Hybrid Deep Learning for Cyberbullying Detection with Recurrent Neural Networks (DEA-RNN) on Twitter

[1] Deverakonda Mallikarjuna, [2] Deverakonda Ashok, [3] K.Vara Prasad, [4] V.Lavanya,
CSE Department,
[1,2,3,4] Assistant Professor, Dhruva Engineering Collage, Hyderabad.
Shree Engineering Collage, Hyderabad.

ABSTRACT

Cyberbullying (CB) is on the rise in today's online communities. With so many people of all ages using social media, it's crucial that these sites be protected from harassment. In order to identify CB on the Twitter platform, this article introduces a mixed deep learning model dubbed DEA-RNN. To fine-tune the Elman RNN's characteristics and shorten training time, the suggested DEA-RNN model blends Elman type RNNs with an improved Dolphin Echolocation Algorithm (DEA). Using a dataset of 10,000 tweets, we conducted extensive testing on DEA-RNN and compared its results to those of other state-of-the-art algorithms like RNNs, SVMs, Multinomial Naive Bayes, and Random Forests. (RF). The testing findings indicate that DEA-RNN performs better than the alternatives in every situation tested. In terms of identifying CB on Twitter, it did better than the other methods that were taken into account. With an average of 90.45% accuracy, 89.52% precision, 88.98% memory, 89.25% F1-score, and 90.94% sensitivity, DEA-RNN performed best in case 3.

Keywords

Social media, twitter categorization, harassment detection, Dolphin Echolocation algorithm, Elman recurrent neural networks, and brief text subject modelling.

INTRODUCTION

People of all ages prefer using social media sites like Facebook, Twitter, Flickr, and Instagram to communicate and connect with one another online. Despite the positive effects of these channels for communication and social interaction, they have also given rise to negative phenomena like cyberbullying. The psychic harm caused by cyberbullying has far-reaching consequences for our culture. Teenagers who spend most of their time online are disproportionately affected by the rise in cyberbullying incidents. The obscurity of the Internet and the widespread use of social media platforms like Twitter and Facebook make them prime targets for CB.

gives those who misuse it. In India, Facebook and Twitter account for 14% of all abuse, with 37% of victims being under the age of 18 [1]. Furthermore, trolling may result in severe mental problems and negative impacts on mental health. Cyberbullying events are a leading cause of worry, melancholy, tension, and social and mental problems [2]-[4]. This fact highlights the importance of developing methods to spot instances of harassment in social media posts. (e.g., posts, tweets, and comments). The issue of identifying instances of harassment on Twitter is the primary topic of this piece. Due to the growing prevalence of cyberbullying on Twitter, the main duties in combating cyberbullying risks [5] are the identification of cyberbullying events from messages and the supply of preventative measures. Therefore, there is a pressing need to deepen our understanding of CB through social networks new ideas and guidance for addressing the cyberbullying crisis with successful methods [6]. Cyberbullying on Twitter is difficult, if not impossible, to manually watch and regulate [7]. It's also not easy to mine social media communications for signs of harassment. It's difficult to infer someone's motivations or context from their social media posts because of the brevity, abundance of vernacular, and potential presence of emoticons and videos in their messages. Furthermore, if the aggressor employs tactics like snarky or passive-aggressiveness to cover up the abuse, it can be hard to spot.

FURTHER READING

Here, we take a look at the current state of the art in CB identification and categorization using Twitter data. The categorization of abuse tweets is a popular application of machine learning (ML) based techniques, with a variety of feature selection methods. To identify instances of harassment in Twitter, Parameswara et al. [26] used a support vector machine (SVM) and Information Gain (IG) based feature selection approach. Classifiers such as AdaBoost(ADB), Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM), Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Multinomial Naive Bayes (MNB) were used by Muneer and Fate [11] to identify instances of harassment in Twitter. In this research, Word2Vec and TF-IDF were used to identify characteristics. To identify instances of harassment in Twitter, Dalvi et al. [12, 27] employed SVM and Random Forests (RF) models that extracted features using TF-IDF. While SVM was successful in these models, the intricacy of the model grew with the number of classes. Cyberbullying detection was studied by Al-gardai et al. [28], who looked at how

different ML classifiers—including RF, Naive Bayes (NB), and SVM—could perform on Twitter data taken from a variety of sources. (Tweet content, activity, network, and user). The method proposed by Huang et al. [29] for detecting CB in social media combined social media and written content characteristics. The IG technique is used to rate the characteristics. Popular categorization systems like NB, J48, and Bagging and Digging are used.

