International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I9P240

Volume 70 Issue 9, 393-399, September 2022

Original Article

An Enhanced Text Mining Approach using Ensemble Algorithm for Detecting Cyber Bullying

Zarapala Sunitha Bai[1], Sreelatha Malempati[2]

[1,2]Department of Computer Science and Engineering, R.V.R. and J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India.
[1]Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

[1]Corresponding Author : zsunithabai@gmail.com

Received: 15 June 2022 Revised: 07 September 2022 Accepted: 20 September 2022 Published: 30 September 2022

Abstract *- Text mining (TM) is most widely used to process the various unstructured text documents and process the data present in the various domains. The other name for text mining is text classification. This domain is most popular in many domains, such as movie reviews, product reviews on various E-commerce websites, sentiment analysis, topic modeling, and cyberbullying on social media messages. Cyberbullying is the type of abusing someone with insulting language. Personal abuse, sexual harassment, and other abuse come under cyberbullying. Several existing systems are developed to detect bullying words based on their situation on social networking sites (SNS). SNS becomes a platform for bullying someone. In this paper, An Enhanced text mining approach is developed by using Ensemble Algorithm (ETMA) to solve several problems in traditional algorithms and improve the accuracy, processing time, and quality of the result. ETMA is the algorithm used to analyze the bullying text within the social networking sites (SNS) such as Facebook, Twitter, etc. The ETMA is applied to a synthetic dataset collected from various data sources consisting of 5k messages belonging to bullying and non-bullying. The performance is analyzed by showing Precision, Recall, F1-Score, and Accuracy.*

Keywords - *Deep Learning, Cyber Bullying, Text Mining, Ensemble Algorithm.*

1. Introduction

Web 2.0 is most widely used to improve user-created platforms for social networking sites (SNS) users [1]. TM is a data mining (DM) sub-domain to mine accurate patterns. TM is most widely used to extract patterns from various text documents or text data. In TM, many types of structured and unstructured are present for analysis. TM is also used to process large datasets by extracting interesting and required information that is useful in various applications. Every day huge amounts of data are generated on social media. Social networking sites (SNS) are most widely used to communicate with various users. In [2], various ML algorithms discuss the challenges faced in detecting cyberbullying. ML algorithms make the prediction more accurate than the behavior of humans [3].

Nowadays, social media messages are generating more and more day by day. Huge messages are generated by the various types of users in SNS. Cyberbullying becomes more complicated on SNS because personal abuse becomes more complicated on social media platforms. Detecting the bullying messages and preventing this message is more complex for the SNS developers. Cyberbullying is a challenging task that can be done in different ways, such as morphing photos, using tough language, uploading personal videos, etc. This research is mainly focused on text bullying. This text bullying can be stopped with the automated technology built-in to find the cyberbullying activities and remove them from the SNS platform. Finding these bullying comments is a very tough task because classifying these bully words or comments is unique. It is very important to know the exact bullying comments. In cyberbullying, there are three types of users: sufferer, forecaster, and analyzer.

In this paper, An Enhanced text mining approach is developed by using Ensemble Algorithm (ETMA) to analyze the messages collected from Twitter data. This dataset contains two folders, a training set, and a testing set. Training contains 10k comments, and the testing set contains 7k Twitter messages with 7 attributes. The ETMA follows the powerful pre-processing after initializing the dataset. The training removes the noise text from the dataset and gives better text based on the bully score. The performance is analyzed by showing the parameters such as precision, recall, accuracy, F1-measure, and duration.

2. Literature Survey

G. M. Abaido [4] proposed the approach that detects cyberbullying among Arab community students. The data is collected from 200 students belonging to UAE. 91% of analysis shows that cyberbullying occurs on social media apps such as Instagram (55.6%) and Facebook (37.9%).

D. Chatzakou et al. [5] proposed a principled and scalable approach that detects bullying and aggressive behaviour on Twitter. This is a very fast approach that extracts the text given by the users and analyzes the users that are with aggressive behaviour. E. Raisi et al. [6] proposed the ML approach that analyses the user roles in harrying-based bullying and new grammatical measures of bullying. This is called participant-vocabulary consistency (PVC). E. Raisi et al. [7] introduced the proposed approach that solves various issues in analyzing bullying using significant properties. The proposed approach is widely used in various applications such as Twitter and Ask. Fm, Instagram data, etc.

Vijay B et al. [8] involved one more technique for distinguishing proof of cyberbullying. This structure used convolution neural framework estimation, which manages various layers and provides careful requests. Like this, a continuously clever way that stood out from the ordinary course of action computations was arranged.

