IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 January 2024, accepted 30 March 2024, date of publication 10 April 2024, date of current version 1 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386637

RESEARCH ARTICLE

RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features

ARWA A. JAMJOOM[1], HANEN KARAMTI[2], MUHAMMAD UMER[3], SHTWAI ALSUBAI[4], TAI-HOON KIM[5], (Member, IEEE), AND IMRAN ASHRAF[6]

[1]Department of Information System, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[3]Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[5]School of Electrical and Computer Engineering, Yeosu Campus, Chonnam National University, Yeosu-si, Jeollanam-do 59626, Republic of Korea
[6]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Corresponding authors: Imran Ashraf (imranashraf@ynu.ac.kr) and Tai-Hoon Kim (taihoonn@chonnam.ac.kr)

ABSTRACT Online platforms are fostering social interaction, but unfortunately, this has given rise to antisocial behaviors such as cyberbullying, trolling, and hate speech on a global scale. The detection of hate and aggression has become a vital aspect of combating cyberbullying and cyberharassment. Cyberbullying involves using aggressive and offensive language including rude, insulting, hateful, and teasing comments to harm individuals on social media platforms. Human moderation is both slow and expensive, making it impractical in the face of rapidly growing data. Automatic detection systems are essential to curb trolling effectively. This research deals with the challenge of automatically identifying cyberbullying in tweets from a publicly available cyberbullying dataset. This research work employs robustly optimized bidirectional encoder representations from the transformers approach (RoBERTa), utilizing global vectors for word representation (GloVe) word embedding features. The proposed approach is further compared with the state-of-the-art machine, deep, and transformer-based learning approaches with the FastText word embedding approach. Statistical results demonstrate that the proposed model outperforms others, achieving a 95% accuracy for detecting cyberbullying tweets. In addition, the model obtains 95%, 97%, and 96% for precision, recall, and F1 score, respectively. Results from k-fold cross-validation further affirm the supremacy of the proposed model with a mean accuracy of 95.07%.

INDEX TERMS Cyberbullying, RoBERTa, GloVe, FastText, transformer based learning.

I. INTRODUCTION

Social media serves as a platform that allows individuals to connect and engage with their friends online, enabling the sharing of photos, videos, and daily updates. In the present day, almost everyone is active on social media, integrating virtual meeting platforms into their daily routines through global online social networks (OSNs) to facilitate communication. This network assists users in discovering new friends and expanding their connections worldwide. Moreover, the sharing of data and opinions constitutes significant features of OSNs [1], [2], [3]. In recent years, the utilization of OSNs has experienced rapid growth. Platforms like Facebook, Google+, LinkedIn, Twitter, VKontakte, Mixi, and Sina Weibo have evolved into preferred modes of communication for billions of daily active users. As per statistics, in 2020, nearly 3.6 billion users engaged with social media networking sites, and this number is projected to reach 4.41 billion by 2025. According to Backlink, 58.11% of the global population is using social media [4]. Users

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao Liu.

dedicate a substantial amount of time to updating their content, interacting with primary users, and browsing others' accounts to find specific information, making social network websites a significant part of their daily activities. OSNs have the potential to eliminate economic and geographical barriers among users, fostering information sharing and communication. Additionally, OSNs play a crucial role in achieving objectives such as entertainment, education, and job searching. The widespread use of OSNs, however, increases the risk of various attacks on users. Many OSN users expose their private data, providing opportunities for attackers to engage in specific malicious activities [5], [6].

In conjunction with the widespread use of social networks, cyberbullying is identified as a troubling phenomenon, encompassing bullying activities that occur through digital devices such as cell phones, computers, and tablets [7]. Presently, the predominant users of the internet are young people, with 95% of teenagers in the U.S. being online, and a significant majority accessing social networks through their mobile devices [8]. In the era of Web 2.0, cyberbullying is more likely to occur compared to traditional bullying. A notable 36.5% of individuals feel they have experienced cyberbullying at some point in their lives, and 17.4% have reported incidents within the last 30 days [9]. These figures have more than doubled since 2007 [10]. Furthermore, a 2019 Google survey revealed that teachers in the U.S. consider cyberbullying as their primary safety concern in classrooms [11]. Over the past few years, online communication has increasingly become user-driven, with social network platforms emerging as the most vulnerable to cyberbullying due to their massive global user base, efficient communication channels, and continuous 24/7 services [12]. The impact of cyberbullying extends beyond affecting individuals' psychological well-being, influencing various aspects of their lives. This is particularly concerning for young people, as cyberbullying has the potential to lead them towards self-harm and even suicide [13].

Cyberbullying has emerged as a significant issue over the past decade, particularly impacting children and adolescents. Notably, a study conducted in the U.S. revealed that more than 43% of teenagers in the country have experienced cyberbullying [14]. European statistics from ER indicate that approximately 18% of youngsters in Europe have encountered bullying or harassment through the Internet and mobile phones [15]. Online Report for EU kids in 2014 further highlights that every fifth student of the 11-16 years age group has faced instances of cyberbullying [16]. Even in developed countries such as Sweden, cyberbullying has increased manifolds, reaching a crucial point and steadily deteriorating [17]. These findings underscore the urgency of devising effective, prompt, and proven solutions to address this pervasive issue on the Internet. Therefore, it is crucial to approach the problem of cyberbullying from various angles, including the development of automated systems for the identification and prevention of such incidents.