The results suggested that social traits might help improve harassment identification. Cyberbullying and cyberbullying forecast on social networks like spring.me and Myspace were identified and predicted by Zuccarini et al. [30] using a decision tree (C4.5) algorithm with a social network, personal, and linguistic characteristics. Different ML algorithms were used by Balakrishnan et al. [31] to identify cyberbullying events from tweets and to categorize tweets into cyberbullying groups such as aggressors, spammers, bullies, and regular users. Researchers found that sentiment analysis had no effect on identification accuracy. While effective, this model can only be used with a modest collection containing few class names. Using the single and double ensemble-based voting model, Allam et al. [32] suggested an ensemble-based categorization strategy. These ensemble-based voting models extracted features using mutual information bigrams and unigram TF-IDF, and used decision trees, LR ensemble model classifiers, and Bagging ensemble model classifiers for classification. The Bagging ensemble model offered the highest accuracy when analysing the Twitter dataset, but it also took into account other factors. However, when used with sarcastic tweets or abbreviation words with multiple meanings, these ensemble models' decreased training and implementation time for categorization becomes a significant restriction. Chia et al. [8] also used ML and feature engineering-based methods to separate cyberbullying tweets with irony and snarky. While this technique significantly identifies snarky and ironic words among cyber-bullying tweets, the detection rate is still very low [33]. Many classifications and feature selection methods were tried.

METHODOLOGY

In Fig. we can see the full DEA-RNN model. 1. (I) data gathering, (ii) data labelling, (iii) pre-processing and data cleaning, (iv) feature extraction and feature selection, and (v) categorization are all components of the model. Each of these factors is emphasized in the sections that follow.

DATA GATHERING

The raw dataset is compiled of around 32 cyber-bullying terms used to gather messages via Twitter API broadcasting. Keywords such as "idiot," "in**er," "le***an," "g*y," "bisexual," "transgender," and "queer," "whore," "pussy," "faggot," "crap," "sucker," "slut," "donkey," "live," "afraid," "moron," "poser," "rape," "fuck," "fucking," "ugly," "bitch," " While [39] suggests other terms like "ban," "kill," "die," "evil," "hate," "attack," "terrible," "threat," "racist," "black," "Muslim," "Islam," and "Islamic." About 130000 tweets are based on terms that include racist, sexist, homophobic, and other offensive language found in the original collection of 435764. There are a lot of unusual tweets in this collection. Since only English-language tweets are required, shares and tweets having words from other languages are omitted. 1. After these kinds of unnecessary tweets have been filtered out, a collection consisting of around 10,000 tweets is arbitrarily chosen from the remaining tweets. These steps are performed mechanically as part of the pre-processing phase. The remaining essential pre-processing steps are carried out in the same manner as described in subsection III-C.

Annotation of Data

Here, we'll focus on tagging and categorizing some of the messages from the initial Twitter dataset. Over the course of 1.5 months, three human annotators chose 10,000 tweets at random from the collected tweets and individually classified them as either "0" (non-cyber bullying) or "1" (cyber bullying). During the tagging process, human annotators assigned labels to each occurrence based on whether or not it met the criteria for cyberbullying, as outlined in [57]. Character attacks, taunts, competence attacks, malediction, verbal abuse, taunting, name calling, ridicule, threats, and physical looks are all factors in determining whether or not an incident of cyberbullying has occurred. At first, two annotators categorized each text, and the degree of consensus between them was roughly 91%. The inconsistencies found during the first round of writing were then addressed by a third person. After addressing inconsistencies and cleaning the data, we were left with a dataset of 10,000 categorized tweets, of which 6,508 (or 0.65) are not cyberbullying and 3492 (or 0.35%) are cyberbullying. The annotated Twitter dataset is skewed if one considers the ratio of cyberbullying to non-cyberbullying messages. There are a wide range of Twitter counts across disciplines. Therefore, methods of achieving a balance, like oversampling or under sampling, are used to address the problem. To address the class disparity between cyberbullying and non-cyberbullying, we have used the Synthetic Minority Oversampling Technique (SMOTE) to artificially increase the

