic algorithm is utilized to cluster the text in order to discover the topics. In contrast, novel deep learning (ECNN) is used to classify the type of cyberbullying into several classes. The proposed methodology consists of the following four major phases: (i) Text pre-processing, (ii) feature extraction, (iii) Unsupervised Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling and (v) classification and detection of cyberbullying using ECNN. The overall workflow of the proposed model is depicted in Fig. 1. Each of these components is highlighted in the following subsection.

3.1 Text pre-processing

Generally, social media data such as posts, tweets, and Instagram comments are short and noisy in nature. Such data may also include a vast range of anomalies such as emotions, acronyms, Elongated, typos, slang, spelling mistakes, Concatenated words, word boundary errors, which affect the original data. Thus, cleansing and pre-processing step is needed to improve short text data quality. The preprocessing [50, 53] is the first phase of the proposed CB detection model. It consists of three sub-phases: (1) Removal of noise such as punctuation/symbol removal, hashtag/mentions removal, emoticon transformation processes, and URL removal, (2) Cleansing OoV (Out of Vocabulary) such as slang modification, expansion of acronym, removal of repeated characters, and spell checking, and (3) Transformation of posts/tweets like word segmentation (tokenization), lower-case conversion, stop word filtering and stemming. These stages are performed to improve the

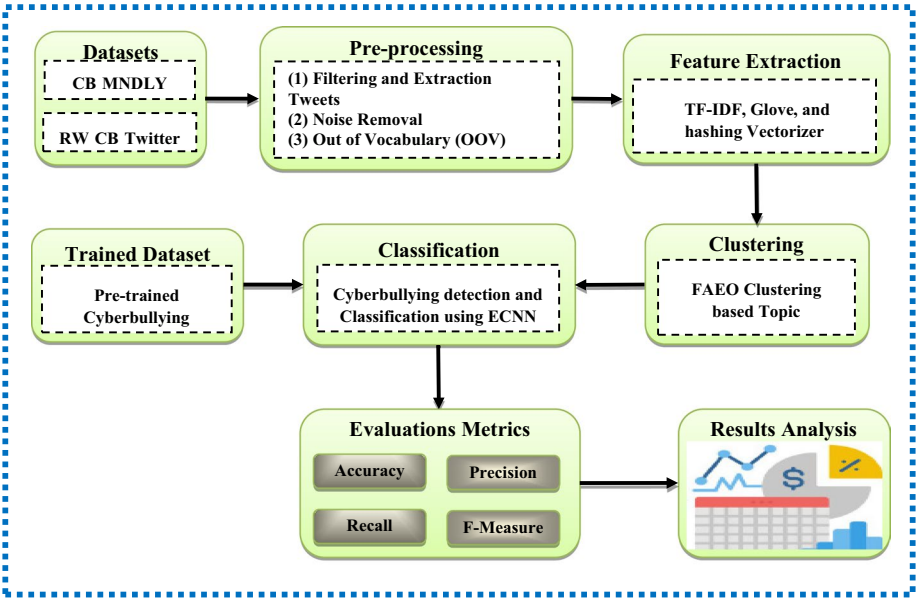


Fig. 1 Proposed Methodology of the Cyberbullying Detection Model

quality of tweets/posts and enhance classification accuracy and feature extraction. Figure 2 illustrates the social media data pre-processing stages.

(a) **Removal of Noise:** Some of the generic steps involved in this sub-phase are URLs replacement, emoticons conversion, removing extra symbols (Hashtags, mention), punctuation and symbol removal.
(b) ***Out of Vocabulary Cleansing*** (irrelevant cleansing vocabulary): This sub-phase includes the following steps: Split Concatenated words, Elongated removal, slang, and acronym modification.
(c) *Split Concatenated words:* This process is used to split the concatenated words into its components words. Most of the time, more than two words are concatenated together to reduce tweets size. For instance, the concatenated word 'IHateBlackPeople" is split into its four components words: ' I','Hate', 'Black', and 'People'.
(d) *Elongated removal:* The words with repeated characters for the given post or tweets should be shortened. Most of the users utilized these elongated words in order to express their emotions or feelings like haaaaaaate', and 'looooooove'. So removing these repeating characters in words is the most important step in pre-processing stage to get standard meaningful words.
(e) *Slang & acronym modification:* social media data such as tweets, etc., have anomalies such as acronyms and slang. Modifying the slang and acronym tends to minimize the post's characters. The words such as 'luv', 'h8', 'Gr8', 'plz' etc., are termed incorrect or non-standard words in English, which must be modified to 'love', 'hate', 'great', and 'please', respectively, by using a slang dictionary.
(f) ***Transformation of tweets/posts:*** This sub-phase contains four steps: lower case conversion, tokenization is dividing the running tweet/post into words or terms named tokens, stop-word removal, and stemming.
(g) *Lower-case conversion:* The process of transforming the terms or words of the post to a lower-case letter is called a lower-case conversion. It is more important to transform all the posts into lower-case to provide a constant format.

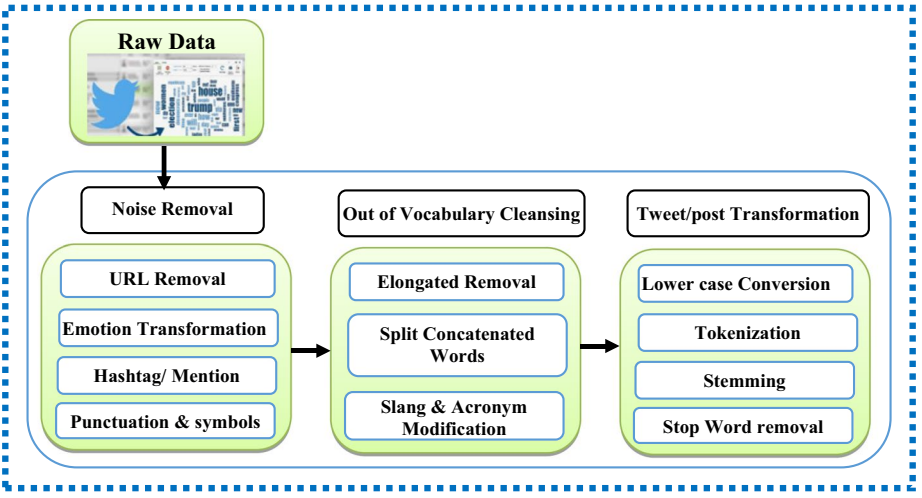


Fig. 2 Steps in Pre-processing

(h) *Tokenization:* The process of transforming a tweet into meaningful data is called tokenization. It can also be defined as the process of splitting a tweet, post, or whole text document into small units named tokens (numbers, words, or punctuations). In a post or tweet, a word is considered a token, whereas, in a paragraph, a sentence is considered as a token.

(i) *Removal of stop-words:* Posts in social media may contain frequent and popular words that contribute little to the meaning or significance of the posts. Such words are called the stop-words, so these words should be removed from the posts or documents. Some of the best examples of stop-words are prepositions, pronouns, and articles. These are the common terms existing in documents, and the words such as "with", "is", "an", "in", "the", "at", "has", "be", etc. are named as stop words.

(j) *Stemming:* it is the process of transforming the variation of words in posts or tweets into their base form stem. For instance, the words 'bully', 'bullied', 'bullying', 'bullier', 'bulliest', and 'bullies' can be converted to the base form 'bully'.

3.2 Feature Extraction

In this subsection, we provides a brief description of the various feature extraction models utilized in this study such as BoW (Bag-of-Words), TF-IDF, Glove [57, 61], and feature hashing.

3.2.1 *Bag of Word (BoW):*

This feature extraction technique is entirely based on a word/term occurrence in the post. BoW [35] is the simpler, more popular and flexible method for extracting the text features from posts (documents). The features are created using bigram, unigram, and trigram parameters. In this research, the unigram method is utilized. However, the number of total words (terms) which occur more repeatedly is defined using the BoW method.

3.2.2 *Term Frequency-Inverse Document Frequency (TF-IDF):*

TF-IDF [35] is a weighting matrix that is used as a weighting factor for Retrieving Information (RI). The importance of a term (weight + count) in each social media tweet/post is assessed using TF-IDF. This technique combines of TF and IDF measures. The TF-IDF can be defined as in Eq. (1)

$$TF - IDF = IDF(t, d) * TF(w, d) \tag{1}$$

$$TF(w, d) = \frac{Total\ number\ of\ a\ word\ \ 'w'\ occur\ in\ doc.\ 'd'}{Total\ words\ in\ doc\ 'd'} \tag{2}$$

$$IDF(t, d) = 1 + log\frac{T}{(1 + DF(t))} \tag{3}$$

where, *TF(w,d)* represents the term frequency, the count of total documents (where the term t appears) is symbolized by *DF(t)*.

Where the term frequency is indicated by $TF(w, d)$ and calculated by the total word count w in the document d by the total number of words in document d, T is indicated by the total number of posts (documents) in the entire dataset, the documents count (where the term t appears) is indicated by $DF(t)$.

3.2.3 *Global Vectors (GloVe)*

The GloVe refers to Global Vectors [57, 60]. The GloVe is characterized as the word embedding context, used to signify the numerical representations of text and signifies the semantic similarity metric among the words. In general, word embedding is used with diverse Deep Learning (DL) tasks, like syntactic parsing, entity recognition, semantic analysis, etc. The expression of GloVe can be provided in Eq. (4).

$$J = \sum_{i=1}^{m} \sum_{j=1}^{m} f(z_{ij})(u_i v_j + b_i^A + b_j^B - \log(z_{ij}))^2 \tag{4}$$

where, m represents the size of vocabulary, the scalar bias terms are symbolized as b_i^A and b_j^B. The weighting parameter is represented by $f\left(z_{ij}\right)$, which can be expressed as given in Eq. (5).

$$f\left(z_{ij}\right) = \begin{cases} (z/z_{max})^{\frac{3}{4}}, z < z_{max} \\ 1, Otherwise \end{cases} \tag{5}$$

Thus, GloVe is deliberated as the framework for illustrating word distribution, it is utilized to show vector representation from words. It can also be utilized for finding the relations among words like synonyms.

3.2.1 Feature Hashing

The process of converting the arbitrary features into indices in a matrix or vector is termed as feature hashing. The hashing trick is the other name of feature hashing. Here, a hash function is applied to the features which work using the hash values. Feature hashing is memory-efficient, fast, simple, and suited for handling high-dimensional and sparse features. Hash function determines the location of the features in a vector. The hash function maps the V (input), which is defined as:

$$f : V \rightarrow \{0, 1, \ldots \ldots, n\} \tag{6}$$

where n is an integer. In feature hashing, a hash function is used to map the feature values to indices. Though it is simple, fast, and memory-efficient but has low accuracy tradeoffs in many machine learning problems. Moreover, these feature extraction models extract the most significant features and minimize the computational complexity. Hence, these four models are analyzed, and the following clustering model is processed by considering the TF-IDF features. The value of TF-IDF represents the importance of the keywords that determine the characteristics of each topic. Some of the significant features of TF-IDF are: Easier to calculate the similarity between the two documents, Easy to extract the most eloquent keywords in the document, and very useful to measure the relevance and uniqueness of contents.