Given the escalating nature of this issue, researchers have actively pursued methods for the detection and prevention of cyberbullying. A particularly promising avenue involves the application of machine learning (ML) algorithms capable of analyzing extensive data sets to discern patterns and relationships among variables [18]. Recognizing the complexity of the task at hand, it is improbable that a sole ML algorithm would prove adequate. In the context of this paper, we propose a model grounded in robustly optimized bidirectional encoder representations from transformers (RoBERTa) to specifically identify cyberbullying in the textual content. The primary contributions of this work can be succinctly summarized as follows

- To enhance the predictive accuracy of cyberbullying, a novel framework is proposed that is based on global vectors for word representation (GloVe) features with the RoBERTa model. The transformer-based RoBERTa model makes use of GloVeword2vec features to give optimal results.
- The study involves an assessment of the performance of established machine, deep, and transformer-based learning algorithms applied to cyberbullying data. These models include random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), Naive Bayes (NB), bidirectional long short-term memory (BiLSTM), convolutional long short-term memory (ConvLSTM), bidirectional encoder representations from transformers (BERT), and convolutional neural network (CNN).
- The effectiveness of the proposed approach is thoroughly examined through extensive experiments, and a comparative analysis with various state-of-the-art methods is conducted. To validate the robustness of the proposed approach, the results are further substantiated using k-fold cross-validation.

The structure of this paper is as follows: Section II dives into existing research, featuring studies that utilize ML and deep learning for cyberbullying detection. Section III outlines the materials, methodologies, and suggested models for cyberbullying identification. Section IV provides comprehensive details about the experiment's results and their corresponding discussions. Finally, Section V concludes the paper with summarizing remarks.

II. LITERATURE REVIEW

The issue of social media cyberbullying, notably Twitter (now referred to as X) and Facebook, is of significant concern due to its profound impact on users' well-being, especially among younger demographics who frequently use these platforms [19]. Ottosson [20] introduced a large language model (LLM) aimed at detecting cyberbullying on social media platforms. Utilizing the GPT-3 LLM, the study sought to minimize the gap in platform moderation. The outcomes indicate that the proposed model performs comparably to the preceding models. The research also suggests that fine-tuning an LLM is an effective strategy for enhancing cyberbullying detection, with the study achieving an accuracy of 90%. In another study by Alhloul and Alam [21], a deep learning-based system was proposed for identifying bullying tweets. The researchers developed a CNN-attention framework that amalgamated an attention layer with a convolutional pooling layer, enabling efficient extraction of cyberbullying-related keywords from users' tweets. The study utilized two sets of combinations. Initially, they combined CNN and ML models where convolutional layers served as feature extractors, and ML models like RF and LR were used for classification. In the subsequent structure, they employed combinations like CNN-XGB and CNN-LSTM for classification. The findings revealed that the proposed CNN-Attention framework outperformed other learning models, achieving an impressive accuracy of 97.10%.

Wang et al. [22] presented a graphical convolutional method for underlying cyberbullying detection. The framework leverages a semi-supervised online dynamic query expansion (DQE) process to automatically generate balanced data. This process extracts additional natural data points of a specific class from Twitter. They also introduced a graph convolutional network (GCN) classifier that operates on a graph, constructed from thresholded cosine similarities between tweet embeddings. The performance of this system was compared against different machine learning models coupled with various embedding techniques. The results show that for the proposed SOSNet, using SBERT, an accuracy score of 92.70% was achieved. In a separate study, Qudah et al. [23] suggested an improved system for cyberbullying detection utilizing an adaptive external dictionary (AED). The authors employed ML models such as RF, XGB, and CatBoost, and introduced ensemble voting models. The findings suggest that the proposed ensemble voting model, when used with AED, provided superior accuracy in detecting cyberbullying incidents.

Mathur et al. [24] introduced a real-time system for detecting cyberbullying on Twitter using natural language processing (NLP) and ML. The system underwent training on a dataset comprising cyberbullying tweets, employing various ML algorithms whose performances were subsequently compared. Through tuning, it was determined that the RF algorithm yielded the most favorable results. The study's outcomes revealed that by carefully selecting preprocessing techniques and fine-tuning the RF model, an impressive accuracy of 94.06% was achieved. In a separate effort, Bokolo and Liu [25] proposed a deep learning-based system designed for automatically detecting cyberbullying on social media platforms. The study involved a comparison of three ML models NB, SVM, and Bi-LSTM using a Twitter dataset focused on cyberbullying. The demonstrated results of the experiments suggested that the Bi-LSTM outclassed the others, attaining a remarkable accuracy of 98%. Following closely, SVM attained a 97% accuracy, while NB lagged with 85%. These findings underscore the effectiveness of ML techniques in unveiling cyberbullying, Bi-LSTM with the deep learning model done and dusted two other ML models.

Nisha and Jebathangam [26] introduced an automated system for classifying and detecting cyberbullying on social media. Metadata and textual content are collectively used to identify social media cyberbullying in this model. This method involves training and testing phases on data related to cyberbullying. NLP functions as the preprocessing tool during these phases, followed by the application of particle swarm optimization for feature selection. The study employs a decision tree (DT) classifier to categorize cyberbullying. DT model performance is evaluated using text-instances-combined features post-classification. The study's results indicate that the proposed DT classifier achieved an impressive accuracy score of 99.1%. In a related endeavor, Mehmud et al [27] proposed an ML-based system focusing on textual features from the data for cyberbullying detection. The study involved an analysis of five distinct ML models LGBM, XGB, LR, RF, and ADA to detect cyberbullying using a dataset of tweets with textual features. The dataset, comprising more than 47,000 tweets divided into six classes, underwent thorough evaluation. The results revealed that LGBM outperformed the other models, demonstrating accuracy as high as 85.5%.