sample size of the minority class (cyberbullying Tweets). Oversampling is carried out by repeatedly duplicating samples of harassment in order to create a more representative dataset, as shown in [15], [16]. Therefore, the sum of all comments.

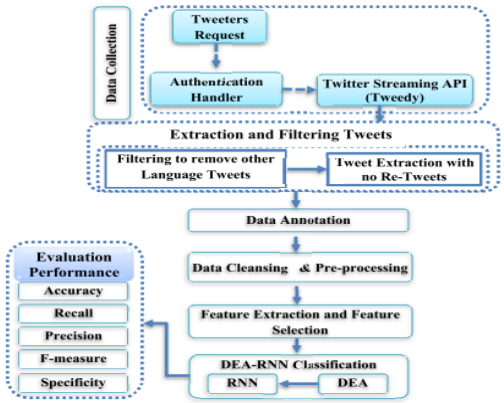

FIGURE 1. Methodology of the proposed model.

TABLE 1. The details of twitter dataset versions.

Dataset	Total number of tweets	Number of Cyberbullying Tweets	# of Non-Cyberbullying Tweets
Original Twitter	10,000	3,492	6,508
Oversampled Twitter	13016	6,508	6,508

after oversampling was 13,016 samples. Table 1 shows the original dataset and the dataset with oversampling.

LABORATORY TESTS

Here, we use the measures of recall, precision, F-measure, accuracy, and sensitivity to assess DEA-RNN's performance on Twitter-crawled datasets. Both the annotated data and the input dataset are discussed in detail in part III-A and III-B, respectively. The suggested DEA-RNN model is compared to three other models: two baseline cyberbullying models based on deep learning (Bi-LSTM [21], RNN [21], and SVM [26], MNB [11], and R [11]) and three benchmark cyberbullying models based on machine learning (SVM [26], MNB [11], and R [11]). These models are representative of the current state of the art in social media abuse identification. The examined standard models' design settings are the same as those used in the initial articles. On the other hand, PyCharm IDE 2020.2.3 and Python 3.7.4 were used in the tests. Kera's, TensorFlow, NumPy, NLTK, Scikit-learn, Tweepy, etc. were used among others in the settings for both the application and the tests. Pre-processing stages are executed using the NLTK Python software, as suggested in [58], and the trial assessments are run on a personal computer with settings of an Intel Core-i5 CPU, Windows 10, and 8 Gigabyte RAM. The raw information is split into two parts, training and assessment. For the purpose of the analysis, it is also split into three groups: 60% (40%), 70% (30%, 10%), and 90% (10%. (Scenario 3). The assessment measures are selected to best illustrate how each approach classifies tweets. Five-fold cross-validation is used, and each technique is evaluated N = 20 times to get an average score.

INDICATORS OF QUALITY

This subsection provides a quick summary of the measures this research used to assess DEA-RNN's efficacy. Metrics such as calculating training time, accuracy, memory, precision, F-measure, and sensitivity are used in the assessment procedure. N = 20 iterations of each technique are performed across all trials to generate a weighted average of the findings. Table 1 details these success indicators.

RESULTS OF EXPERIMENTS

In this subsection, we compare the DEA-RNN classifier's trial findings to those of the Bi-LSTM, RNN, and MNB, RF, and SVM baseline deep learning models and baseline machine learning models. Cyberbullying forecasts are checked against three different input datasets with success rates of 60% (Scenario 1), 70% (Scenario 2), and 90% (Scenario 3). (Scenario 3). The aforementioned measures are used to assess success. Each classifier's tests were run M = 20 times for each input situation in the dataset. Then, the formulae in Table 2 are used to determine the mean of the success indicators. Table 3 shows the aggregate performance comparative findings of different classifications under varying input situations for various datasets.