3.3 FAEO Clustering based Topic Modelling

In this sub-section, the FCM (Fuzzy C Means) [13] with Adaptive Equilibrium Optimization (AEO), named FAEO is utilized to perform topic modelling. Where FCM divides a set of posts into a number of groups in which every group contains a number of identical objects, and different groups contain different objects, the object may fit a different number of clusters. The AEO is used to optimize the fuzzifier parameter of the fuzzy model. The procedure of FCM clustering is explained as follows:

Let us assume a set of objects posts/tweets in FCM as $D = d_1, d_2, d_3, ..., d_n$ with a fuzzy set S which is a subset of D. The function of a fuzzy set can be modeled as in Eq. (7).

$$F_S : D \rightarrow [0, 1] \tag{7}$$

The main aim of fuzzy systems is to provide a logical foundation for representing uncertainty and imprecision and use it for making-decision. The whole dataset is clustered into C groups by the FCM clustering algorithm [81]. The process of clustering is performed based on the objective function J_{Min}. The objective function J_{Min} of FCM is computed for each iteration as given in Eq. (8).

$$J_{Min}(\mu, U, D) = \sum_{g=1}^{c} \sum_{j=1}^{n} \left(\mu_{gj}\right)^{w} Di_{gj}^{2} \tag{8}$$

Subject to conditions

$$0 \leq \mu_{gj} \leq 1 \tag{9}$$

$$\sum_{g=1}^{c} \mu_{gj} = 1 \tag{10}$$

$$0 \leq \sum_{j=1}^{n} \mu_{gj} \leq n \tag{11}$$

where c signifies the number of clusters, w represents the fuzzifier ($1 < w \leq \infty$), n indicates the number of posts/tweets/comments, U signifies the center vector of cluster and $Di_{gj} = dis(d_j, u_g)$ represents the distance between d_j and u_g. The membership degree is updated and expressed as given in Eq. (12).

$$u_i = \frac{\sum_{j=1}^{n} \left(\mu_{gj}\right)^{w} d_j}{\sum_{j=1}^{n} \left(\mu_{gj}\right)^{w}} \tag{12}$$

$$\mu_{gj} = \frac{1}{\sum_{i=1}^{c} \left(\frac{Di_{gj}}{Di_{ij}}\right)^{\frac{2}{w-1}}} \tag{13}$$

Here, the membership function for each post/tweet with reference to each cluster is μ_{gj} which has the membership value between [0, 1]. The term topic modelling refers to an unsupervised ML (Machine Learning) model with the ability to scan the posts or

tweets, and extracting the latent topics, and cluster the word sets automatically based on the similarity with other posts. The steps from Eqs. (8) to Eq. (13) are reiterated till J_{Min} a specific maximum number of the iterations. In according to document or post probability, post-term matrix are utilized with the weighting matrix (TF-IDF) technique (Words × posts matrix) to find the probability of posts $P(D)$. The other two matrices such as $P(W|T)$ and $P(T|D)$ which mean the probability of words for each topic and the probability of topics for each post, respectively. The probability of document j $P\left(D_j\right)$ is defined as in Eq. (14).

$$P\left(D_j\right) = \frac{\sum_{i=1}^{m}(W_i, D_j)}{\sum_{i=1}^{m}\sum_{i=1}^{n}(W_i, D_j)} \tag{14}$$

$$P\left(W_i|D_j\right) = \frac{P(W_i, D_j)}{\sum_{j=1}^{m}(W_i, D_j)} \tag{15}$$

$$P\left(D_j|T_g\right) = \frac{P(D_j, T_g)}{\sum_{j=1}^{n}(D_j, T_g)} \tag{16}$$

The $D, W, andT$ represent the documents, words, and topics, respectively. Finally, from the Eqs. (15), (16), we can get the $P(T|D)$ matrix, and it can be formulated as in Eq. (17).

$$P\left(W_i|T_g\right) = \sum_{j=1}^{n} P\left(W_i|D_j\right) \times P\left(D_j|T_g\right) \tag{17}$$

The quality of clusters is measured using J. In order to optimize the fuzzifier parameter of fuzzy models, the optimization algorithm, named Adaptive Equilibrium Optimization (AEO) [30] is used. Generally, AEO is motivated using the control volume-mass balance mechanism, which can be utilized in the estimation of equilibrium and dynamic states. AEO uses a group of elements or particles (keywords), each of which represents a concentration vector (Topic). The primary concentrations are made according to the number of dimensions and particles with random initialization. Here, the random generation of the initial Concentration Vector (CV) is expressed as given in Eq. (18).

$$V_{ci} = C_{min} + \left(C_{max} - C_{min}\right) * S \tag{18}$$

where, $i = 1, 2, 3, \ldots \ldots, n$, the initial CV in the ith particle is represented by V_{ci}, the lower bound dimension is represented by C_{min}, whereas the upper bound dimension in the problem is represented by C_{max}. In Eq. (18), S indicates a random vector, and the value of S is between [0, 1]. The overall number of particles appears within the group is denoted by n.

Like all optimization algorithms, AEO strives to improve optimization outcomes. It consistently looks for the equilibrium state of the system. As soon as it reaches equilibrium, as a result, it compels to progress toward a solution that is close to optimal for the optimization problem at hand. AEO does not know the level of concentration required to reach the equilibrium state during the optimization process. As a result, five particles must be assigned. Four of the five particles are the best in the population, while the fifth particle is the average of the other four. With the aid of these five equilibrium particles, further exploration and exploitation of an operator is conducted. The five chosen particles are

saved as vectors, which are commonly referred to as an equilibrium pool. However, the typical representation of the equilibrium pool is indicated as given in Eq. (19).

$$\vec{E}_{eq} = \left[\vec{E}_{eq(1)} + \vec{E}_{eq(2)} + \vec{E}_{eq(3)} + \vec{E}_{eq(4)} + \vec{E}_{eq(Avg)}\right] \tag{19}$$

$$\vec{E}_{eq(avg)} = \frac{\vec{E}_{eq(1)} + \vec{E}_{eq(2)} + \vec{E}_{eq(3)} + \vec{E}_{eq(4)}}{4} \tag{20}$$

where, $\vec{E}_{eq}$ signifies the equilibrium pool. The term $\vec{F}_E$ is named as the exponential term which contributes to the balance among exploration, exploitation and formulated as,

$$\vec{F}_E = e^{-(t-t_0)\vec{\lambda}} \tag{21}$$

where, λ defines the random vector in the range [0, 1], t be the time which is the iteration function, hence reduces with the number of iterations expressed as,

$$t = \left(1 - \frac{itern}{\max\_itern}\right)^{\left(\frac{itern}{\max\_itern} c_2\right)} \tag{22}$$

$$t_0 = {}^{1}\Big/_{\vec{\lambda}}\, ln\left(-c_1 sign(\overrightarrow{rn} - 0.5)\left[1 - e^{-t\vec{\lambda}}\right]\right) + t \tag{23}$$

where, *itern* specifies the present iteration and *max_itern* refers to maximum iterations, c_1, c_2 defines the constant value and the random vector $\left(\overrightarrow{rn}\right)$ lies between [0,1]. The revised form of the exponential term is formulated as in Eq. (24).

$$\vec{F}_E = c_1 sign(\overrightarrow{rn} - 0.5)\left[e^{-t\vec{\lambda}} - 1\right] \tag{24}$$

Another important parameter is the generation rate (*Gn*), which is used to gain an exact solution. The expression of *Gn* as 1st order exponential is stated as in Eq. (25).

$$\overrightarrow{Gn} = \overrightarrow{Gn_0}\left(e^{-(t-t_0)\vec{k}}\right) \tag{25}$$

where, Gn_0 specifies the initial value, k refers to decay constant. If $k = \lambda$, then the previous Eq. (25) becomes as in Eq. (26).

$$\overrightarrow{Gn} = \overrightarrow{Gn_0} * \vec{F}_E \tag{26}$$

$$\overrightarrow{Gn_0} = \left(\vec{E}_{eq} - \vec{\lambda}\vec{E}\right) * \overrightarrow{GCP} \tag{27}$$

$$\overrightarrow{GCP} = \begin{cases} 0.5rn_1, rn_2 \geq GP \\ 0, rn_2 < GP \end{cases} \tag{28}$$

where, $\overrightarrow{GCP}$ stands for Generation rate Control Parameter, $rn_1 and rn_2$ are the random numbers in the range[0,1]. Here the random numbers (rns) used are the pseudo-rns. Hence, the chaos model is introduced to replace the rns, which is further considered to have fast-convergent, stable and have lesser residual errors. Recently, chaos is presented to solve the

issues related with ergodicity and pseudo-randomness. Tent map is the most commonly used chaos to improve the EO algorithm performance.

$$x_{k+1} = \begin{cases} \frac{x_k}{0.7}, x_k < 0.7 \\ \frac{10}{3}(1 - x_k), x_k \geq 0.7 \end{cases} \tag{29}$$

Finally, based on this the new location of candidates in the next iteration can be computed using Eq. (30).

$$\vec{E} = \vec{E}_{eq} + \left(\vec{E} - \vec{E}_{eq}\right) * \vec{F}_E + \frac{\overrightarrow{Gn}}{V\vec{\lambda}}\left(1 - \vec{F}_E\right) \tag{30}$$

where V indicates the unit and $\vec{F}_E$ is the exponential term. The FAEO clustering-based topic modelling clusters the topics by the TF-IDF value leads to finding the group of topics with similar subjects in accordance with the importance of keywords. Thus, FAEO performs topic modelling and estimates the total number of topics that appears in the dataset.

3.4 FAEO-ECNN Cyberbullying Classification and Detection

Cyberbullying has created a deeper negative effect on the victim and many automated techniques are introduced by researchers to identify the conversations related to cyberbullying. CB detection is recognized as a classification issue whereas the main motive is to identify and classify each offensive or abusive post, message, and comment as a cyberbullyng or non-cyberbullying. Hence, the detection of such instances requires the use of automated intelligent systems. Recently, DL models have gained significant popularity and resulted in enhanced detection performance of multimedia-based cyberbullying events.

This section describes the FAEO-ECNN model utilized for classifying and detecting cyberbullying. Cyberbullying detection makes use of CNN [29] due to its advanced applications in object detection and text classification. After data clustering process and extracting topics, the Extended CNN (ECNN) is used in order to identify the type of cyberbullying. The input layer is the first layer. The wavelet concept is incorporated with CNN and called ECNN, mainly used for reducing the features by compressing the number of layers. The wavelet domains are mentioned as HL (high-low), LH (low–high), LL (low-low), HH (high-high).

An alternative to the conventional pooling mechanism is the Wavelet Pooling (WP) strategy, which can be used to minimize the feature dimensions. The problem of overfitting can be solved using max-pooling; however, the features can be minimized in a structurally compressed form than the normal pooling. The key modules of CNN are the convolution layer (CL) and pooling layer (PL), which undergoes consistent innovation and modification to promote the efficiency and accuracy of CNNs. PL subsamples the outcomes of the CL and reduces the spatial dimensions of the data gradually over the entire network. However, the wavelet pooling can act as a dimensionality reduction procedure which has certain applications such as regulating overfitting, parameter reduction, and improving the efficiency of computation. Figure 3 shows the architecture of ECNN.

In this study, the structure of ECNN differs from traditional CNN, such as the activation function softmax is utilizeed for multi-class classification problems. The architecture of ECNN involves the following layers: convolution, pooling, fully-connected, and output.

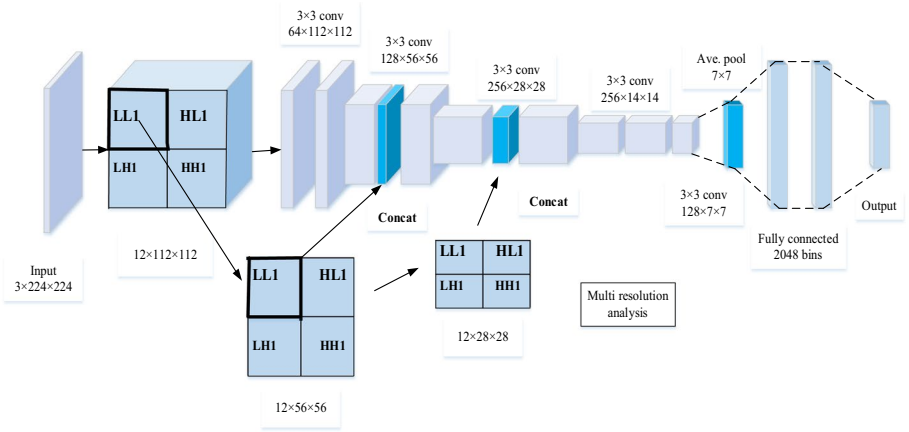


Fig. 3 Architecture of ECNN

3.4.1 Convolution layers

The output of clustering is the input to CNN. The outcome of the convolution layer is the vector of a similar number of components, which can be expressed as given in Eq. (31).

$$S = \left(s_0, s_1, s_2, \ldots, s_{n-1}\right) \in \mathcal{R}^m \tag{31}$$

$$s_i = \sum_{j \in M_i} we_j z_j \tag{32}$$

where the set of indices is represented as M_i, the weight factor is signified as we_j, it includes a bias value equivalent to 1. $\mathcal{R}^m$ signifies the cluster groups. CNN minimizes the total number of parameters and gains translational invariance. The Equation of the convolution layer can be re-written by means of a convolution operator expressed as:

$$S = z * we \tag{33}$$

where the $we = \left(we_0, we_1, we_2, \ldots, we_{m-1}\right) \in \mathcal{R}^m$. In CNN, the convolution layers typically use a diverse set of weights for a similar set of input and output as a concatenated vector representation.