Muneer et al. [28] proposed a modified BERT and stacking ensemble model to identify social media cyberbullying. A continuous bag of words (CBOW), along with a word2vec-like feature extraction method is employed to establish weights in the embedding layer. An outstanding accuracy of 97.4% was achieved using the stacking ensemble learning method as revealed by experiment results for discovering cyberbullying on the tweet dataset. A related study used CBOW feature extraction and attention mechanism based on deep learning focussed on detecting cyberbullying. Various deep learning models, including Conv1DLSTM, LSTM, CNN, BiLSTM_Pooling, BiLSTM, and gated recurrent unit (GRU), were employed. Results underlined the dominance of the attention-based Conv1DLSTM classifier over the other applied approaches, achieving the highest accuracy of 94.49%.

The existing approaches are marked by several limitations. First, feature engineering approaches are not very well studied in the context of machine learning models for cyberbullying detection. Secondly, the imbalanced class distribution is not investigated with respect to its impact on classification accuracy. Finally, BERT and its variants are not widely studied for the topic at hand. This study aims to fill this gap by proposing the RoBERTa model and investigating several feature engineering approaches. The complete summary of state-of-the-art related works is shown in Table 1.

III. MATERIAL AND METHODS

This segment outlines the crucial aspects of the envisioned framework for detecting cyberbullying, known as the Cyberbullying detection system (CDS). It delves into the

TABLE 1. Summary of state-of-the-art related work.

Ref	Classifiers	Dataset	Performance
[20]	BERT, GPT-3	Kaggle	90% GPT-3 Ada
[21]	LR, RF, XGB, CNN, CNN+LR, CNN+RF, CNN+XGB, CNN+LSTM, CNN-Attention	Kaggle	97.10% CNN-Attention
[22]	LR, NB, k-NN, SVM, XGB, MLP, SOSNet	Kaggle	92.70% SOSNet with SBERT
[23]	RF, XGB, CatBoost, Ensemble Voting	Kaggle	98.55% Ensemble voting
[24]	Vanilla RF, AdaBoost, GBC, Tuned RF	Kaggle	94.06% Tuned RF
[25]	NB, SVM, Bi-LSTM	Kaggle	98% Bi-LSTM
[26]	SVM, ANN, DT	Kaggle	99.1% DT
[27]	LGBM, XGB, LR, RF, ADA	Kaggle	85.5% LGBM
[28]	Conv1DLSTM, BiLSTM, LSTM, BERT, Tuned-BERT, Stacked and CNN	Kaggle (37373)	97.4% stacked
[29]	LSTM, Conv1DLSTM, CNN, BiLSTM, BiLSTM_Pooling, GRU	Kaggle (37373)	94.49% Conv1DLSTM

examination and identification of cyberbullying activities across various social media platforms. Additionally, this section details the dataset, machine learning, and deep learning models employed in the study. A concise overview of the feature extraction techniques utilized is also incorporated within this section.

A. DATASET

As the utilization of social media becomes widespread across diverse age groups, a significant portion of the population relies on this essential tool for daily communication. The pervasive nature of social media has made cyberbullying a pervasive issue, capable of affecting individuals at any time and place. The Internet's inherent anonymity adds to the complexity of combating such attacks compared to traditional forms of bullying. The recent pandemic (COVID-19), marked by extensive school closures, declined in-person social contact, and explosive screen time, led UNICEF to warn the users of the potential exposure to cyberbullying. Statistics reveal that cyberbullying was experienced by 36.5% of middle and high school students. This has affected the academic performance of the students drastically, feelings of sadness, and even suicidal thoughts.

The dataset under consideration has been sourced from the Kaggle website and encompasses over 47,000 tweets classified as cyberbullying [30]. The categories within this dataset include

- Gender
- Age
- Religion
- Ethnicity
- Other forms of cyberbullying
- No online bullying

It is noteworthy to point out that the dataset is balanced, with 8,000 instances in each class. This balance ensures equal representation. As the dataset is explored, it is found that the tweets are either offensive in their entirety or actually describe incidents of bullying.

B. FEATURE EXTRACTION METHODS

Feature extraction methods were employed on the training subset, encompassing both the training and testing data. These techniques were applied to train the selected models using the training data and were subsequently utilized during the classification of the testing data. In the context of cyberbullying detection in this study, two feature extraction methods, namely Word2Vec and FastText, were utilized. A brief overview of these techniques is provided below.

1) WORD2VEC

To improve the current form of word implanting, Word2Vec, a robust statistical model, was developed. Word representations were elevated to the subsequent level through Word2Vec, allowing the model to efficiently capture understated distinctions [31]. These representations significantly influence semantics, shaping the understanding of language by depicting the relationships between words. Word2Vec encompasses two distinct models: the continuous bag-of-words (CBoW) and the continuous skip-gram model. The CBoW model generates embeddings by predicting the current word based on its context. This process is mathematically expressed by the following

$$J_\theta = \frac{1}{T}\sum logp(w_t|w_{t-n}, \cdots, w_{t+n}) \tag{1}$$

The continuous skip-gram model is derived by predicting the neighbouring words assuming the current word. This model is known for its efficiency in training compared to the word embedding model. It is expressed mathematically by the following

$$J_\theta = \frac{1}{T}\sum\sum logp(w_{j+1}|w_t) \tag{2}$$

2) GLOBAL VECTORS FOR WORD REPRESENTATION

Numerous statistical models have been devised to address the concept of topic modeling. One notable algorithm in this domain is Latent semantic analysis, which utilizes matrix factorization to extract meaningful semantics [32]. While effective, it has certain limitations when compared to Word2Vec. Subsequently, to create a more efficient model that integrates the strengths of mutual approaches, GloVe was introduced. In many cases, GloVe tends to outperform Word2Vec. It operates on the entire corpus by employing a word context matrix and a word co-occurrence matrix.