ROUGHLY ACCURATE

By calculating the average precision across all situations, the suggested DEA-RNN model is contrasted to the current models that were taken into consideration. Figure 3 displays the results of various models, with the DEA-RNN model achieving the best average accuracy of 90.45% in case 3, followed by the Bi-LSTM model at 88.74%, the RNN model at 87.15%, the SVM model at 85.21%, the MNB model at 82.26%, and the RF model at 83.45%. Results show that deep learning models (Bi-LSTM and RNN) outperform their machine learning counterparts. (SVM, RF, and MNB). When compared to other versions, the MNB type performs the worst. Comparing the

accuracy findings of other existing Bi-LSTM, RNN, SVM, MNB, and RF models, we find that the DEA-RNN achieves 87.14%, with case 2. This is the best accuracy number. The suggested model also outperformed the current models in the assessment procedure and produced the best outcomes (82.25%) in case 1. Among all the models, Bi-LSTM performed the best, while MNB performed the worst. According to Fig. 1, it is clear that Scenario 3 yields the most accurate findings from the suggested model and the other approaches. 3.

Algorithm 1: DEA-RNN

```
Begin
  Initializes DEA population, size dimension and Elman RNN structure
  Load the training data
  While (MSE < stopping criteria)
    Pass the DEA locations as weights to the network
    Feed-forward network runs using the weights initialized with DEA
    For each solution candidate
      Compute the error utilizing Eq. (15)
      Minimize the error using adjusting network parameter by utilizing DEA
      Generate DEA next loop location (i + 1). (From random locations using Eq. (6))
      Eliminate a fraction of the worst solutions.
      Find new solutions to replace the old ones.
      Assess the fitness function to select the best configuration of RNN
      If new location (i + 1) > old location (i)
          Replace old location (i) with the new location (i + 1)
      End if
    End for
    DEA estimates weight and bias at each iteration Until the network is converged
    Update weights and bias utilizing Eq. (20) & (21)
  End While
End
```

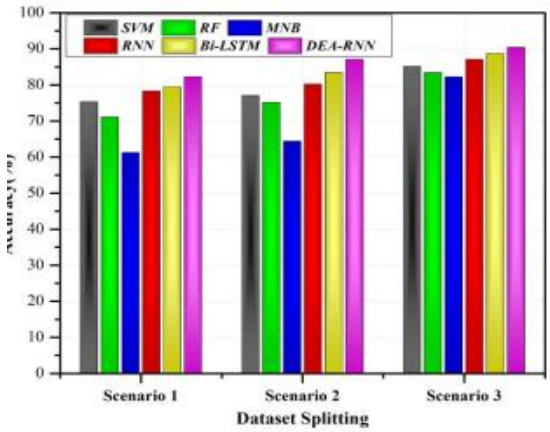

FIGURE 2. Performance evaluations in terms of average accuracy.

ABOUT ACCUATE

The average accuracy findings of the suggested DEA-RNN model are displayed in Fig. 4 in comparison to the current models that were studied. While the other current models, including the Bi-LSTM, RNN, SVM, MNB, and RF, all received results in the 87.9%-86.62%-84.25%-80.01%-83.87% range for Scenario 3, the DEA-RNN achieved 89.52%. Comparing the findings of other existing Bi-LSTM, RNN, SVM, MNB, and RF models, we find that DEA-RNN's 87.02%, with case 2, is the best accuracy value. This is in comparison to results of 82.88%, 80.09%, 76.6%, 75.78%, and 78.96%. Among the models tested, Bi-LSTM achieved the second-highest accuracy score in cases 2 and 3, while MNB performed the worst. Figure 4 shows that using (case 3) yields better outcomes in terms of accuracy measure than any of the other situations.