3.4.2 Pooling Layers

In order to make the information simpler, the pooling layers are utilized just after the convolution layers, whereas the output is in the form of dimensions. Sub-sampling is the objective of the pooling layer. The basic dual forms of pooling are mean and max-pooling. In order to utilize the wavelets for pooling operation, the dual propagation approaches (forward, backward) need to be defined. The forward propagation (FP) makes use of second-order decomposition, whereas the backward propagation (BP) is reversible of FP. The 2-dimensional Discrete Wavelet Transform (DWT) utilized for FP, and it can be described as in Eq. (34).

$$Wav_{\psi(j_0,l,k)} = \frac{1}{\sqrt{PQ}} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} f(p,q)\psi_{(j_0,l,k)}(p,q) \tag{34}$$

$$Wav_{\phi(j,l,k)} = \frac{1}{\sqrt{PQ}} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} f(p,q)\phi_{j,l,k}^{i=h,v,d}(p,q) \tag{35}$$

where,ϕ represents the scaling function and ψ signifies the wavelet function, $f(p,q)$ signifies the position of text feature, The feature dimensions are indicated as (P,Q), and the position of wavelets is characterized as p,q,l,k. The vertical, horizontal, and diagonal scaling coefficients can be signified as $V,H,andD$. The 2-dimensional Inverse DWT utilized for BP can be described as:

$$f(p,q) = \frac{1}{\sqrt{PQ}} \sum_{p} \sum_{q} Wav_{\psi}(j_0,l,k)\psi_{(j_0,l,k)}(p,q) + \frac{1}{\sqrt{PQ}} \sum_{j=j_0}^{inf} \sum_{p} \sum_{q} Wav_{\phi}^{j}(j_0,l,k)\phi_{j,l,k}^{i}(p,q), j = j_0 = 2 \tag{36}$$

Here, the wavelet named Harr Wavelet is used, which is easy and simple to implement. The expression for scaling and wavelet function using Harr wavelet is given as:

$$\phi(t) = \begin{cases} 1, 0 \le t \le 1 \\ 0, o.w. \end{cases} \tag{37}$$

$$\psi(t) = \begin{cases} 10 \le t \le \frac{1}{2} \\ -1, \frac{1}{2} \le t \le 1 \\ 0, O.w. \end{cases} \tag{38}$$

3.4.3 Fully-connected and output layer

In this fully-connected layer, the entire feature maps are signified using one-dimensional vector representation which is then connected to the output layer. The expression for the output layer is:

$$OL_i = f\left(\sum_{j=1}^{d} z_j^{FC} we_{ij}^{OL} + b_i^{OL} \right) \tag{39}$$

where, OL_i indicates the ith unit value in OL (output layer), the weight-related to OL_i is represented as we_{ij}^{OL}, jth unit in FC(Fully-connected) layer is denoted as z_j^{FC} and bias is signified as b_i^{OL}. Since the output specifies multi-class classification, the loss function considered is the Categorical Cross-Entropy (CCE), and the activation function utilized is softmax. The CCE loss function is expressed as,

$$\begin{aligned} BCE &= -\sum_{i}^{C} t_i log\left(f(S)_i\right) \\ f(S)_i &= \frac{e^{s_i}}{\sum_j^C e^{s_j}} \end{aligned} \tag{40}$$

where, the target vector is t and the softmax activation function is $f(S)_i$. Thus, the loss function (Training loss) in the classification layer is reduced using the Rain Optimization (RO) Algorithm [47]. RO algorithm is based on the principle of rain behaviour. Using this algorithm, the problem (loss function) can be modelled using the functionality of raindrops. Here the raindrops are considered as features, and the main objective is minimizing the loss function. If dual features are near to each other, both can be connected to form a larger feature which is represented as:

$$R = \left(ra_1^m + ra_2^m\right)^{\frac{1}{m}} \tag{41}$$

where, $ra_1 and ra_2$ denote the radius and m signifies the total number of variables. The percentage of texture feature volume α absorbed in every iteration ranges from 0 to 100 and can be expressed as in Eq. (42).

$$R = \left(\alpha ra_1^m\right)^{\frac{1}{m}} \tag{42}$$

The details of the hyperparameters are described in Table 1. For the training hyperparameter, the optimal values considered are the learning rate (0.001) and epochs (150) for effective network convergence.

As the iteration number increases, the loss function and the weaker texture features disappear and increase the speed of finding the cyberbullying classification. Finally, the classification based on ECNN for cyberbullying detection classifies the cyberbully posts/texts into insult, sexism, racism, aggression and non-bullying. Thus, the optimization based loss reduction attains better outcomes without extra tuning. The pseudo code of cyberbullying classification is provided in Algorithm 1.

In recent years, online cyberbullying detection has obtained increased social importance. With technological development, students and young adults rely on the internet for various applications such as playing games, browsing, project work, job interviews etc. More often, people depend on the internet with increasing frequency for social communication. Cyberbullying remains an alarming issue all over the world among students and adolescents. Furthermore, CB may cause serious mental problems and lead to suicide, and such suicides may happen as a result of the depression and stress caused by cyberbullying events [69]. This highlights the importance of developing a method for detecting CB in social media short text messaging such as comments, posts, and tweets. Therefore, this work presents an automated cyberbullying detection model named FAEO-ECNN using a DL framework for detecting CB in SMPs. The major advantage of the proposed

Table 1 Details of proposed Hyperparameters

Hyperparameters	Value
Activation	Softmax
Learning rate	0.001
Epoch	150
Dropout rate	0.01
Loss	Categorical cross-entropy
Optimizer	RO
Batch size	32

Input: Cyberbullying Mendeley (CB-MNDLY) or Real-world Cyberbullying Twitter (RW-CB-Twitter) datasets consist of *n* posts/tweets
Output: Classification of cyberbullying data (insult, sexism, racism, aggression and non-bullying)
Begin
Step 1: **Data cleansing and Pre-processing**
1.1 Noise Removal,
1.2 Out of Vocabulary cleansing,
1.3 post-Transformation
Step 2: **Compute TF-IDF score Method for each term using Equations (2) and (3)**
Step 3: **Clustering-based topic-modelling (FAEO)**
3.1 Set the cluster number C, error $\in$ and degree of fuzziness w>1
3.2 Randomly initialize the cluster membership data μ_{gj}
3.3 Update the cluster center u_i using Equation (12)
3.4 Update the calculation of the Membership function μ_{gj} utilizing Equation (13)
3.5 Updating the fuzzifier parameter using AEO algorithm
3.6 Calculate the objective function using J_{Min} Equation (8)
3.7 Reiterated the steps from3.3 to 3.6 Eqn. (8) to Eqn. (13) are reiterated till J_{Min} a specific maximum number of the iterations
3.8 Compute the posts probability for each TF-IDF
3.9 Perform topic modelling with a probability of words for each topic and probability of topics for each document using Equations (15) to (17).
Step 4: **Classification (ECNN)**
4.1 Clustering output is resized and given as an input to CNN input layer
4.2 Pooling layer is replaced with wavelet HL (high-low), LH (low-high), LL (low-low), and HH (high-high) decompositions.
4.3 Fully Connected layer extracts the feature maps and the Training loss minimization using ROA Equations (41) and (42).
4.4 Output layer classifies the posts to the exact class.
End

Algorithm 1 Pseudo code for FAEO-ECNN cyberbullying model

FAEO-ECNN detection model is that help to mitigate committing suicide or being affected with other psychological effects because of cyberbullying activities. This work is a key factor in preventing online harassment and creating awareness among children and adults always to be cautious from their early stages. Also, this research aims to minimize the harassing instances in cyberspace.

4 Experimental Analysis

In this section, we present the performance analysis of the FAEO-ECNN model compared with the baseline cyberbullying models. The evaluation is conducted on two short-text datasets: RW-CB-Twitter dataset gathered from the platform of Twitter using Twitter API streaming and also the second dataset called CB-MNDLY dataset collected from various social media platforms by Elsafoury [25] to evaluate the superiority of the suggested cyberbullying detection model. the description of both datasets is explained in Sub-Sect. 4.1. Four baseline cyberbullying detection models, namely LSTM [34], CNN + LSTM [62], RNN [2], and BLSTM [34], are selected for the comparison with the suggested FAEO-ECNN cyberbullying model. Besides, four baseline topic modelling models, namely PTM [82], CME-DMM, BTM [18], GLTM [43], and [43] are utilized for comparing with the proposed FAEO topic modelling. The same settings of the setup parameters as those used in the considered original papers are

utilized. The experiments were carried out utilizing Pycharm IDE 2020.2.3 and Python 3.7.4, and some needed Packages, including NLTK, NumPy, TensorFlow, Keras, Sklearn, Tweepy, Scikit-learn, etc., on a system with the information: Intel Core-i7 processor with 8 Gigabytes RAM and windows 10. The datasets utilized, baseline models, the metrics considered for evaluation, and experimental results are described in the upcoming subsections.

4.1 Dataset Descriptions

In this sub-section, the datasets, which are utilized in the experimental analysis, are described briefly. The evaluation is conducted on two short-text datasets: the RW-CB-Twitter dataset gathered from Twitter and CB-MNDLY dataset gathered from various social media platforms by Elsafoury [25] to assess the superiority of the proposed cyberbullying detection model.

4.1.1 Real-world Cyberbullying Twitter (RW-CB-Twitter) Dataset

This dataset is gathered from Twitter SMP using Twitter API streaming with the help of several key terms related to cyberbullying, including Nigger, sucker, Idiot, slut, faggot, LGBTQ, donkey, moron, afraid, ass, fucking, fuck, poser, live, whale, bitch, rape, ugly, pussy, whore, and shit are some of the key terms as suggested in psychology literature [20, 54, 17], and [66]. Whereas the other key terms such as Islamic, Islam, ban, Islam, terrorist, hate, black, kill, attack, threat, racism, and evil were recommended in [77]. The RW-CB-Twitter is expanded to include the dataset that was used in [49]. The overall number of collected tweets is 435764 tweets included in the collected dataset, and out of those tweets, 20000 tweets are used and selected randomly for experiments. This dataset contains ten features: tweet date created, tweet id, user name, a screen name (account's name on Twitter), user location, text (tweet), retweet count, hashtag, number of followers, and friends. The collected tweets have numerous outliers. Because only English language tweets are required, tweets of the other languages have been filtered out, and retweets in the dataset have been removed. These tasks are conducted automatically in the pre-processing phase. Then the rest key preprocessing steps are performed as described in sub-Sect. 3.1.

4.1.2 Cyberbullying Mendeley (CB-MNDLY) dataset

It is publicly available from Mendeley data repository. The posts in this dataset is gathered from various social media platforms by Elsafoury [25], including YouTube, Twitter, Kaggle, Wikipedia, and Talk pages, corresponding to automatic cyberbullying detection. We combined all these datasets to construct one dataset called CB-MNDLY dataset. Then, we consider each dataset as one class then we labelled the dataset into 6 classes: racism, insults, aggression, sexism, toxicity, and Non-cyberbullying. The CB-MNDLY dataset consists of 448880 short text post, and out of those tweets, 30000 posts are used and selected randomly for experiments.

4.2 Baselines Approaches

This sub-section briefly describes all the baseline models of cyberbullying detection and topic modelling, which are used to compare with the proposed model.