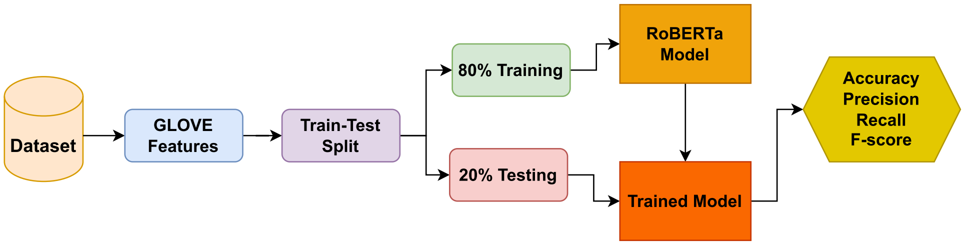

FIGURE 1. Architecture of proposed RoBERTaNet model.

3) FASTTEXT

FastText is a word illustration library, that encompasses 300-dimensional 2 million common crawl words, resulting in 600 billion word vectors [32]. Particularly, hand-crafted n-grams are used as features alongside solo words. Its straightforward architecture contributes to the effective and efficient performance of text classification tasks. Various word implanting methods are used in diverse text classification endeavors, with pre-trained unsupervised word embedding to predict word context. FastText excels in this context by leveraging morphological features to identify challenging words, enhancing its suitability for vector representation. This characteristic also bolsters its generalizability. Furthermore, FastText word embedding generates vectors using n-grams, a feature that proves beneficial in handling unknown words.

C. PROPOSED METHODOLOGY

This section introduces the framework suggested for this study, as depicted in Figure 1. Recently, a growing trend in using deep learning-based classifiers has emerged. The RoBERTa model, in particular, has garnered attention for its potential to enhance traditional classifiers for their classification accuracy. Therefore, this study seeks to leverage RoBERT to detect cyberbullying in text data.

Experiments are conducted using benchmark datasets related to cyberbullying. Various word embedding techniques, including Word2Vec and FastText, are employed. Subsequently, the suggested method, which combines Word2Vec word embedding with RoBERTa, is applied for training. The effectiveness of the projected methodology is assessed using four evaluation measures including precision, recall, accuracy, and F1 score.

D. SUPERVISED MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

This section focuses on the ML algorithms employed for cyberbullying detection using the Twitter dataset. The execution of these ML models involves the utilization of the Sci-Kit learn library and NLTK. Specifically, four ML algorithms and three deep learning algorithms were applied in Python. To assess the effectiveness of the proposed system, a combination of regression-based, tree-based, and ensemble-based models was employed. The following ML algorithms, in combination with the proposed methodology, evaluated the performance of the ML classifiers.

1) RANDOM FOREST

RF stands as a sophisticated decision tree (DT) iteration. It operates as a supervised learning algorithm, employing numerous DTs that operate independently to predict a class label. The ultimate prediction is determined by the class that garners the majority of votes across the individual trees [33]. When assessing the error rate in RF and comparing it to other models, it emerges as notably low. This low error rate is attributed to the diminished correlation between the constituent trees. In this study, the RF was trained using various parameters, and for decision tree splits, multiple algorithms were employed depending on the specific problem at hand.

2) K-NEAREST NEIGHBOR

KNN is categorized as a lazy learner as it requires all the data during the training process. In KNN, a new sample is attributed to a class using a distance metric [34]. KNN measures the distance of new data points to their closest neighbors, and the value of K dictates the number of neighbors. For instance, if K is set to 3, new data points are assigned to a class by considering its distance to 3 nearest neighbors. The calculation of the distance between two points can be carried out using a variety of distance estimation metrics and the following are the most commonly used ones.

$$Euclideandistance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{3}$$

$$Manhattandistance = \sum_{i=1}^{k}|x_i - y_i| \tag{4}$$

$$Minkowskidistance = (\sum_{i=1}^{k}|x_i - y_i|^q)^{\frac{1}{q}} \tag{5}$$

3) NAIVE BAYES

GNB, a Naive Bayes variant grounded in Bayes' theorem, operates on conditional probabilities to forecast the outcome of an event [35]. In this context, if a sample is categorized into k classes represented by $k = c1, c2, \cdots, ck$, the output results are denoted as c. The GNB function is outlined below, where c represents the class and d signifies the sample

$$P(c|d) = \frac{P(d|c) \cdot P(c)}{P(d)} \tag{6}$$

This formula encapsulates the class c probability given the sample d, computed as the product of the likelihood of the sample given the class ($P(d|c)$) and the prior probability of the class

$$(P(c)),$$

normalized by the probability of the sample ($P(d)$).

4) SUPPORT VECTOR MACHINE

The primary purpose of the SVM classifier is to adapt to the provided data and determine the optimal hyperplane that effectively separates the data [36]. Once this hyperplane is obtained, certain features are input into the classifier to derive the final class results. This characteristic makes SVM a distinctive algorithm. SVM employs a linear kernel function and is implemented in liblinear as opposed to libSVM. Notably flexible in terms of penalty and loss function choices, SVM yields favorable scaling results, particularly for large sample sizes. As a widely utilized supervised learning algorithm, SVM excels in handling both classification and regression problems. Within the SVM module, two functions are available: linear SVM and SVM. The linear function of SVM is employed when developing an SVM model with a linear kernel.

5) BIDIRECTIONAL LONG SHORT-TERM MEMORY

BiLSTM is a subtype of recurrent neural network (RNN) specifically designed for handling sequence processing tasks, commonly applied in natural language processing and speech recognition [37]. In contrast to traditional LSTMs, BiLSTM processes input data in both the forward and backward directions, allowing it to capture contextual information from both preceding and subsequent states. This bidirectional strategy enhances the network's ability to comprehend and learn long-range dependencies within sequential data. Comprising two LSTM layers, one processing the input sequence from start to finish and the other in reverse, BiLSTM concatenates the outputs from both directions. This amalgamation yields a more comprehensive representation of the input sequence, making BiLSTM particularly adept at tasks requiring a nuanced understanding of context and dependencies in both directions. As a result, BiLSTM proves to be a potent tool for various applications in the analysis of sequential data.