Monirah A et al. [9] explored the current Twitter cyber-bullying revelation frameworks and proposed one more request strategy subject to deep learning. The proposed approach (OCDD) was collected using planning data set apart by a human understanding organization, and subsequently, a word introduction was delivered for each word using (GloVe) procedure. They devised a game plan of word embedding that was subsequently supported by CNN computation for portrayal. Batoul H et al. [10] discussed several approaches to detecting cyberbullying messages in Arabic. The authors in this paper made their research on multilingual cyberbullying detection and finally proposed a solution for the issue of Arabic cyberbullying.

Xiang Z et al. [27] proposed the novel pronunciation-based convolutional neural network (PCNN) to solve cyberbullying issues. The proposed approach corrects the spelling mistakes and pronunciation of the given bullying text by solving the noise and bullying data sparsity. Various integrated approaches are included to solve these issues and achieve better results.

A. S. Srinath et al. [12] introduced the three-stage approach known as BullyNet. It mainly focuses on detecting the bully's words on a social network like Twitter. A robust approach is also developed as a cyberbullying signed network (SN). This approach also analyzed the situations in which context by analyzing their bully score. The proposed approach shows the high accuracy of detecting bullies from many tweets. R. Zhao et al. [13] proposed a new approach to solve the detection of cyberbullying words. The proposed approach is extended based on the semantic deep learning approach stacked denoising autoencoder (SDA). This approach also focused on extracting the hidden features from the bullying data and training the fast and discriminative text initialization. H. Rosa et al. [14] proposed the fuzzy fingerprints that detect textual cyberbullying in social media. This approach also solves the extraction the accurate cyberbullying data. X. Zhang et al. [15] proposed the novel pronunciation-based convolutional neural network (PCNN) to overcome several challenges by observing the misspelled words in the dataset by using CNN. This approach dynamically corrects the spelling exceptions that didn't change the pronunciation. The proposed system is applied to two datasets collected from Twitter and Formspring. Results show that the PCNN shows high performance. Johnson et al. [16] proposed the segmentation approach that divides sentences into words to understand the meaning of every word. The segmentation is done based on the length of the word and analyzes the sensitivity of the words. S. Salawu et al. [17] reviewed various cyber-bullying detection approaches. These approaches mainly focus on finding the bullying patterns from the complex social media datasets. Many existing approaches are analyzed based on the given dataset. An accurate cyber-bullying dataset is defined based on the labeled and unlabelled datasets. I. Nazar et al. [28] proposed the novel hierarchical approach that detects cyberbullying based on the types of words. Based on the number of messages, the decision is taken regarding the type of cyberbullying. Y. Win [19] introduced the supervised method that detects cyberbullying in Myanmar social media posts containing the Myanmar language. This approach contains the integrated approaches containing pre-processing, word segmentation, and applying SVM to detect the bullying words in the specific language. This approach mainly achieved a very low accuracy of 75.40%.

3. Materials and Methods

The dataset consists of 16k bullying and non-bullying messages (Sentences). As per the dataset description, there are 6135 bullying messages, 7235 non-bullying messages, and 2630 normal messages. The proposed algorithm is applied to these Twitter datasets. The algorithm process the overall dataset for accurate analysis. The messages are divided into five types such as attack, sexual harassment, personal abuse, flaming, and cyber-stalking.

Table 1. Dataset Description

Message Type	Training	Testing
Bullying	10k	7k
Non-Bullying	4k	6k
Normal Messages	2k	2k

The above table shows the training and testing set of the given dataset. Here we are finding the bully and non-bully comments or words.

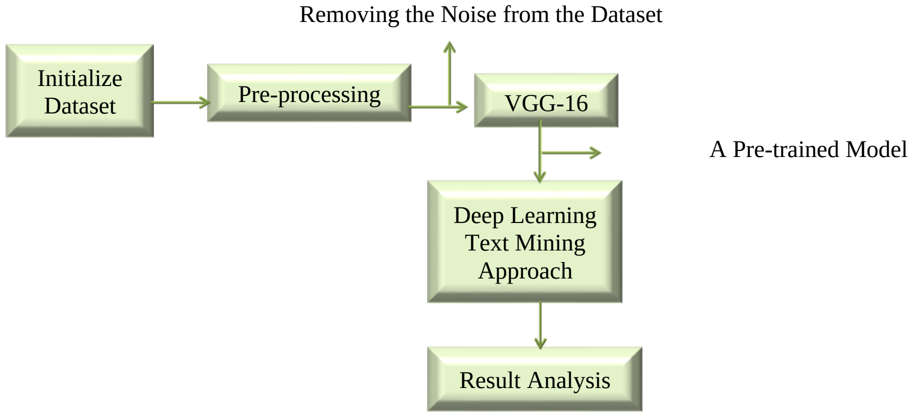

Fig. 1 Process of Enhanced Text Mining Approach is developed by using Ensemble Algorithm (ETMA)