4.2.1 Baseline Approaches for Cyberbullying Detection

- ***CNN-LSTM*** [62]: CNN-LSTM is a hybrid model that extracts higher-level phrase representations utilizing CNN and is then considered input into an LSTM for obtaining sentence representation.
- ***Bi-LSTM*** [34]: The Bi-LSTM model works on bi-direction and differs from the traditional LSTM network. It is used for detecting cyberbullying. Bi-LSTM combines two LSTMs, which retain doubled input gates, forget gates, output gates, and forward input states.
- ***RNN***-based approach [2]: used Bi-LSTM layers combined with Max-Pooling and an attention layer to capture contextual information in the text. Besides, class weighting and UnderSampling are utilized to reduce the influence of the imbalance problem in cyberbullying classification.
- ***LSTM*** [34]: LSTM is a special kind of RNN used to detect cyberbullying, which consists of three components (1) Input gate, (2) Output gate, and (3) a forget gate.

4.2.2 Baseline Approaches for Topic Modelling

- ***PTM*** [82]: PTM provides more benefits in learning topic distributions by introducing the pseudo document to self-combine many short texts without considering auxiliary contextual data. PTM obtained the highest precision and efficacy outcomes by analyzing the topic distributions of covert pseudo documents.
- ***BTM*** [18]: The BTM is utilized to discover hidden topics or themes from short texts by generating word co-occurrence patterns in the entire dataset to learn the latent topics in order to solve data sparsity issues at the document level.
- ***GLTM*** [43]: GLTM method utilized the negative sampling along with the (SGNS) skip-gram model to achieve local word embedding's and train the global word embedding's from the enormous external corpus. It can find semantic information among words by utilizing both local and global word embedding's, and it can also be used Gibbs sampler to support the semantic coherence of topics.
- ***CME-DMM*** [43]: It can discover coherent latent topics from social media short text, and it integrates word and topic embedding's through the attention strategy, which enhances the hidden topic quality.

4.3 Performance Metrics

This subsection presents various performance metrics utilized to evaluate the performance of FAEO-ECNN model in comparison with baseline cyberbullying models, like recall (sensitivity), precision, accuracy, and F1-score (F-Measure). Moreover, the other metrics which use to evaluate the performance of Topic modelling such as topic coherence which uses Normalized Pointwise Mutual Information (NPMI), purity, and Normalized Mutual Information (NMI)

used to analyze the performance of the clustering-based topic modelling. Most of these metric have been described in [52].

4.3.1 Topic coherence (TC)

TC is a metric used to compute the quality of the learning themes or topics constructed by topic models. It is an average pair-wise word semantic similarity degree among high-scoring probability formed by the topmost words of a given theme or topic. For each K topic of posts produced, the TC is applied to the topmost P words. In all our experiments, we choose $N = 10$ topmost words (terms) with high probabilities $(Wd_1, Wd_2, \ldots, Wd_p)$ as a sliding window. This measure can be calculated using NPMI, which specifies the relation between the topic k and most possible words p. Here in this experiments, the TC can be formulated as given in Eq. (43).

$$Topic\ Coherence\ (K) = \frac{2}{p(p-1)} \sum_{j=1}^{P-1} \sum_{l=j+1}^{P} \frac{log \frac{P(wd_j, wd_l)}{P(wd_j)P(wd_l)}}{-logP(wd_j, wd_l)} \tag{43}$$

where the occurrence of words wd_l and wd_j are defined as in $p(wd_l)$ and $p(wd_j)$, respectively. Whereas the terms (words) probability wd_j and wd_l co-occurring in random posts can be denoted by $p(wd_j, wd_l)$.

4.3.2 Purity

The purity metric [78] is an external metric utilized to determine the quality of the clusters. Purity can be defined as the ratio of the total number of short texts correctly clustering to all the labelled short texts in the dataset. The formula of purity can be defined as given in Eq. (44).

$$Purity = \frac{1}{n} \sum_{j=1}^{|A|} \sum_{l=1}^{|B|} max \left| a_j \cap b_l \right| \tag{44}$$

Here, the group of posts derived clusters can be represented as $A = \{a_1, a_2, a_3, \ldots, a_{|A|}\}$, while the group of ground-truth clusters in the corpus is $B = \{b_1, b_2, b_3, \ldots, a_{|B|}\}$, the number of posts in a dataset is indicated as n.

4.3.3 NMI

NMI [64] is a good metric to assess the clustering quality, and it is called an external metric because the class label of the short texts is required to determine the NMI. It computes the Mutual Information $I(A, B)$ shared among both of A and B, where the range of this metric is normalized to the value between 0 and 1 by using the mean entropy metric of clusters denotes $H(A)$ and the entropy metric of classes denotes $H(B)$. The NMI permits us to trade-off between the numbers of clusters against clustering quality. NMI can be expressed as in given Eq. (45).

$$NMI(A, B) = \frac{2 * I(A, B)}{[H(A) + H(B)]} \tag{45}$$

Here $I(A, B)$ signifies the mutual information of sets A and B, which can be calculated as in Eq. (46).

$$I(A,B) = \sum_{j=1}^{|A|}\sum_{l=1}^{|B|}\left[P(a_j \cap b_l) log \frac{P(a_j \cap b_l)}{P(a_j)P(b_l)}\right] = \sum_{j=1}^{|A|}\sum_{l=1}^{|B|}\left[\frac{\left|a_j \cap b_l\right|}{n} \log \frac{M\left|a_j \cap b_l\right|}{\left|a_j\right|\left|b_l\right|}\right] \tag{46}$$

$$H(A) = -\sum_{j=1}^{|A|} P(a_j) log P(a_j) = -\sum_{j=1}^{|A|} \frac{\left|a_j\right|}{n} log \frac{\left|a_j\right|}{n} \tag{47}$$

$$H(B) = -\sum_{l=1}^{|B|} P(b_l) log P(b_l) = -\sum_{l=1}^{|B|} \frac{|b_l|}{n} log \frac{|b_l|}{n} \tag{48}$$

Here, $P(a_j \cap b_l)$, $p(b_l)$, and $\mathrm{P}(a_j)$ indicate the probability of the post (short text) possibly appearing in both clusters a_j and b_l, possibly appear in a cluster b_l, and cluster a_j, respectively. n indicates the number of posts in the original corpus, $\left|a_j \cap b_l\right|$, $\left|a_j\right|$, and $|b_l|$ indicate the number of the posts (short texts) occurring in both clusters a_j and b_l, occurs in the cluster a_j, and occurs in b_l, respectively. The information entropy of A and B are represented by $H(A)$ and $H(B)$, respectively.

4.3.4 ARI

It is a data clustering measure that evaluates the similarity between two clusters. The mathematical expression of ARI can be expressed as in given Eq. (49),

$$\mathrm{ARI} = \frac{\sum_{\mathrm{j=1}}^{|\mathrm{A}|}\sum_{\mathrm{l=1}}^{|\mathrm{B}|}\binom{\mathrm{m_{j,l}}}{2} - \left[\sum_{\mathrm{j=1}}^{|\mathrm{A}|}\binom{\mathrm{a_j}}{2}\cdot\sum_{\mathrm{l=1}}^{|\mathrm{B}|}\binom{\mathrm{b_l}}{2}\right]/\binom{\mathrm{n}}{2}}{\frac{1}{2}\left[\sum_{\mathrm{j=1}}^{|\mathrm{A}|}\binom{\mathrm{a_j}}{2} + \sum_{\mathrm{l=1}}^{|\mathrm{B}|}\binom{\mathrm{b_l}}{2}\right] - \left[\sum_{\mathrm{j=1}}^{|\mathrm{A}|}\binom{\mathrm{a_j}}{2}\cdot\sum_{\mathrm{l=1}}^{|\mathrm{B}|}\binom{\mathrm{b_l}}{2}\right]/\binom{\mathrm{n}}{2}} \tag{49}$$

4.3.5 Accuracy

It is defined as the ratio of the correctly total predicted observations to all observations [1]. Accuracy is defined as in given Eq. (50).

$$Accuracy = \frac{T_p + T_N}{T_p + T_N + F_P + F_N} \tag{50}$$

4.3.6 Precision

The term precision is known as the ratio of the correctly predicted positive observations to the total predicted positive observations. It can be formulated as in Eq. (51).

$$Precision = P = \frac{T_p}{T_p + F_P} \tag{51}$$

4.3.7 Recall

It is called as sensitivity; it is defined as the ratio of correctly predicted positive observations to total observations when the actual class is 'Yes'. Recall is formulated as given in Eq. (52).

$$Recall = R = \frac{T_p}{T_p + F_N} \tag{52}$$

4.3.8 F-measure

It can be defined as the weighted average of recall (sensitivity) and precision values. It is formulated as given in Eq. (53).

$$F - Measure = F - M = \frac{2 * P * R}{P + R} \tag{53}$$

4.3.9 Performance Improvement Rate (PIR)

PIR describes the rate of improvement defined by the suggested model compared to the existing model. It can be expressed as in Eq. (54)

$$P_{IR} = \frac{\left| \sum_{j=1}^{n} Perf_M\left(proposed_j\right) - \sum_{j=1}^{n} Perf_M\left(existing_j\right) \right|}{\sum_{j=1}^{n} Perf_M\left(existing_j\right)} * 100 \tag{54}$$

where $j = 1, .., n$, M denotes the performance metrics such as F-Measure, accuracy, recall, and precision.

4.4 Experimental Results

This section analyses the obtained results of the FAEO-ECNN and FAEO models. The obtained results were validated based on two short text datasets: the CB-MNDLY dataset and RW-CB-Twitter dataset. We divide this section into two sub-sections: Sub-Sect. 4.4.1 discusses and analyses the experimental results of the FAEO clustering-based topic modelling approach. Sub-Sect. 4.4.2 discusses experimental results of the proposed FAEO-ECCN cyberbullying detection model.

4.4.1 FAEO Clustering-based Topic Modelling Results

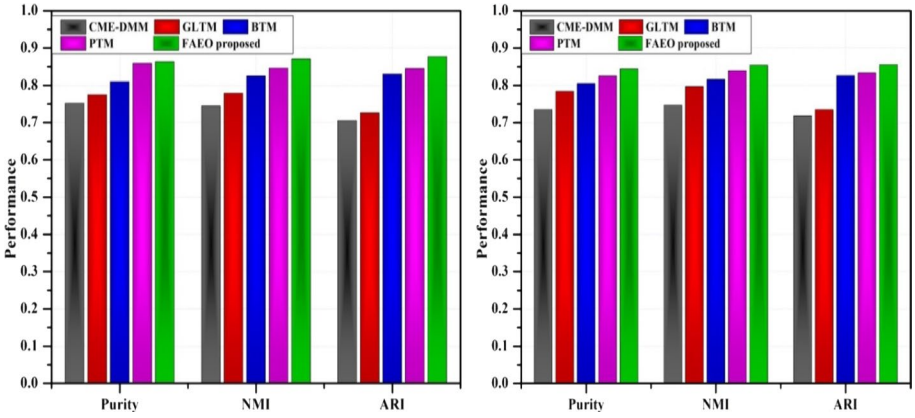
The experimental results of the FAEO clustering-based topic modeling are compared with baseline models such as PTM [82], BTM [43], GLTM [18], and CME-DMM [43]. The performance of the FAEO model is assessed over the CB-MNDLY and RW-CB-Twitter datasets based on topic coherence, purity, and NMI metrics. We fixed the number of iterations as 1000 for both existing and proposed methods. The hyperparameters of all the considered models are initialized as in the original paper. The values of hyperparameters are

Table 2 Performance results of topic coherence, purity, and NMI on FAEO Topic modelling over both datasets

Datasets	Technique/Metrics	Topic Coherence	Purity	NMI	ARI
Cyberbullying Mendeley	FAEO proposed	0.426	0.863	0.871	0.877
	PTM	0.379	0.859	0.846	0.845
	BTM	0.366	0.810	0.825	0.830
	GLTM	0.355	0.775	0.779	0.726
	CME-DMM	0.348	0.752	0.745	0.705
Real-world Twitter	FAEO proposed	0.409	0.844	0.854	0.855
	PTM	0.356	0.825	0.839	0.833
	BTM	0.341	0.805	0.816	0.826
	GLTM	0.332	0.784	0.797	0.735
	CME-DMM	0.326	0.735	0.747	0.718

$\beta = 0.01$ for all models, $\alpha = 0.1$ for PTM, whereas we set $\alpha = 50/k$ for both GLTM and BTM. We fixed $\lambda = 0.1$ and $\lambda = 0.5$ for PTM and GLTM models. The number of topics is 10. The overall findings of the FAEO model and existing models on both datasets are presented in Table 2.