6) CONVOLUTIONAL LONG SHORT TERM MEMORY

ConvLSTM represents a neural network architecture that merges the spatial hierarchies learned by convolutional layers with the capacity to capture temporal dependencies provided by LSTM networks [38]. This architecture proves particularly effective in handling spatiotemporal data, such as video sequences or time-series data characterized by spatial dependencies. ConvLSTM seamlessly incorporates convolutional operations into the LSTM cells, enabling the model to simultaneously learn both spatial patterns and temporal dynamics. This integration makes ConvLSTM well-suited for tasks demanding an understanding of intricate patterns in both spatial and temporal dimensions, including applications like video prediction, anomaly detection, and weather forecasting. ConvLSTM has demonstrated success in capturing complex relationships within sequences, establishing itself as a valuable tool in various domains that necessitate the analysis of dynamic, structured data.

7) CONVOLUTIONAL NEURAL NETWORK

CNN stands as a prevalent artificial neural network employed across a diverse range of tasks [39]. Conceptually, a CNN bears a resemblance to a multi-layer perceptron (MLP) in that it encompasses an activation function for each neuron, mapping the weighted outputs. Notably, an MLP transitions into a deep MLP when multiple hidden layers are introduced. The architecture of CNN imparts it with the ability to be invariant to translation and rotation. Three fundamental layers constitute a CNN: a core layer, a pooling layer, and a fully connected layer, each incorporating an activation function.

E. TRANSFORMER BASED ARCHITECTURE

1) BERT

BERT is a transformer-based model featuring an attention mechanism that actively processes input text [40]. Comprising two components, an encoder, and a decoder, BERT takes textual input and generates predictions as output. Particularly in NLP-suited question-and-answering and sentiment analysis tasks, BERT's strength lies in its training on extensive word-based data. Unlike traditional models that consider word context in only one direction, typically left to right, BERT comprehensively accounts for word context in both directions. This unique feature distinguishes BERT from earlier deep learning models, endowing it with a nuanced understanding of word meanings. Trained on a large corpus of data, BERT excels at producing accurate results and acquiring a profound grasp of intricate patterns and structures.

2) ROBUSTLY OPTIMIZED BERT APPROACH

RoBERTa stands as a cutting-edge natural language processing (NLP) model introduced by Facebook AI in 2019. Rooted in the BERT architecture, RoBERTa incorporates multiple optimizations aimed at enhancing performance [40]. Pre-trained on an extensive corpus of text data, RoBERTa excels in grasping contextual relationships within language. Key optimizations include the utilization of larger mini-batches, dynamic masking during pre-training, and the elimination of the next sentence prediction objective. By undergoing training on diverse and comprehensive datasets, RoBERTa demonstrates superior performance across a spectrum of NLP tasks, including question answering, sentiment analysis, and named entity recognition. Its robustness and versatility have propelled its widespread adoption, serving as a foundational model for fine-tuning and transfer learning in various natural language understanding applications.

F. PERFORMANCE EVALUATION METRICS

The evaluation of deep, machine, and transformer-based models also involves assessing their performance using various metrics such as accuracy, precision, recall, and the F1 score. Accuracy is determined by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP represents true positive, TN denotes true negative, FP signifies false positive, and FN represents false negative.

Precision is an additional performance metric that gauges the actual positives to the total number of positive predictions ratio, expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

The recall metric is also employed to evaluate model performance and is calculated by dividing true positives by the sum of true positives and false negatives, given by

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

In situations where class imbalance is present, the F1 score emerges as a superior metric, as it considers both precision and recall, providing a more comprehensive assessment of the model's performance. The F1 score is computed using the formula

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

IV. RESULTS AND DISCUSSION

This segment provides comprehensive insights into the experiments conducted using the Twitter dataset for bullying employing DL, ML, and transformer-based models. The assessment of research experiments is carried out using a Python-based Colab Notebook, leveraging Tensorflow, Keras, and Sklearn libraries. To gauge the performance of diverse models, multiple metrics, comprising precision, recall, accuracy, and the F1 score, are employed. For transformer-based and deep model training, a graphics processing unit (GPU) and 16 GB of RAM are used to expedite this process. Subsequent sections present the outcomes of the conducted experiments.

A. MODEL RESULTS USING WORD2VEC FEATURES

In this study, a series of experiments are undertaken to detect cyberbullying tweets, employing a combination of ML and DL models. The identification of bullying instances in the data was facilitated by incorporating Word2vec features into the feature set. The results of employing the Word2vec technique for multi-class classification are presented in the accompanying Table 2.

TABLE 2. Models results using Word2Vec features.

Classifier	Accuracy	Precision	Recall	F1 score
RF	90	90	88	89
KNN	70	75	69	72
NB	83	81	82	81
SVM	80	82	80	81
BiLSTM	82	81	82	81
Conv LSTM	82	82	83	83
BERT	89	89	90	90
RoBERTa	91	91	92	91
CNN	77	78	75	76

TABLE 3. Models results using GloVe features.

Classifier	Accuracy	Precision	Recall	F1 score
RF	94	94	94	94
KNN	67	79	67	69
NB	85	85	85	84
SVM	82	86	84	85
BiLSTM	86	84	86	85
Conv LSTM	85	85	85	84
BERT	93	93	93	93
RoBERTa	95	95	97	96
CNN	79	81	73	77

Table 2 presents the outcomes of the utilization of Word2vec features. Notably, RoBERTa achieves the highest accuracy of 91%, and similarly, it secures top values for precision, recall, and F1 scores. In the realm of machine learning-based classification, the tree-based model RF attains an accuracy of 90%, accompanied by corresponding values of 90% for precision, 88% for recall, and 89% for the F1 score. Among the DL models, the CNN for multi-class grouping records the lowest accuracy, standing at 77%. Conversely, the ML model KNN emerges as the least performing classifier, achieving an accuracy of 70%.