4. Feature Extraction using TF-IDF (Term Frequency (TF) and Inverse Document Frequency (IDF))

TF-IDF Lin L., [20] is one of the mathematical approaches that counts the relevant word in a specific document and is also used in many documents. Based on the two parameters, the count of words that occur in a document and IDF counts the number of relevant words in a group of documents. Here every document is considered a message in the dataset. TF measures the high term frequency value based on the number of words. Thus the TF is used to measure the overall all relevant documents from the irrelevant documents is low because of its innocence in frequency collection. To overcome this issue, the IDF is proposed better to classify text by A. Onan [21]. IDF is extracted from DF, which shows the number of terms that occur in the total number of documents M. Lan et al., [29].

$$\mathrm{IDF}\,(\mathrm{t_{er}}, \mathrm{d_{oc}}, \mathrm{D_o}) = \log \frac{|\mathrm{D_o}|}{\mathrm{DF}(\mathrm{t_{er}}, \mathrm{D_o})} \qquad (1)$$

In (2), DF shows the term. ‘$\mathrm{t_{er}}$’ in corpus ‘$\mathrm{D_o}$’. The symbol in Equ-2 represents the overall tweets in the corpus $\mathrm{D_o}$. To prevent abnormal cases, the formula is given:

$$\mathrm{IDF}\,(\mathrm{t_{er}}, \mathrm{d_{oc}}, \mathrm{D_o}) = \log \frac{|\mathrm{D_o}| + 1}{\mathrm{DF}(\mathrm{t_{er}}, \mathrm{D_o}) + 1} \qquad (2)$$

To improve the performance of IDF, the TF is to be merged, called FT-IDF. It is also called a global statistical measure. The final equation is represented below.

$$\mathrm{TF} - \mathrm{IDF}\,(\mathrm{t_{er}}, \mathrm{d_{oc}}, \mathrm{D_o}) = \mathrm{TF}(\mathrm{t_{er}}, \mathrm{d_{oc}}) * \mathrm{IDF}\,(\mathrm{t_{er}}, \mathrm{d_{oc}}, \mathrm{D_o}) \qquad (3)$$

In equation (3), initializes the t as the tweet $\mathrm{d_{oc}}$ in corpus $\mathrm{D_o}$ and TF value of the term $\mathrm{t_{er}}$ is present in the document $\mathrm{d_{oc}}$. Finally, this approach measures the total number of stop words, bad words, and the total number of word count.

4.1. An Enhanced Text Mining Approach is developed by using Ensemble Algorithm (ETMA)

In this paper, the proposed algorithm focused on detecting cyberbullying with the combination of robust pre-processing, TF-IDF, with the merging of the convolutional neural network (CNN) algorithm, which is called ETMA, and this can solve complex issues. The main aim of this approach is to develop the efficient detection of cyberbullying based on accurate meanings and reduce computational time and cost. By using CNN, efficient classification of bullying words is done R. Johnson et al., [23], X. Zhang et al., [24], and A. J. McMinn et al., [25]. The significant feature of this ESTM is to reduce the workflow of classical detection, which makes detection without any features. ESTM transforms text into word embeddings as an input. In the existing approaches, the process starts with feature extraction followed by feature selection. The proposed approach in this paper starts with robust pre-processing and effective training with VGG-16, called a pre-trained model, to get better results.

4.1.1. Training

The training is done by using VGG-16. VGG-16 is the pre-trained model that trains any type of data, such as image format, text format, etc. In this paper, the training is done using stop words, word count, average words, etc. Some define the negatives, and some words define the positives.

Data Cleaning: In this step, the data cleaning is done based on the given tweet. The following are the steps that are used to process the data cleaning.

- On the white space, the tokens are Split.
- Punctuation from words is removed.
- All the known stop words are removed.
- A length <= 1 character is removed.

4.1.2. Example

Real Donald trump, you are the man Donald trump donâ€™t listen to anyone else, ever follow your own instincts and god-given ability; thank god for Donald trump; much love.