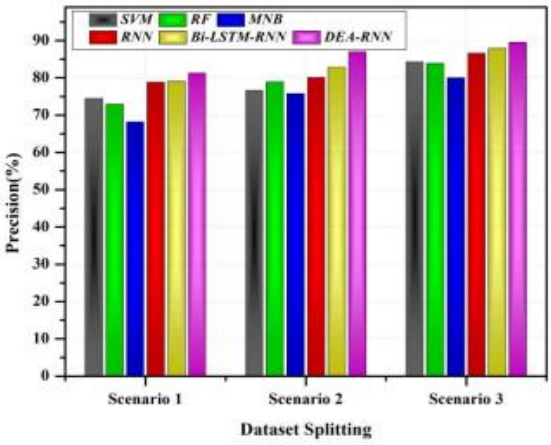

FIGURE 3. Performance evaluations in terms of average precision.

DISCUSSION

The proposed model's performance has improved in terms of sensitivity, f-measure, precision, memory, and accuracy, as measured by the Performance Improvement Rate (PIR). Two deep learning models and three machine learning models are taken into account for comparison, with the aggregate results of each model used to calculate the PIR. In Scenario 2, the suggested model outperforms the benchmark models Bi-LSTM [21], RNN [21], SVM [26], RF [11], and MNB [11] by 3,69%, 6,91%, 10,04%), 12%, and 22,69% in terms of accuracy. When comparing Bi-LSTM, RNN, SVM, RF, and MNB, the corresponding PIRs of precision for Scenario 3 are 1.71, 3.3, 5.2, 7, and 8.19%, respectively. Scenario 2's suggested model outperforms the benchmark models Bi-LSTM, RNN, SVM, RF, and MNB by 4.14 percentage points, 6.93 percentage points, 10.42 percentage points, 8.06 percentage points, and 11.24 percentage points, respectively, in terms of accuracy. Similarly, the accuracy rates of Bi-LSTM, RNN, SVM, RF, and MNB are 1.62%, 2.9%, 5.65%, and 9.51 in Scenario 3. In Scenario 2, the suggested model outperforms Bi-LSTM, RNN, SVM, RF, and MNB with recollection rates of 4.33 percent, 10.03 percent, 17.24 percent, 6.48 percent, and 10.09 percent, respectively. Scenario 3

precision performance is enhanced by 1.46 percentage points, 3.08 percentage points, 6.26 percentage points, 6.48 percentage points, and 10.09% when utilizing Bi-LSTM, RNN, SVM, RF, or MNB, respectively. Scenario 2 F-Measure gains for the suggested model versus Bi-LSTM, RNN, SVM, RF, and MNB are 5.74 percent, 6.72 percent, 10.21 percent, 9.06 percent, and 14.32 percent, respectively. In Scenario 3, the performance increase rates of accuracy for Bi-LSTM, RNN, SVM, RF, and MNB are respectively 1.54%, 2.99%, 5.77%, 6.08%, and 9.8%.

The Final Thoughts and Future Plans

This study improved the ability of subject models to identify instances of cyberbullying by creating a robust Twitter categorization model. In order to efficiently tune RNN parameters, the DEA algorithm was combined with the Elman type RNN to create DEA-RNN. In addition, it was put through its paces against the likes of the established Bi-LSTM, RNN, SVM, RF, and MNB techniques by way of an artificial Twitter dataset mined for CB terms. The trial evaluation revealed that the DEA-RNN outperformed all other available techniques across a wide range of measures, including accuracy, memory, F-measure, precision, and sensitivity. This represents how DEA affects RNN functionality. The combination suggested model outperformed the other existing models that were taken into account, but the feature compatibility of DEA-RNN degrades as more data is fed into it. The current research only used data from Twitter, but other SMPs like Instagram, Flickr, YouTube, Facebook, etc. should be looked into as well in order to spot the growing problem of harassment. Then, in the future, we will look into the potential of using data from a variety of sources to identify instances of cyberbullying. In addition, we were only able to analyse the tweets' substance, rather than the tweets' connection to the users' actions. This is planned for upcoming projects. While the suggested model is effective at detecting cyberbullying using the written content of tweets, further study into the detection of cyberbullying using other types of media, such as pictures, video, and audio, remains an open research field and potential path for future work. We also hope to identify and categorize CB comments in real-time streams.