Purity, NMI, and ARI Results over CB-MNDLY Dataset This subsection illiterates the obtained results of the FAEO model compared to other existing topic models on CB-MNDLY dataset as presented in Fig. 4a. The performance of FAEO approach is assessed on CB-MNDLY dataset in terms of NMI, ARI, and purity metrics. The purity value of the FAEO topic model is 0.863, which is the best result compared with baselines topic modelling approaches. The purity values of the existing models PTM, BTM, GLTM, and CME-DMM are 0.859, 0.810, 0.775, and 0.752, respectively. In according to the NMI metric, the value of NMI in FAEO is 0.871 whereas the NMI values of the existing models are PTM (0.846), BTM (0.825), GLTM (0.779), and CME-DMM (0.745). Similarly, the FAEO model's result in ARI is 0.877, which is the optimal result in comparison with the baseline model. The ARI value of the existing models PTM, BTM, GLTM, and CME-DMM are 0.845, 0.830, 0.726, and 0.705. It can be noted that from Fig. 4a, the FAEO model



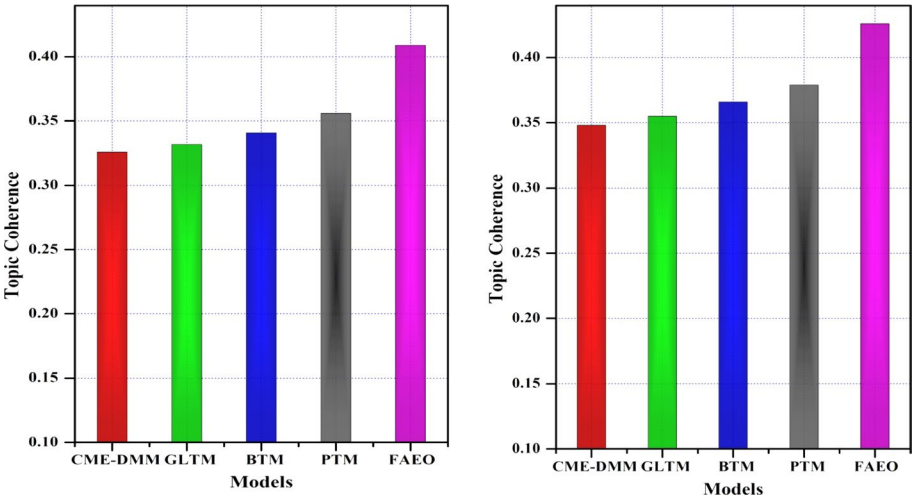
(a) Cyberbullying Mendeley Dataset (b) Real-world Twitter Dataset

Fig. 4 Performance results on both datasets in terms of ARI, NMI and purity

outperforms the existing topic modelling approaches. The GLTM has got the second best values, followed by the CME-DMM model, while BTM has got the fewer values among all the existing models based on two performance metrics NMI, purity, and ARI.

Purity, NMI, and ARI Results over RW-CB-Twitter dataset This part presents the results of the FAEO model compared with baseline topic modelling approaches based on NMI, purity, and ARI on the RW-CB-Twitter dataset, as depicted in Fig. 4b. The purity value of the suggested FAEO topic model is 0.844, which is the optimal result in comparison with the existing models, whereas the purity values of the existing models are PTM (0.825), BTM (0.805), GLTM (0.784), and CME-DMM (0.735). Besides, the proposed FAEO topic model has obtained a 0.854 value of NMI over the RW-CB-Twitter dataset. According to ARI values, the experimental results of FAEO in terms of ARI are 0.855, which has the highest value compared with the comparison models. The existing models, such as PTM, BTM, GLTM, and CME-DMM, obtained ARI values of 0.833, 0.826, 0.735, and 0.718, respectively. From Fig. 4b, it is observed that the FAEO model has optimal outcomes in comparison with existing models, while the PTM model has scored the second-order value of ARI, NMI, and purity among all the existing models.

Topic Coherence Results over CB-MNDLY and RW-CB-Twitter datasets The TC of the suggested FAEO topic modeling is evaluated utilizing the NPMI metric in order to evaluate the topics quality over two datasets: cyberbullying Mendeley and Real-world Twitter datasets. The proposed FAEO performance outperforms the existing Topic modeling approaches models in terms of TC over both datasets. Hence the topic coherence of the FAEO model is 0.409 over RW-CB-Twitter dataset, while the topic coherence of the baseline topic models PTM, BTM, GLTM, and CME-DMM are 0.356, 0.341, 0.332, and 0.326, respectively, as shown in Fig. 5a. Likewise, from Fig. 5b, the FAEO has achieved coherence 0.426 over the cyberbullying Mendeley, higher than the existing models, while the performance of the current models: PTM, BTM, GLTM, and CME-DMM are 0.379, 0.366, 0.355, and 0.348, respectively. It can be concluded that from Fig. 5, the FAEO topic modeling has obtained the optimal results than other models over both datasets. In contrast,



(a) RW-CB-Twitter Dataset (b) CB-MNDLY Dataset

Fig. 5 Topic coherence results on both datasets with 10 topics

the CME-DMM model has less TC value on both datasets than all other models, including the proposed model.

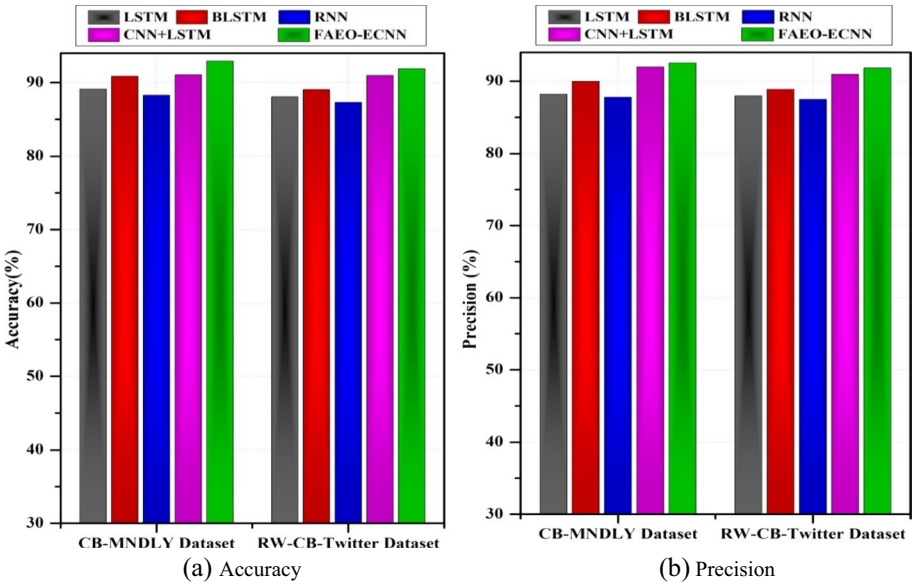
4.4.2 FAEO-ECNN Cyberbullying detection Results

This subsection presents the results of the suggested FAEO-ECNN cyberbullying detection in comparison with baseline models, namely LSTM [34], CNN + LSTM [62], RNN [2], and Bi-LSTM [34]. Cyberbullying detection evaluations are carried out utilizing F-score, precision, recall, and accuracy. The obtained results were validated based on CB-MNDLY and RW-CB-Twitter datasets**.** The overall performance results of the proposed of FAEO-ECNN Cyberbullying detection compared with all existing models are presented in Table 3.

Table 3 Performance results of accuracy, F-measure, recall, precision of Cyberbullying detection on both datasets

Datasets	Technique	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
CB-MNDLY	LSTM	89.11	88.19	88.36	88.27
	BLSTM	90.86	89.97	89.34	89.65
	RNN	88.27	87.76	86.90	87.33
	CNN + LSTM	91.04	91.95	90.57	91.26
	FAEO-ECNN	**92.91**	**92.53**	**92.28**	**92.40**
RW-CB-Twitter	LSTM	88.03	87.96	88.01	87.98
	BLSTM	89.04	88.85	88.83	89.84
	RNN	87.29	87.45	86.40	86.92
	CNN-LSTM	90.95	90.92	90.96	90.94
	FAEO-ECNN	**91.89**	**91.81**	**91.32**	**91.56**

Bold values illusrtrate the values obtained by proposed FAEO-ECNN technique

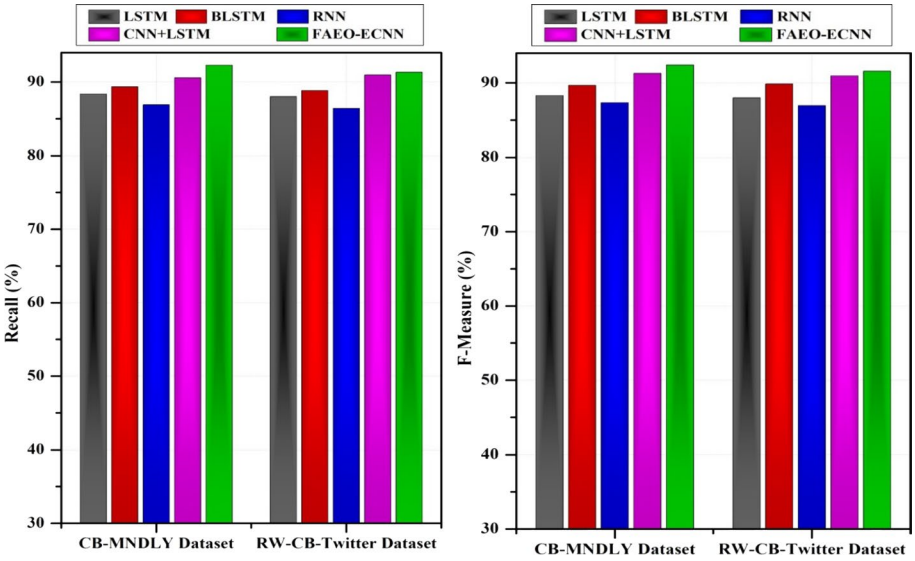


(a) Accuracy (b) Precision

Fig. 6 Performance evaluation of the cyberbullying detection in terms of overall precision and accuracy

Accuracy Results The suggested FAEO-ECNN is assessed using the accuracy metric over CB-MNDLY and RW-CB-Twitter datasets. The FAEO-ECNN model's performance outperforms the existing cyberbullying methods with respect to accuracy on both datasets. Hence the accuracy of the FAEO-ECNN model is 92.91% on CB-MNDLY dataset, while the accuracy of the baseline models LSTM, BLSTM, RNN, CNN + LSTM are 89.11%,90.86%, 88.27%, and 91.04%, respectively, as illustrated in Fig. 6a and Table 3. Similarly, the FAEO-ECNN model has achieved 91.89% on RW-CB-Twitter dataset, higher than the existing models, while the performance of the current models: LSTM, BLSTM, RNN, CNN + LSTM are 88.03, 89.04%, 87.29%, and 90.95, respectively. It can be noticed that the performance of FAEO-ECNN has more optimal results compared with existing models in terms of accuracy, as depicted in Fig. 6a.

Precision Results Figure 6b displays the comparison results of the suggested FAEO-ECNN model with other baseline cyberbullying detection methods in terms of precision on the RW-CB-Twitter dataset and another dataset namely CB-MNDLY. As illustrated in Fig. 6b, the proposed FAEO-ECNN model has got the highest precision of 92.53% on the CB-MNDLY dataset, while other existing methods CNN + LSTM, RNN, BLSTM, LSTM have obtained 91.95%, 87.76%, 89.97%, and 88.19%, respectively. That means the precision of the FAEO-ECNN model outperforms the existing models. Similarly, the obtained findings of the suggested model over RW-CB-Twitter dataset are 91.81% value of precision as provided in Table 3, which is the optimal result, compared with the baseline models. However, as depicted in Table 3, the existing models like CNN + LSTM, RNN, BLSTM, and LSTM, acquired lower values than FAEO-ECNN 90.92%, 87.45%, 88.85%, and 87.96%, respectively, on the real-world cyberbullying Twitter dataset. From Fig. 6b, it can be concluded that the precision of FAEO-ECNN has an optimal result than other existing



(a) Recall (b) F-Measure

Fig. 7 Performance evaluation of the cyberbullying detection in terms of overall recall and F-measure

models, and the CNN + LSTM has obtained the second-order optimal among all the considered current models. In contrast, RNN has scored less precision than all models.