The above sentence is a normal tweet containing noise, such as special words. (#$%). So, removing these types of noises from this tweet is important. A normal word is considered as the A length <=1.

4.2. Noise Removal from the Dataset

In social networking data, noise removal plays a major role. Several finite schemes are based on the interpolation of polynomials used to remove the noise from the datasets. But these noise filters are not fit for Twitter datasets. In this paper, to remove the noise from the Twitter datasets, the 9-point technique is the high number of points that shows the noise reduction for the selected dataset. The 9-point formula is given as:

$$\mathcal{F}_0' \approx \frac{(\mathcal{F}_1 - \mathcal{F}_{-1}) + 2(\mathcal{F}_1 - \mathcal{F}_{-2}) + 3(\mathcal{F}_1 - \mathcal{F}_{-3}) + 4(\mathcal{F}_1 - \mathcal{F}_{-4})}{60\text{h}} \quad (4)$$

Another 9-point finite difference model is represented as:

$$\mathcal{F}_0' \approx \frac{1}{4}\left(w_1 \frac{\mathcal{F}_1 - \mathcal{F}_{-1}}{2h} + w_2 \frac{\mathcal{F}_1 - \mathcal{F}_{-2}}{4h} + w_2 \frac{\mathcal{F}_1 - \mathcal{F}_{-3}}{6h} + w_4 \frac{\mathcal{F}_1 - \mathcal{F}_{-4}}{8h}\right) \quad (5)$$

5. Convolutional Neural Networks (CNN) for Classification of Cyber-bullying

In CNN, the input is given to the input layer. For the embedding_layer, the embedding matrix is passed. Different sizes of filters are applied to the Twitter dataset, and for every layer, the GlobalMaxPooling1D is applied. Then all the outputs are merged. A dropout and dense layer are applied and called a final dense layer.

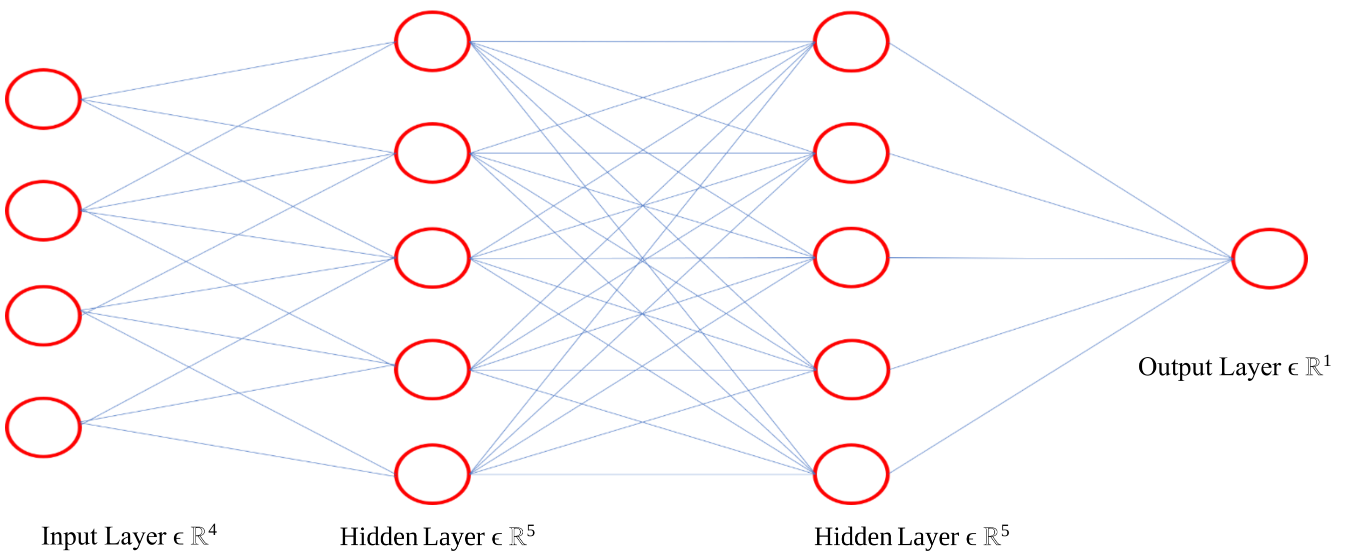

Fig. 2 CNN Architecture

5.1. Input Layer

In this layer, the input_data selects the two factors, such as shape and the name for the input layer. This layer becomes the component that feeds data to the neural network. This paper's shape is one-dimensional, consisting of long sentences and Nil for batch size. The shape of the input data (text or bully) is represented as [Nil, max_words], where Nil is the batch size. This layer considers the input data as parameters and input/output dim. The input_dimension initializes the overall vocabulary indexes, and the output_dimension initializes the embedding size. The output_dimension can change for various approaches.