REFERENCES

[1] F. Mishna, M. Khoury-Kasseri, T. Gardella, and J. Daciuk, ''Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims,'' Children Youth Services Rev., vol. 34, no. 1, pp. 63–70, Jan. 2012, doi: 10.1016/j.childyouth.2011.08.032.

[2] K. Miller, ''Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress,'' Southern California Interdiscipl. Law J., vol. 26, no. 2, p. 379, 2016.

[3] A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, ''A systematic review and content analysis of bullying and cyber-bullying measurement strategies,'' Aggression Violent Behav., vol. 19, no. 4, pp. 423–434, Jul. 2014, doi: 10.1016/j.avb.2014.06.008.

[4] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ''Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren,'' PLoS ONE, vol. 9, no. 7, Jul. 2014, Art. no. e102145, doi: 10.1371/journal.pone.0102145.

[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ''Improving cyberbullying detection with user context,'' in Proc. Eur. Conf. Inf. Retr., in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7814, 2013, pp. 693–696.

[6] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ''BullyNet: Unmasking cyberbullies on social networks,'' IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332–344, Apr. 2021, doi: 10.1109/TCSS.2021.3049232.

[7] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan, and M. Prasad, ''Identification and classification of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting,'' in Neural Information Processing (Communications in Computer and Information Science), vol. 1333, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 113–120.

[8] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, ''Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection,'' Inf. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102600, doi: 10.1016/j.ipm.2021.102600.

[9] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, ''Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking,'' Math. Problems Eng., vol. 2021, pp. 1–12, Feb. 2021, doi: 10.1155/2021/6644652.

[10] B. A. Talpur and D. O'Sullivan, ''Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter,'' Informatics, vol. 7, no. 4, p. 52, Nov. 2020, doi: 10.3390/informatics7040052.

[11] A. Muneer and S. M. Fati, ''A comparative analysis of machine learning techniques for cyberbullying detection on Twitter,'' Futur. Internet, vol. 12, no. 11, pp. 1–21, 2020, doi: 10.3390/fi12110187.

[12] R. R. Dalvi, S. B. Chavan, and A. Halbe, ''Detecting a Twitter cyberbullying using machine learning,'' Ann. Romanian Soc. Cell Biol., vol. 25, no. 4, pp. 16307–16315, 2021.

[13] R. Zhao, A. Zhou, and K. Mao, ''Automatic detection of cyberbullying on social networks based on bullying features,'' in Proc. 17th Int. Conf. Distrib. Comput. Netw., Jan. 2016, pp. 1–6, doi: 10.1145/2833312.2849567.

[14] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, ''XBully: Cyberbullying detection within a multi-modal context,'' in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339–347, doi: 10.1145/3289600.3291037.

[15] K. Reynolds, A. Kontostathis, and L. Edwards, ''Using machine learning to detect cyberbullying,'' in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), vol. 2, Dec. 2011, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[16] S. Agrawal and A. Awekar, ''Deep learning for detecting cyberbullying across multiple social media platforms,'' in Advances in Information Retrieval (Lecture Notes in Computer Science), vol. 10772, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham, Switzerland: Springer, 2018, pp. 141–153.

[17] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, ''Careful what you share in six seconds: Detecting cyberbullying instances in vine,'' in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2015, pp. 617–622, doi: 10.1145/2808797.2809381.

[18] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, and A. R. Rajan, ''Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification,'' Comput. Electr. Eng., vol. 92, Jun. 2021, Art. no. 107186, doi: 10.1016/j.compeleceng.2021.107186.

[19] A. Al-Hassan and H. Al-Dossari, ''Detection of hate speech in Arabic tweets using deep learning,'' Multimedia Syst., Jan. 2021, doi: 10.1007/s00530-020-00742-w.

[20] Y. Fang, S. Yang, B. Zhao, and C. Huang, ''Cyberbullying detection in social networks using bi-GRU with self-attention mechanism,'' Information, vol. 12, no. 4, p. 171, Apr. 2021, doi: 10.3390/info12040171.

[21] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, ''Cyberbullying detection solutions based on deep learning architectures,'' Multimedia Syst., 2020, doi: 10.1007/s00530-020-00701-5.