Recall Results The overall recall of the FAEO-ECNN model with baseline cyberbullying detection approaches is depicted in Fig. 7a. It can be noticed that from Fig. 7a, the performance of the FAEO-ECNN approach over real-word cyberbullying Twitter dataset scored 91.32%, which is the highest result compared with all baseline models. The recall of the baseline models LSTM, BLSTM, RNN, CNN + LSTM 88.01%, 88.83%, 86.40%, and 90.96%, respectively, as illustrated in Fig. 7a. Likewise, the FAEO-ECNN achieved 92.28%, with the CB-MNDLY dataset as the best recall value compared with current models LSTM (88.36%), BLSTM (89.34%), RNN (86.90%), and CNN + LSTM (90.57%), respectively. Finally, it is concluded that the performance of the FAEO-ECNN model in both datasets outperforms the current approaches in terms of recall.

F-Measure Results The suggested FAEO-ECNN is evaluated utilizing F-Measure (F-score) metric over two datasets. The FAEO-ECNN model's performance outperforms the baseline cyberbullying models in terms of F-Measure over the CB-MNDLY dataset and another dataset namely RW-CB-Twitter. Hence the F-Measure value of the FAEO-ECNN model is 91.56% over the RW-CB-Twitter dataset, while the F-Measure of the baseline approaches LSTM, BLSTM, RNN, CNN + LSTM are 87.98%, 89.84%, 86.92%, and 90.94%, respectively, as shown in Fig. 7b and Table 3. Likewise, the proposed FAEO-ECNN has achieved 92.40% over the CB-MNDLY, higher than the existing models, while the performance of the current models: LSTM, BLSTM, RNN, CNN + LSTM are 88.27, 89.65%, 87.33%, and 91.26%, respectively. As illustrated in Fig. 7b, the FAEO-ECNN outperforms baseline models in terms of F-Measure across both datasets. In contrast, the RNN

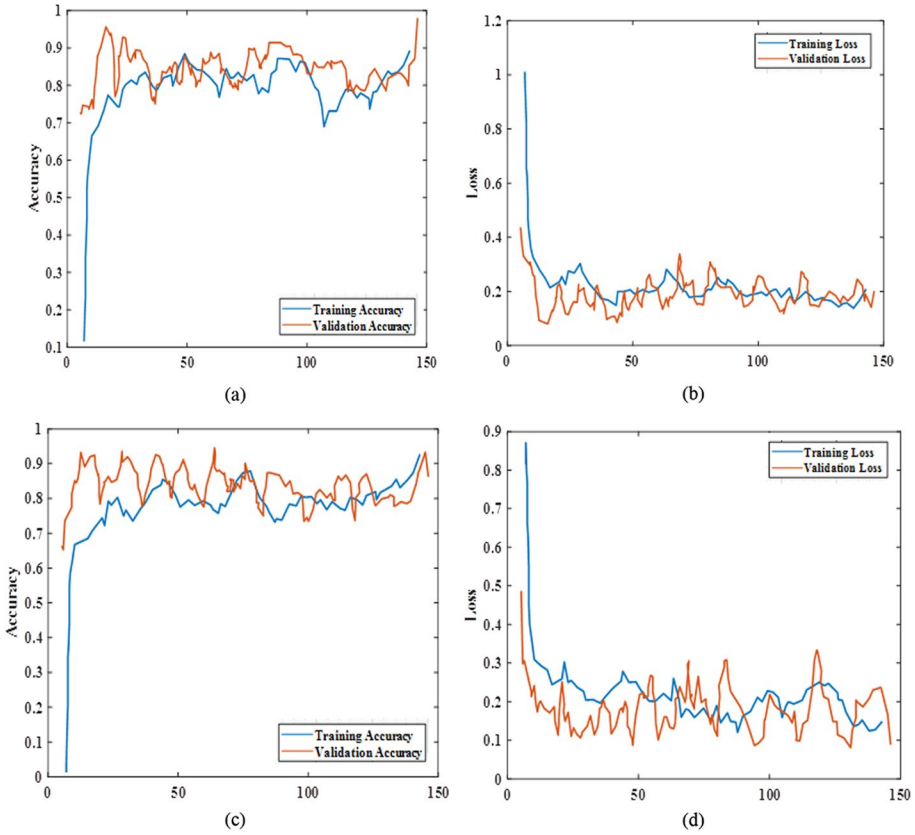


Fig. 8 Accuracy and Loss curves over CB-MNDLY and RW-CB-Twitter

model has less F-measure value on both datasets than all other models, including the proposed model.

Loss Curves and Accuracy This subsection presents the accuracy and the loss curves with different number of epoch value 0 to 150. Figure 8 illustrates the accuracy and loss plots of CB-MNDLY and RW-CB-Twitter datasets. Where Fig. 8a and b represents the accuracy and loss curves of the CB-MNDLY dataset, While Fig. 8c and d illustrates RW-CB-Twitter datasets' accuracy and loss curves.

5 Discussion

This section discusses and presents the proposed model's improvement rate over the considered State-of-The-Art (SoTA) models on CB-MNDLY and the other dataset namely RW-CB-Twitter in terms of F-measure, precision, accuracy, and recall. To show the improvement of suggested FAEO-ECNN model over baseline, we have used the PIR metric. The overall PIR is computed by the performance of the suggested FAEO-ECNN model in comparison with the current models. Equation (54) is used to calculate the PIR of the

Table 4 FAEO-ECNN performance improvement rate (%) of F-Measure, precision, recall, and Accuracy over two datasets

Metric	Models	CB-MNDLY Dataset					RW-CB-Twitter Dataset				
		LSTM	BLSTM	RNN	CNN+LSTM	FAEO-ECNN	LSTM	BLSTM	RNN	CNN+LSTM	FAEO-ECNN
Accuracy	*OveralAccuracy*	89.11	90.86	88.27	91.04	**92.91**	88.03	89.04	87.29	90.95	**91.89**
	PIR over LSTM	-	1.96	-0.94	2.16	**4.26**	-	1.15	-0.84	3.32	**4.38**
	PIR over BLSTM	-	-	-2.85	0.20	**2.26**	-	-	-1.97	2.15	**3.20**
	PIR over RNN	-	-	-	3.14	**5.25**	-	-	-	4.19	**5.27**
	PIR over CNN+LSTM	-	-	-	-	**2.05**	-	-	-	-	**1.03**
Precision	*OveralPrecision*	88.19	89.97	87.76	91.95	**92.53**	87.96	88.85	87.45	90.92	**91.81**
	PIR over LSTM	-	2.01	-0.49	4.26	**4.92**	-	1.01	-0.58	3.37	**4.37**
	PIR over BLSTM	-	-	-2.46	2.20	**2.85**	-	-	-1.58	2.33	**3.33**
	PIR over RNN	-	-	-	4.77	**5.44**	-	-	-	3.97	**4.99**
	PIR over CNN+LSTM	-	-	-	-	**0.63**	-	-	-	-	**0.98**
Recall	*OveralRecall*	88.36	89.34	86.90	90.57	**92.28**	88.01	88.83	86.40	90.96	**91.32**
	PIR over LSTM	-	1.11	-1.65	2.50	**4.44**	-	0.93	-1.83	3.35	**3.76**
	PIR over BLSTM	-	-	-2.73	1.38	**3.29**	-	-	-2.74	2.40	**2.80**
	PIR over RNN	-	-	-	4.22	**6.19**	-	-	-	5.28	**5.69**
	PIR over CNN+LSTM	-	-	-	-	**1.89**	-	-	-	-	**0.40**
F-Measure	*OveralF − Measure*	88.27	89.65	87.33	91.26	**92.40**	87.98	89.84	86.92	90.94	**91.56**
	PIR over LSTM	-	1.56	-1.07	3.39	**4.68**	-	**2.11**	-1.20	3.36	**4.07**
	PIR over BLSTM	-	-	-2.58	1.80	**3.07**	-	-	-3.25	1.22	**1.91**
	PIR over RNN	-	-	-	4.50	**5.81**	-	-	-	4.62	**5.34**
	PIR over CNN+LSTM	-	-	-	-	**1.25**	-	-	-	-	**0.68**

Bold values illusrtrate the values obtained by proposed FAEO-ECNN technique

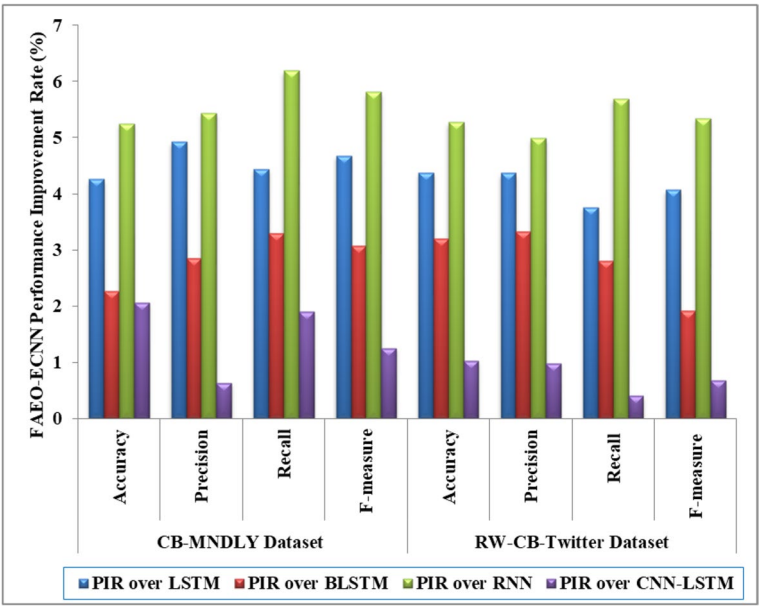


Fig. 9 PIR of FAEO-ECNN model in comparison with existing models

suggested model. Table 4 illustrates the PIR of FAEO-ECNN over the existing models on both two datasets.

This study considers four baselines cyber-bullying detection approaches for for the evaluation process. In case of accuracy, the suggested FAEO-ECNN produced 4.26%, 2.26%, 5.25%, and 2.05% of accuracy improvements on LSTM, BLSTM, RNN, and CNN + LSTM models over CB-MNDLY and the other dataset namely RW-CB-Twitter as presented in Table 4, respectively. While it showed 4.38%, 3.20%, 5.27%, and 1.03% of accuracy improvements over RW-CB-Twitter dataset. In the same way, The PIR of FAEO-ECNN model in terms of precision on the CB-MNDLY dataset are 4.92%, 2.85%, 5.44%, and 0.63%, over the baseline approaches LSTM, BLSTM, RNN, and CNN + LSTM, respectively. Whereas the improvements of the suggested model in terms of precision over baseline models LSTM, BLSTM, RNN, and CNN + LSTM are 4.37%, 3.33%, 4.99%, and 0.98%, respectively, over RW-CB-Twitter dataset. Similarly, The suggested FAEO-ECNN shows 4.44%, 3.29%, 6.19%, and 1.89% of recall improvements over LSTM, BLSTM, RNN, and CNN + LSTM approaches on CB-MNDLY dataset, respectively whereas it generates 3.76%, 2.80%, 5.69%, and 0.40% of recall improvements on LSTM, BLSTM, RNN, and CNN + LSTM models over RW-CB-Twitter dataset. Lastly, the PIR of the suggested model in terms of F-Measure on the CB-MNDLY dataset are 4.68%, 3.07%, 5.81%, and 1.25% over the LSTM, BLSTM, RNN, and CNN + LSTM. Whereas the PIR of the FAEO-ECNN model in terms of F-Measure on RW-CB-Twitter dataset are 4.07%, 1.91%, 5.34%, and 0.68% over baselines models LSTM, BLSTM, RNN, and CNN + LSTM as provided in Table 4. Figure 9 presents the PIR of the FAEO-ECNN model over existing models clearly.