5.2. Convolutional Layer

This layer mainly focuses on performing the convolution operation to extract the new matrix with convolved features. The filter is slid (convolve operation) over the matrix in the matrix. This matrix initializes the input data (text with words), so this matrix consists of digits. These digits filter the neuron's weights (parameters) that update during training. This operation creates a new matrix consisting of convolved features sent to the next layer for analysis.

5.3. Fully Connected Layer

In this layer, the main function is fully_con which takes the previous layer output as the input and predicts the number of classes such as Aggression, Attack, Bad words, Racism, Sexism, Toxicity, and the activation function is called as softmax function. A regression layer is present in this layer, and all the previous layers are combined, and the classification sentences are measured.

$$\mathrm{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \tag{6}$$

It is considered the final output layer in NN that performs the multi-class classification for a given Twitter dataset, such as Aggression, Attack, Bad words, Racism, Sexism, and Toxicity. Based on the given tweet, the softmax function measures the scores of the tweet and assigns the values for each type of tweet, which is considered bully text. The overall results are analyzed using a confusion matrix and analyzing the performance of the proposed approach.

5.4. Algorithm Steps

Step-1: Initialize the input_layer, Conv_layer, hidden-layer and Output_layer.
Step-2: Training with VGG-16
Step-2: Remove the noise
Step-4: Filter/Kernel for feature extraction using Equ (6).
Step-5: Classify the text
Step-6: Show Results
Step-7: Apply the confusion matrix.
Attack, sexual harassment, personal abuse, flaming, and cyber-stalking.

6. Experimental Results

Experiments are conducted by using the python programming language. The most powerful and popular libraries used to process the dataset are numpy, keras, pandas, etc. The performance metrics that are used to analyze the proposed approach

6.1. Precision

This is one of the significant metrics that shows the correct positive predictions based on the type of Twitter.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

6.2. F1 Measure

F1-measure is the metric that merges recall and precision.

$$\mathrm{F1\ Measure} = 2 \times \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{8}$$

6.3. Accuracy

This parameter plays a major role in showing the overall accuracy.

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{9}$$

6.4. Recall

This metric is mainly focused on reducing false negatives.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{No.\,of\,TP} + \mathrm{No.\,of\,FN}} \tag{10}$$

Table 2 shows the comparison between SVM, CNN, and TF-IDF-CNN. Among all the approaches, the proposed approach TF-IDF+CNN achieved high performance for classifying Twitter data.

Table 1. Detecting the performance for TF-IDF+CNN

Bully Type	Precision	F1-Measure	Accuracy	Recall	Duration (Sec)
Attack	90.23%	96.23%	96.78%	96.34%	34.34
Sexual	91.34%	96.78%	97.56%	96.7%	24.1
Harassment	95.45%	96.34%	97.34%	94.9%	23.5
Personal Abuse	94.54%	95.12%	98.34%	93.12%	26.4
Flaming	94.23%	94.34%	99.23%	96.34%	27.8
Cyber-staking	93.23%	93.23%	98..34%	96,78%	32.5

Table 2. Performance of Existing and Proposed Algorithms

	SVM	CNN	TF-IDF+CNN
Precision	78.98%	82.12%	89.89%
F1-Measure	80.12%	84.32%	90.12%
Accuracy	81.23%	85.12%	92.32%
Recall	82.34%	87.12%	94.12%
Duration (Sec)	2.56	1.56	58.34