B. RESULTS USING GLOVE FEATURES

Several experiments were conducted to identify cyberbullying tweets, employing both ML and DL models in this study. The detection of bullying in the data was facilitated by utilizing GloVe features as part of the feature set. The outcomes of the GLOVE technique for multi-class classification are showcased in Table 3.

Table 3 displays the results obtained through GloVe features. Notably, the highest accuracy of 95% is attained by RoBERTa. Similarly, RoBERTa also secures the top values for precision, recall, and F1 scores. In the context of a tree-based machine learning model, RF achieves an accuracy score of 94%, along with corresponding values of 94% for precision, recall, and F1 score. Among the deep learning models, CNN for multi-class grouping records the lowest accuracy, standing at 79%. Conversely, the machine learning model KNN emerges as the least performing classifier, achieving an accuracy of 67%.

C. RESULTS USING FASTTEXT FEATURES

In the subsequent set of experiments, trials were conducted utilizing FastText features. The outcomes for both ML and

TABLE 4. Models results using FastText functions.

Classifier	Accuracy	Precision	Recall	F1 score
RF	88	84	87	85
KNN	74	74	70	72
NB	76	74	71	72
SVM	78	80	81	80
BiLSTM	75	73	77	74
Conv LSTM	84	87	88	87
BERT	80	86	84	85
RoBERTa	92	90	91	90
CNN	74	75	75	75

TABLE 5. K-fold cross-validation results.

K-fold	Accuracy	Precision	Recall	F1 score
Fold 1	94.23	94	96	95
Fold 2	94.89	94	96	96
Fold 3	95.99	95	96	96
Fold 4	95.27	96	97	97
Fold 5	94.98	95	97	96
Average	**95.07**	**96**	**95**	**96**

DL models, incorporating FastText features, are showcased in Table 4. The results suggest a decline in the performance of the learning models. Additionally, the findings underscore that models employing Word2vec features outperform those utilizing FastText features in terms of performance.

Table 4 displays the outcomes obtained using FastText features. RoBERTa stands out with the highest accuracy of 92.25%, along with superior values for precision, recall, and F1 scores. In the context of multi-class classification, the tree-based model RF achieves an accuracy of 88.97%, with corresponding values of 84% for precision, 87% for recall, and 85% for F1 score. Among the DL models, CNN for multi-class classification records the lowest accuracy on FastText features, standing at 74%. Conversely, the ML model KNN again emerges as the least performing classifier, achieving an accuracy of 74%. It is worth noting an improvement in the performance of KNN on FastText features. In summary, the overall performance of the learning models demonstrates a decline using FastText features when compared to Word2vec features.

D. K-FOLD CROSS-VALIDATION RESULTS

K-fold cross-validation was utilized to enrich the investigation of the models. Five-fold cross-validation results in Table 5 showcase the superior performance in terms of F1 score, accuracy, recall, and precision of the model to other alternatives.

E. PERFORMANCE COMPARISON WITH EXISTING STUDIES

The results of this method were evaluated against multiple approaches for robustness and performance from the available literature. In pursuit of meaningful outcomes, miscellaneous DL and ML models were employed here. For instance, in [25], bi-directional LSTM models were utilized for predicting cyberbullying on social media. Additionally, [22] incorporated a feature selection approach to enhance model performance. Other studies, such as [24], [26], and [27], employed ML models for a similar purpose. Ensemble models were also explored in [21] and [23], making a significant contribution to the efficient detection of cyberbullying. The comparative outcomes given in Table 6 demonstrate that the proposed approach yields superior results when compared to these existing approaches.

TABLE 6. Performance comparison of the proposed approach with state-of-the-art techniques.

Ref	Proposed approach	Performance
[20]	GPT-3 Ada	90.2%
[22]	SOSNet with SBERT	92.7%
[24]	Tuned RF	94.6%
[27]	LGBM	85.5%
Proposed	**RoBERTa with Word2vec**	**95.9%**

F. DISCUSSION

This study proposes a novel RoBERTa model for cyberbullying detection. The model incorporates Word2vec features for improved performance. Extensive experiments have been performed to check the efficiency of the proposed model in connection to existing approaches as well as, several machine learning and deep learning models. In addition, the study performs experiments with Word2vec, GloVe, and FastText features. Experimental results indicate superior performance of the proposed model when used with the Word2vec features. A superior accuracy of 95.9% is obtained using the proposed RoBERTa model which is better than existing state-of-the-art approaches for cyberbullying detection. Moreover, results from k-fold cross-validation also confirm the robustness and generalizability of the proposed approach.

Nonetheless, the proposed model has a few limitations. First, it has high computational complexity, which we intend to work on in the future. Secondly, the dataset used in the study is taken from Twitter. Taking multi-domain data containing text from multiple social media platforms is needed for extensive evaluation. Thirdly, the impact of class imbalance is not investigated in this study and might be a good avenue for future research. Lastly, using other feature engineering approaches like term frequency-inverse document frequency and bag of words, can provide further insights into the performance of the proposed approach.