Therefore, the overall PIR of the FAEO-ECNN model proves its effectiveness in detecting CB from social media data. In summary, we observe that the proposed FAEO-ECNN model has achieved the optimal results for detecting CB over both datasetsin terms of all

Table 5 Comparative analysis on cyberbullying classification

Author Name & Ref No	Technique	Dataset	Class	Performance (%)
FAEO-ECNN (Proposed)		CB-MNDLY	6	Accuracy (92.91%)
		RW-CB-Twitter	5	Accuracy (91.89%)
Balakrishnan et al. [11]	J48	#Gamergate	4	Accuracy (91.88%)
Kumari et al. [40]	Single-layer CNN	Real-time	3	Recall (74%)
Lu et al. [45]	CNN	Chinese Weibo	3	Precision (79.0%)
H. Chen and Li [14]	HENIN	Vine, Instagram	2	Accuracy (80.4%), (90.2%)
	Multichannel deep learning	Kaggle	2	Accuracy (87.99)
Wang et al. [71]	MMCD	Vine	2	Accuracy (83.8%)
Roy and Mali [63]	2DCNN	Social media data	2	Accuracy (89%)

performance measures. In addition, the proposed FAEO topic modelling approach has the optimal results compared with the existing topic modeling approaches in terms of ARI, purity, topic coherence, and NMI. Therefore, the proposed FAEO-ECNN model can be an efficient model for detecting CB in social media. Where the efficient solution was attained in the proposed FAEO-ECNN model, which can be attributed to the utilize of topic modeling for discovering the hidden (latent) topics from social media datasets and generate clusters of words. This signifies the effect of topic modelling on the ECNN performance based CB detection model. Besides, combining the Wavelet Pooling with CNN architecture reduces the dimensionality problem and minimizes the loss in CNN by utilizing a meta-heuristic Rain Optimization (RO) algorithm. In addition in this subsection, we also have included the comparative analysis of specific comparison study among similar types of work. Table 5. depicts the comparative study on cyberbullying classification.

Bullying via social media posts seems to be a growing trend and is perceived to have a severe negative impact. It can be considered a destructive and threatening act which can be the major cause of life-long problems for victims. The problem behind cyberbullying is the automated detection of cyberbullying from Twitter data. In addition to that, it is really hard to identify the type of cyberbullying. The cyberbullying type identification is one of the serious issues which create dangerous long-term effects on victims. Cyberbullying issues have multiple aspects, which bring self-doubt, harm, depression, and insecurities, disrupt the peace of mind, spread false rumors and destroy victims' lives. Existing techniques on cyberbullying detection exhibit several limitations in which a smaller dataset is utilized for testing the performance, and fewer words related to 'curse' are considered. The detection accuracy is quite low and the approaches used can only detect if the post is related to one of the cyberbullying categories, such as harassment or non-harassment. Sometimes, the chosen dataset is limited to tweet categorizes and size, affecting cyberbullying detection. The major issue is the Twitter data may not have tweets related to the cyberbullying categories like bystanders and victims. Existing studies on cyberbullying detection identify the tweet only as bully or non-bully. The different bullying categories (aggression, sexual, etc.) are not detected. The Twitter dataset used for multi-class imbalance classification is content-specific, which must be enhanced for each content. One of the drawbacks is that other social networking platforms are not used. The major issue is with the dataset, which

is limited to any community, so most of the tweets are classified into binary form (normal or spam). This work presents textual information-based cyberbullying detection using ECNN to overcome the above stated issues. Thus, compared to the baseline approaches, the proposed FAEO-ECNN based cyberbullying detection solves the problem of multi-class classification by effectively tolerating overfitting issues, computation complexity and accuracy challenges.

6 Conclusion

This article proposed an effective CB classification and detection model named FAEO-ECNN to detect CB from social media short text data. The proposed approach integrates a Fuzzy Adaptive Equilibrium Optimization (FAEO) clustering-based topic modelling and ECNN to enhance the accuracy of detecting CB. In this study, data cleansing is achieved in the pre-processing stage. Then, the features are extracted utilizing TF-IDF feature extractor; next, FAEO creates the word clusters by examining the text data. Finally, the ECNN performed classification under different CB categories such as insult, sexism, racism, aggression, and etc. The proposed ECNN model is evaluated using two social media datasets: CB Mendeley (CB-MNDLY) and another dataset gathered from Twitter namely RW-CB-Twitter dataset, which proved to be an effective classification technique gaining a higher accuracy of 92.91% and 91.89%, respectively, and F-Measure of 92.40% and 91.56 over both datasets, respectively. In addition, more simulations were carried out on dual datasets to evaluate the improved detection performance of FAEO-ECN approach. The main advantage of this approach is the early and accurate CB detection that helps schools and parents easily identify the problem with the students and take proper action. The limitations of this research are ECNN based CB detection is only limited to the English language and it will be extended to work with a multi-language dataset. The proposed FAEO-ECNN model mainly focus on detecting CB from textual content only, other type of media such as (images, video, and audio) are not considered Therefore, these media type such as images, audio, and video are still open research area, and future work aims to develop multi-modal CB detection with multi-language datasets. Besides, we aim to categorize and identify CB of Twitter data in a real-time stream and it is considered promising for future research direction.

Authors Contributions Belal Abdullah Hezam Murshed: Methodology, Conceptualization, Validation, Data curation, Formal analysis, Investigation, Software, Visualization, Writing—Original Draft, Writing—Review & Editing. Suresha: Supervision, Methodology, Resources, Data curation. Jemal Abawajy: Methodology, Supervision, Formal analysis, Data curation, Writing—Review & Editing, Supervision. Mufeed Ahmed Naji: Investigation, Formal analysis, Data curation, Writing—Review & Editing. Hudhaifa Mohammed Abdulwahab: Resources, Writing—Review & Editing. Fahd A Ghanem: Resources, Writing—Review & Editing.

Funding No funding is provided for the preparation of the manuscript.

Data Availability No data Availability.

Declarations

Conflict of Interest Authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to Publish. Reviewer and Editors can publish this work.

References

1. Abdulwahab HM, Ajitha S, Saif MAN (2022) Feature selection techniques in the context of big data: taxonomy and analysis. Appl Intell 52:13568–13613. https://doi.org/10.1007/s10489-021-03118-3
2. Agarwal A, Chivukula AS, Bhuyan MH, Jan T, Narayan B, and Prasad M (2020) Identification and Classification of Cyberbullying Posts: A Recurrent Neural Network Approach Using Under-Sampling and Class Weighting. in Yang H. et al. (eds) Neural Information Processing ICONIP 2020. Communications in Computer and Information Science. Cham: Springer, Cham, vol. 1333, pp. 113–120. https://doi.org/10.1007/978-3-030-63823-8_14.
3. Agrawal S, Awekar A (2018) Deep learning for detecting cyberbullying across multiple social media platforms. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Cham, vol. 10772 LNCS, pp. 141–153. https://doi.org/10.1007/978-3-319-76941-7_11
4. Aind AT, Ramnaney A, and Sethia D (2020) Q-Bully: A reinforcement learning based cyberbullying detection framework, in 2020 International Conference for Emerging Technology, INCET 2020. IEEE, pp. 1–6. https://doi.org/10.1109/INCET49848.2020.9154092
5. Akhter MP, Jiangbin Z, Naqvi IR, AbdelMajeed M, Zia T (2021) Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimed Syst, (0123456789). https://doi.org/10.1007/s00530-021-00784-8
6. Al-Ajlan MA, Ykhlef M (2018) Optimized Twitter Cyberbullying Detection based on Deep Learning. in 2018 21st Saudi Computer Society National Computer Conference (NCC). IEEE, pp. 1–5. https://doi.org/10.1109/NCG.2018.8593146
7. Alam KS, Bhowmik S and Prosun PRK. (2021) Cyberbullying Detection: An Ensemble Based Machine Learning Approach. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE. IEEE, pp. 710–715. https://doi.org/10.1109/ICICV50876.2021.9388499
8. Alduailaj AM, Belghith A (2023) Detecting Arabic Cyberbullying Tweets Using Machine Learning. Mach Learn Knowl Extract 5(1):29–42. https://doi.org/10.3390/make5010003
9. Al-Hassan A, Al-Dossari H (2021) Detection of hate speech in Arabic tweets using deep learning. Multimed Syst 28:1963–1974. https://doi.org/10.1007/s00530-020-00742-w
10. Balakrishnan V, Khan S, Arabnia HR (2020) Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput Secur 90:101710. https://doi.org/10.1016/j.cose.2019.101710
11. BalakrishnanV Khan S, Fernandez T, Arabnia HR (2019) Cyberbullying detection on twitter using Big Five and Dark Triad features. Person Individ Differ 141(September 2018):252–257. https://doi.org/10.1016/j.paid.2019.01.024
12. Banerjee V, Telavane J, Gaikwad P, Vartak P (2019) Detection of Cyberbullying Using Deep Neural Network. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, pp. 604–607. https://doi.org/10.1109/ICACCS.2019.8728378
13. Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm'. Comput Geosci 10(2–3):191–203. https://doi.org/10.1109/IGARSS.1988.569600
14. Chen H, Li C-T (2020) HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media. Available at: http://arxiv.org/abs/2010.04576
15. Chen Z, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013) Leveraging multi-domain prior knowledge in topic modelsIJCAI Twenty-Third International Joint Conference on Artificial Intelligence International Joint Conference on. Artif Intell 13:2071–2077
16. Chen J, Yan S, Wong KC (2020) Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. Neural Comput Appl 32(15):10809–10818. https://doi.org/10.1007/s00521-018-3442-0
17. Cheng L, Li J, Silva Y, Hall D, Liu H (2019) PI-Bully: Personalized Cyberbullying Detection with Peer Influence. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Calif Int Joint Conf Artif Intell Organ, pp. 5829–5835. https://doi.org/10.24963/ijcai.2019/808