V. CONCLUSION

This study delves into the domain of cyberbullying tweet detection, presenting a novel approach that integrates the powerful RoBERTa model with GloVe features. The aim was to enhance the accuracy and efficiency of cyberbullying detection in the ever-evolving landscape of online interactions. The proposed model underwent rigorous evaluation, pitting its performance against a spectrum of existing models encompassing traditional machine learning, deep learning, and transformer-based architectures, as well as other state-of-the-art models. The results obtained from extensive experimentation and evaluation highlight the prowess of the RoBERTa model in tandem with GloVe features. The proposed model exhibits a significant improvement in cyberbullying detection accuracy compared to established benchmarks. Its ability to leverage contextual embeddings and nuanced linguistic features provided by RoBERTa, coupled with the complementary information encoded in GloVe features, contributes to a robust and effective detection mechanism. The model obtains a 95.9% accuracy using the word2vec features which is better than other used models, as well as, existing approaches. Furthermore, cross-validation results provide a mean accuracy of 95.07% indicating the robustness of the proposed RoBERTa model. This study not only advances the field of cyberbullying detection but also underscores the importance of leveraging state-of-the-art models in addressing contemporary challenges in online social spaces. The comparative analysis emphasizes the superiority of the proposed RoBERTa and GloVe-based approach over conventional methods, signifying its potential as a reliable tool for identifying and combating cyberbullying in the digital sphere. As the online landscape continues to evolve, this research lays a foundation for further exploration and refinement of models that can adapt to the dynamic nature of cyberbullying instances across various platforms and contexts. For future endeavors, there is an envisioned development of a customized deep learning model optimized for small datasets. Furthermore, there is contemplation of combining multiple datasets to create a complex and high-dimensional dataset for conducting experiments with the proposed approach.

ABBREVIATIONS

Table 7 shows the acronyms used in this study.

TABLE 7. Acronyms used in the study.

Acronym	Full form
RoBERTa	robustly optimized bidirectional encoder representations from the transformers
GloVe	global vectors for word representation
OSNs	online social networks
ML	machine learning
RF	random forest
SVM	support vector machine
KNN	K-nearest neighbor
NB	Naive Bayes
BiLSTM	bidirectional long short-term memory
ConvLSTM	convolutional long short-term memory
BERT	bidirectional encoder representations from transformers
CNN	convolutional neural network
LLM	large language model
DQE	dynamic query expansion
GCN	graph convolutional network
AED	adaptive external dictionary
NLP	Natural language processing
DT	decision tree
CBOW	continuous bag of words
CDS	Cyberbullying detection system
GNB	Gaussian Naive Bayes
MLP	multi-layer perceptron
RNN	recurrent neural network
GPU	graphics processing unit

REFERENCES

[1] S. R. Sahoo and B. B. Gupta, "Classification of various attacks and their defence mechanism in online social networks: A survey," *Enterprise Inf. Syst.*, vol. 13, no. 6, pp. 832–864, Jul. 2019.

[2] S. Neelakandan, R. Annamalai, S. J. Rayen, and J. Arunajsmine, "Social media networks owing to disruptions for effective learning," *Proc. Comput. Sci.*, vol. 172, pp. 145–151, Sep. 2020.

[3] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection detecting spammers and fake profiles in social networks based on topology anomalies," *Human J.*, vol. 1, no. 1, pp. 26–39, 2012.

[4] B. Dean. (2021). *How Many People Use Social Media in 2021*. [Online]. Available: https://backlinko.com/social-media-users

[5] S. Neelakandan and D. Paulraj, "A gradient boosted decision tree-based sentiment classification of Twitter data," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 18, no. 4, Jul. 2020, Art. no. 2050027.

[6] H. Kefi and C. Perez, "Dark side of online social networks: Technical, managerial, and behavioral perspectives," *Encyclopedia Social Netw. Anal. Mining*, vol. 143, pp. 535–556, Aug. 2018.

[7] Stopbullying.gov. (2020). *What is Cyberbullying*. [Online]. Available: https://www.stopbullying.gov/cyberbullying/what-is-it

[8] S. Hinduja and J. W. Patchin, *Cyberbullying: Identification, Prevention, and Response*, 2018. [Online]. Available: https://cyberbullying.org/cyberbullying-fact-sheet-identification-prevention-and-response

[9] C. R. Center. (2019). *Cyberbullying Data*. [Online]. Available: https://cyberbullying.org/2019-cyberbullying-data

[10] (2007). *Summary of Our Cyberbullying Research (2007–2019)*. [Online]. Available: https://cyberbullying.org/summary-of-our-cyberbullying-research

[11] Google. (2019). *Be Internet Awesome: Online Safety & Parents*. [Online]. Available: https://beinternetawesome.withgoogle.com/en_us

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

[13] A. John, A. C. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton, "Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review," *J. Med. Internet Res.*, vol. 20, no. 4, p. e129, Apr. 2018.

[14] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, pp. 656–666, 2012.

[15] H. Keinänen and O. Kuivalainen, "Antecedents of social media B2B use in industrial marketing context: Customers' view," *J. Bus. Ind. Marketing*, vol. 30, no. 6, pp. 711–722, Jul. 2015.

[16] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Eur. Conf. Inf. Retr.*, 2018, pp. 141–153.

[17] N. Selwyn, "Social media in higher education," *The Europa world Learn.*, vol. 1, no. 3, pp. 1–10, 2012.

[18] J. F. Hair and M. Sarstedt, "Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing," *J. Marketing Theory Pract.*, vol. 29, no. 1, pp. 65–77, Jan. 2021.

[19] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in *Proc. ACM Web Sci. Conf.*, Jun. 2017, pp. 13–22.

[20] D. Ottosson, *Cyberbullying Detection on Social Platforms Using Large Language Models*, 2023. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1786271&dswid=5491

[21] A. Alhloul and A. Alam, *Bullying Tweets Detection Using CNN-attention*, document SSRN 4338998, 2023.

[22] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2020, pp. 1699–1708.

[23] H. Qudah, M. A. Alhija, and H. Tarawneh, *Improving Cyberbullying Detection Through Adaptive External Dictionary in Machine Learning*, 2023. [Online]. Available: https://www.researchsquare.com/article/rs-3306599/v1

[24] S. A. Mathur, S. Isarka, B. Dharmasivam, and J. C. D., "Analysis of tweets for cyberbullying detection," in *Proc. 3rd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2023, pp. 269–274.

[25] B. George Bokolo and Q. Liu, "Cyberbullying detection on social media using machine learning," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2023, pp. 1–6.

[26] M. Nisha and J. Jebathangam, "Detection and classification of cyberbullying in social media using text mining," in *Proc. 6th Int. Conf. Electron., Commun. Aerosp. Technol.*, 2022, pp. 856–861.

[27] M. I. Mahmud, M. Mamun, and A. Abdelgawad, "A deep analysis of textual features based cyberbullying detection using machine learning," in *Proc. IEEE Global Conf. Artif. Intell. Internet Things (GCAIoT)*, Dec. 2022, pp. 166–170.

[28] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT," *Information*, vol. 14, no. 8, p. 467, Aug. 2023.

[29] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, "Cyberbullying detection on Twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction," *Mathematics*, vol. 11, no. 16, p. 3567, Aug. 2023.

[30] AndrewMVD. (2022). *Cyberbullying Classification Dataset*. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

[31] M. Umer, S. Sadiq, H. Karamti, A. Abdulmajid Eshmawi, M. Nappi, M. Usman Sana, and I. Ashraf, "ETCNN: Extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification," *Pattern Recognit. Lett.*, vol. 164, pp. 224–231, Dec. 2022.

[32] M. Karim, M. M. S. Missen, M. Umer, S. Sadiq, A. Mohamed, and I. Ashraf, "Citation context analysis using combined feature embedding and deep convolutional neural network model," *Appl. Sci.*, vol. 12, no. 6, p. 3203, Mar. 2022.

[33] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018.

[34] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Water quality prediction using KNN imputer and multilayer perceptron," *Water*, vol. 14, no. 17, p. 2592, Aug. 2022.

[35] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, 2001, pp. 41–46.

[36] S. Sarwat, N. Ullah, S. Sadiq, R. Saleem, M. Umer, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Predicting students' academic performance with conditional generative adversarial network and deep SVM," *Sensors*, vol. 22, no. 13, p. 4834, Jun. 2022.

[37] M. Ahmad, S. Sadiq, A. A. Eshmawi, A. S. Alluhaidan, M. Umer, S. Ullah, and M. Nappi, "Industry 4.0 technologies and their applications in fighting COVID-19 pandemic using deep learning techniques," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105418.

[38] L. Cascone, S. Sadiq, S. Ullah, S. Mirjalili, H. U. R. Siddiqui, and M. Umer, "Predicting household electric power consumption using multi-step time series with convolutional LSTM," *Big Data Res.*, vol. 31, Feb. 2023, Art. no. 100360.

[39] U. Hafeez, M. Umer, A. Hameed, H. Mustafa, A. Sohaib, M. Nappi, and H. A. Madni, "A CNN based coronavirus disease prediction system for chest X-rays," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 10, pp. 13179–13193, Oct. 2023.

[40] X. Chen, T. Aljrees, M. Umer, H. Karamti, S. Tahir, N. Abuzinadah, K. Alnowaiser, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "A novel approach for explicit song lyrics detection using machine and deep ensemble learning models," *PeerJ Comput. Sci.*, vol. 9, p. e1469, Aug. 2023.

ARWA A. JAMJOOM received the master's degree in computer science from the University of Southern California, Los Angeles, CA, USA, in 1997, and the Ph.D. degree in computer science from the University of Surrey, Guildford, U.K., in 2011. She is currently an Associate Professor with the Department of Information System, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include data analytics and decision support systems.

HANEN KARAMTI is associated with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University. Her research interest includes text mining and computer vision tasks.

MUHAMMAD UMER received the B.S. degree from the Department of Computer Science, Khwaja Fareed University of Engineering and IT (KFUEIT), Pakistan, in October 2018. He is currently pursuing the Ph.D. degree in computer science with KFUEIT. He is a Research Assistant with the Fareed Computing and Research Center, KFUEIT. He is currently serving as the Head of the Computer Science Department, The Islamia University of Bahawalpur. His recent research interests include data mining, mainly working machine learning and deep learning-based IoT, text mining, and computer vision tasks.

SHTWAI ALSUBAI received the bachelor's degree in information system from King Saud University, Saudi Arabia, in 2008, the master's degree in computer science from CLU, USA, in 2011, and the Ph.D. degree from The University of Sheffield, U.K., in 2018. He is currently an Assistant Professor of computer science with Prince Sattam bin Abdulaziz University. His research interests include XML, XML query processing, XML query optimization, machine learning, and natural language processing.

TAI-HOON KIM (Member, IEEE) received the M.S. and Ph.D. degrees in electrics, electronics, and computer engineering from Sungkyunkwan University, Seoul, South Korea, and the second Ph.D. degree in information science from the University of Tasmania, Hobart, Australia, in December 2011. He is currently a Professor with Chonnam National University, Gwangju, South Korea. His research interests include statistical analysis, image processing, and system design.

IMRAN ASHRAF received the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2018, and the M.S. degree (Hons.) in computer science from Blekinge Institute of Technology, Karlskrona, Sweden, in 2010. He was a Post-doctoral Fellow with Yeungnam University. He is currently an Assistant Professor with the Information and Communication Engineering Department, Yeungnam University. His research interests include positioning using next generation networks, communication in 5G and beyond, location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data analytics.