18. Cheng X, Yan X, Lan Y, Guo J (2014) BTM: Topic Modeling over Short Texts. IEEE Trans Knowl Data Eng 26(12):2928–2941. https://doi.org/10.1109/TKDE.2014.2313872
19. Chia ZL, Ptaszynski M, Masui F, Leliwa G, Wroczynski M (2021) Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. Inf Process Manag 58(4):102600. https://doi.org/10.1016/j.ipm.2021.102600
20. Cortis K, Handschuh S (2015) Analysis of cyberbullying tweets in trending world events. In Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business. New York, NY, USA: ACM, pp. 1–8. https://doi.org/10.1145/2809563.2809605
21. Dadvar M, Eckert K (2018) Cyberbullying Detection in Social Networks Using Deep Learning Based Models: A Reproducibility Study. In arXiv preprint arXiv:1812.08046. arXiv:1812.08046
22. Dadvar M, Trieschnigg D, Ordelman R, Jong FD (2013) Improving Cyberbullying Detection with User Context. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7814 LNCS, pp. 693–696. https://doi.org/10.1007/978-3-642-36973-5_62
23. Dalvi RR, Baliram Chavan S, Halbe A (2020) Detecting A Twitter Cyberbullying Using Machine Learning. In: Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020. ICICCS, pp 297–301. https://doi.org/10.1109/ICICCS48265.2020.9120893
24. Edo-Osagie O, Lake I, Edeghere O, Iglesia BDL (2019) Attention-Based Recurrent Neural Networks (RNNs) for Short Text Classification: An Application in Public Health Monitoring. In 15th International Work-Conference on Artificial Neural Networks. Springer, Cham., pp. 895–911. https://doi.org/10.1007/978-3-030-20521-8_73
25. Elsafoury F (2020) Cyberbullying datasets. Mendeley Data V1. https://doi.org/10.17632/jf4pzyvnpj.1
26. Eronen J, Ptaszynski M, Masui F, Pohl A, Leliwa G, Wroczynski M (2021) Improving classifier training efficiency for automatic cyberbullying detection with Feature Density. Inf Process Manage 58(5):02616. https://doi.org/10.1016/j.ipm.2021.102616
27. Fang Y, Yang S, Zhao B, Huang C (2021) Cyberbullying Detection in Social Networks Using Bi-GRU with Self-Attention Mechanism. Inf 12(4):171. https://doi.org/10.3390/info12040171
28. Galán-García P, Puerta JGDL, Gómez CL, Santos I, Bringas PG (2014) Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. Logic J IGPL 24(1):42–53. https://doi.org/10.1093/jigpal/jzv048
29. Gambäck B, Sikdar UK (2017) Using Convolutional Neural Networks to Classify Hate-Speech. In Proceedings of the first workshop on abusive language online. ACL, pp. 85–90
30. Gao Z, Zhao J, LI S-R, HU R-R (2020) The Improved Equilibrium Optimization Algorithm with Tent Map. In 2020 5th International Conference on Computer and Communication Systems (ICCCS). IEEE, pp. 343–346. https://doi.org/10.1109/ICCCS49078.2020.9118477
31. Hee C et al (2018) Automatic detection of cyberbullying in social media text'. Plos One 13(10):1–22. https://doi.org/10.1371/journal.pone.0203794
32. Huang Q, Inkpen D, Zhang J, Van Bruwaene D (2018) Cyberbullying Intervention Based on Convolutional Neural Networks. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 42–51. Available at: https://www.bbc.co.uk/news/10302550. Accessed 7 Nov 2014
33. Huang Q, Singh VK, Atrey PK. (2014) Cyber Bullying Detection Using Social and Textual Analysis. In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia - SAM '14. ACM Press, New York, USA, pp 3–6. https://doi.org/10.1145/2661126.2661133
34. Iwendi C, Srivastava G, Khan S, and Maddikunta PKR (2020) Cyberbullying detection solutions based on deep learning architectures. Multimed Syst 1–14. https://doi.org/10.1007/s00530-020-00701-5
35. Kaur S, Kumar P, Kumaraguru P (2020) Automating fake news detection system using multi-level voting model'. Soft Comput 24(12):9049–9069. https://doi.org/10.1007/s00500-019-04436-y
36. Khodabakhsh M, Kahani M, Bagheri E (2020) Predicting future personal life events on twitter via recurrent neural networks. J Intell Inf Syst 54(1):101–127. https://doi.org/10.1007/s10844-018-0519-2
37. Kumar A, Sachdeva N (2021b) Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimed Syst (0123456789). https://doi.org/10.1007/s00530-020-00747-5
38. Kumar A, Sachdeva N (2021) A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web. https://doi.org/10.1007/s11280-021-00920-4
39. Kumari K, Singh JP (2021) Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Trans Emerg Telecommun Technol 32(2):1–13. https://doi.org/10.1002/ett.3907

40. Kumari K, Singh JP, Dwivedi YK, Rana NP (2020) Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Comput 24(15):11059–11070. https://doi.org/10.1007/s00500-019-04550-x
41. Kumari K, Singh JP, Dwivedi YK, Rana NP (2021) Bilingual Cyber-aggression detection on social media using LSTM autoencoder. Soft Comput 25(14):8999–9012. https://doi.org/10.1007/s00500-021-05817-y
42. l-garadi MA, Varathan KD, Ravana SD (2016) Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Comput Hum Behav 63:433–443. https://doi.org/10.1016/j.chb.2016.05.051
43. Liang W, Feng R, Liu X, Li Y, Zhang X (2018) GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts. IEEE Access 6:43612–43621. https://doi.org/10.1109/ACCESS.2018.2863260
44. Liu Z, Qin T, Chen K-J, Li Y (2020) Collaboratively Modeling and Embedding of Latent Topics for Short Texts. IEEE Access 8:99141–99153. https://doi.org/10.1109/ACCESS.2020.2997973
45. Lu N, Wu G, Zhang Z, Zheng Y, Ren Y, Choo KR (2020) Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurr Comput Practice Exp 32(23):1–11. https://doi.org/10.1002/cpe.5627
46. Mishna F, Khoury-Kassabri M, Gadalla T, Daciuk J (2012) 'Risk factors for involvement in cyber bullying: Victims, bullies and bully-victims. Child Youth Serv Rev 34(1):63–70. https://doi.org/10.1016/j.childyouth.2011.08.032
47. Moazzeni AR, Khamehchi E (2020) Rain optimization algorithm (ROA): A new metaheuristic method for drilling optimization solutions'. J Pet Sci Eng 195:107512. https://doi.org/10.1016/j.petrol.2020.107512
48. Muneer A, Fati SM (2020) A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Int 12(11):1–21. https://doi.org/10.3390/fi12110187
49. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-ariki HDE (2022) DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. IEEE Access 10:25857–25871. https://doi.org/10.1109/ACCESS.2022.3153675
50. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-Ghuribi SM, Ghanem FA (2020) Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling. IEEE Access 10:105328–105351. https://doi.org/10.1109/ACCESS.2022.3211396
51. Murshed BAH, Al-ariki HDE, Mallappa S (2020) Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study. Comput Syst Sci Eng 35(6):495–512. https://doi.org/10.32604/csse.2020.35.495
52. Murshed BAH, Mallappa S, Abawajy J, Mallappa S, Saif MAN, Al-ariki HDE, Abdulwahab HM (2022) Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. Artif Intell Rev. https://doi.org/10.1007/s10462-022-10254-w
53. Murshed BAH, Mallappa S, Ghaleb OAM, Al-ariki HDE,. (2021) Efficient Twitter Data Cleansing Model for Data Analysis of the Pandemic Tweets. In Studies in Systems, Decision and Control. Springer, Cham, 348, pp. 93–114https://doi.org/10.1007/978-3-030-67716-9_7
54. Nand P, Perera R, Kasture A (2016) " How Bullying is this Message ?": A Psychometric Thermometer for Bullying. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, pp. 695–706. Available at: https://aclanthology.org/C16-1067
55. Paul S, Saha S (2020) CyberBERT: BERT for cyberbullying identification. Multimed Syst (0123456789). https://doi.org/10.1007/s00530-020-00710-4
56. Paul S, Saha S, Hasanuzzaman M (2020) 'Identification of cyberbullying: A deep learning based multimodal approach. Multimed Tools Appl 81(19):26989–27008. https://doi.org/10.1007/s11042-020-09631-w
57. Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation Jeffrey. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543
58. Pericherla S, Ilavarasan E (2021) Transformer network-based word embeddings approach for autonomous cyberbullying detection. Int J Intell Unmanned Syst. https://doi.org/10.1108/IJIUS-02-2021-0011
59. Pitsilis GK, Ramampiaro H, Langseth H (2018) Effective hate-speech detection in Twitter data using recurrent neural networks. Appl Intell 48(12):4730–4742. https://doi.org/10.1007/s10489-018-1242-y

60. Purnamasari NMGD, Fauzi MA, Indriati Dewi LS (2020) Cyberbullying identification in twitter using support vector machine and information gain based feature selection. Indones J Electric Eng Comput Sci 18(3):1494–1500. https://doi.org/10.11591/ijeecs.v18.i3.pp1494-1500
61. rochier R, Guille A, Velcin J (2019) Global Vectors for Node Representations. In The World Wide Web Conference on - WWW '19. New York, New York, USA: ACM Press, pp. 2587–2593. https://doi.org/10.1145/3308558.3313595
62. Rosa H, Matos D, Ribeiro R, Coheur L, Carvalho JP (2018) A 'Deeper' Look at Detecting Cyberbullying in Social Networks. In 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489211
63. Roy PK, Mali FU (2022) Cyberbullying detection using deep transfer learning. Complex Intell Syst 8(6):5449–5467. https://doi.org/10.1007/s40747-022-00772-z
64. Singh J, Singh AK (2020) NSLPCD: Topic based tweets clustering using Node significance based label propagation community detection algorithm. Ann Math Artif Intell 1–37. https://doi.org/10.1007/s10472-020-09709-z.
65. Squicciarini A, Rajtmajer S, Liu Y, and Griffin C (2015) Identification and characterization of cyberbullying dynamics in an online social network. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. New York, NY, USA: ACM, pp. 280–285. https://doi.org/10.1145/2808797.2809398
66. Srinath AS, Johnson H, Dagher GG, Long M (2021) BullyNet: Unmasking Cyberbullies on Social Networks. IEEE Trans Comput Soc Syst 8(2):332–344. https://doi.org/10.1109/TCSS.2021.3049232
67. Talpur BA, O'Sullivan D (2020) Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter. Inf 7(4):52. https://doi.org/10.3390/informatics7040052
68. Tripathy JK, Chakkaravarthy SS, Satapathy SC, Sahoo M, and Vaidehi V (2020) ALBERT-based fine-tuning model for cyberbullying analysis. Multimed Syst 28:1941–1949. https://doi.org/10.1007/s00530-020-00690-5
69. Vivolo-Kantor AM, Martell BN, Holland KM, Westby R (2014) A systematic review and content analysis of bullying and cyber-bullying measurement strategies. Aggress Violent Beh 19(4):423–434. https://doi.org/10.1016/j.avb.2014.06.008
70. Wang XD, Chen RC, Yan F, Zeng ZQ, Hong CQ (2019) Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data. IEEE Access 7:42639–42651. https://doi.org/10.1109/ACCESS.2019.2907043
71. Wang K, Xiong Q, Wu C, Gao M, Yu Y (2020) Multi-modal cyberbullying detection on social networks. Proceed Int Joint Conf Neural Netw. https://doi.org/10.1109/IJCNN48605.2020.9206663
72. Yan F, X-dong W, Z-qiang Z, C-qun H (2020) Adaptive multi-view subspace clustering for high-dimensional data'. Pattern Recogn Lett 130:299–305. https://doi.org/10.1016/j.patrec.2019.01.016
73. Yuvaraj N et al (2021) Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Comput Electric Eng 92:107186. https://doi.org/10.1016/j.compeleceng.2021.107186
74. Yuvaraj N et al (2021) Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking. Math Probl Eng 2021:1–12. https://doi.org/10.1155/2021/6644652
75. Zhang X et al (2016) Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 740–745. https://doi.org/10.1109/ICMLA.2016.0132
76. Zhang Y, Ramesh A (2018) Fine-Grained Analysis of Cyberbullying Using Weakly-Supervised Topic Models. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 504–513. https://doi.org/10.1109/DSAA.2018.00065
77. Zhang Z, Robinson D, Tepper J (2018) Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In GangemiAnna, A. et al. (eds) The European semantic web conference. ESWC 2018. Lecture Notes in Computer Science. Springer, Cham, pp. 745–760. https://doi.org/10.1007/978-3-319-93417-4_48
78. Zhao Y, Karypis G (2001) Criterion functions for document clustering: Experiments and analysis
79. Zhao R, Mao K (2016) Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. IEEE Trans Affect Comput 8(3):328–339. https://doi.org/10.1109/TAFFC.2016.2531682
80. Zhou C, Sun C, Liu Z, Lau FCM (2015) A C-LSTM Neural Network for Text Classification. arXiv preprint arXiv:1511.08630. arXiv:1511.08630

81. Zhou K, Yang S (2020) Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Anal Appl 23(1):455–466. https://doi.org/10.1007/s10044-019-00783-6
82. Zuo Y, Wu J, Zhang H, Lin H, Xu K, Xiong H (2016) Topic Modeling of Short Texts: A Pseudo-Document View. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD 2016), pp. 2105–2114. https://doi.org/10.1145/2939672.2939880

Authors and Affiliations

Belal Abdullah Hezam Murshed[1,2] · Suresha[2] · Jemal Abawajy[3] · Mufeed Ahmed Naji Saif[4] · Hudhaifa Mohammed Abdulwahab[5] · Fahd A. Ghanem[6]

1 Department of Computer Science, College of Engineering and Information Technology, University of Amran, Amran, Yemen

2 Department of Studies in Computer Science, University of Mysore, Mysore, Karnataka 570006, India

3 School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Geelong, VIC 3220, Australia

4 Department of Computer Applications, Sri Jayachamarajendra College of Engineering, (Affiliated to VTU), Mysore, Karnataka 570006, India

5 Department of Computer Application, Ramaiah Institute of Technology, (Affiliated to VTU), Bangalore 560054, India

6 Department of Computer Science and Engineering, PES College of Engineering, (Affiliated to Mysore University), Mandya 571401, India