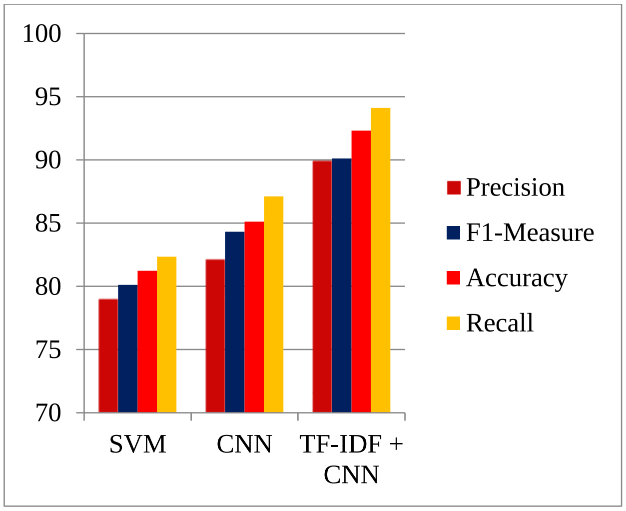

Fig. 3 Comparison Graph between Existing and Proposed Algorithms

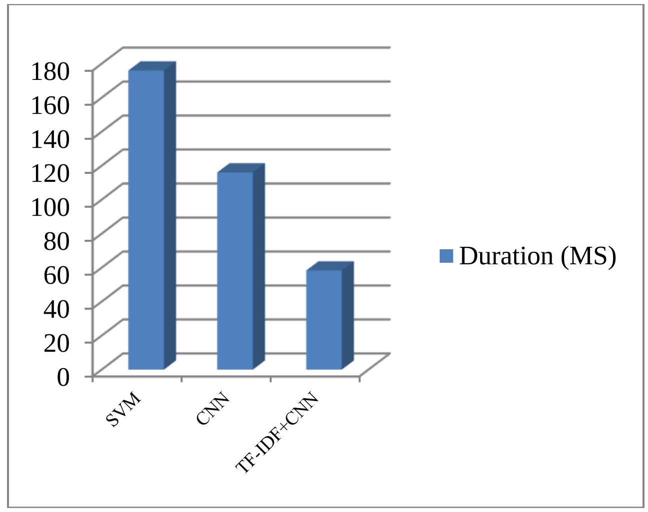

Fig. 4 Comparison of time between existing and proposed algorithms

6. Conclusion

Cyberbullying becomes more complicated for OSNS users. Cyberbullying is also considered a personal attack, mainly occurring in OSNS using text messages, videos, and images. This paper mainly focused on classifying the several types of cyber-bullying 'attacks' such as Aggression, Attack, Bad words, Racism, Sexism, and Toxicity. The text messages that are given by various types of users indicate whether cyberbullying is occurring or not. The ETMA in this paper showed a huge performance compared with existing approaches such as SVM and CNN. The proposed approach ESTM is the combination of TF-IDF and CNN, achieving the performance of Precision-89.89%, F1-Measure-90.12%, Accuracy-92.32%, Recall-94.12%, Duration (Sec)-58.34. The performance is increased using the strong pre-trained model integrated with a robust pre-processing approach.

References

[1] F. Elsafoury, S. Katsigiannis, Z. Pervez and N. Ramzan, "When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection," *in IEEE Access*, vol. 9, pp. 103541-103563, 2021, doi: 10.1109/ACCESS.2021.3098979.

[2] 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.

[3] M. A. Al-Garadi, K. D. Varathan and S. D. Ravana, "Cybercrime Detection in Online Communications: the Experimental Case of Cyberbullying Detection in the Twitter Network," *Computers in Human Behavior*, vol. 63, pp. 433-443, 2016.

[4] G. M. Abaido, "Cyberbullying on Social Media Platforms Among University Students in the United Arab Emirates," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 407-420, 2020.

[5] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," *Proc. ACM Conference Web Science (WebSci)*, pp. 13-22, 2017.

[6] E. Raisi and B. Huang, "Cyberbullying Detection with Weakly Supervised Machine Learning," *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 409-416, Jul. 2017.

[7] E. Raisi and B. Huang, "Weakly Supervised Cyberbullying Detection using Co-Trained Ensembles of Embedding Models," *Proceedings IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 479-486, Aug. 2018.

[8] Vijay B, Jui T, Pooja G, Pallavi V, "Detection of Cyberbullying using Deep Neural Network," *in 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp.604-607, 2019

[9] Monirah A., Mourad Y, "Optimized Twitter Cyberbullying Detection based on Deep Learning," *in 21st Saudi Computer Society National Computer Conference (NCC)*, 2018.

[10] Batoul H, Maroun C, Fadi Y, "Cyberbullying Detection: A Survey on Multilingual Techniques" *in European Modelling Symposium (EMS)*, pp. 165–171, 2016.

[11] Ms. Anushree Negi, "A Brief Survey on Text Mining, Its Techniques, and Applications," *SSRG International Journal of Mobile Computing and Application*, vol. 8, no. 1, pp. 1-6, 2021. Crossref, https://doi.org/10.14445/23939141/IJMCA-V8I1P101.

[12] Xiang Z, Jonathan T, Nishant V, Elizabeth W., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," *in 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 740-745, 2016.

[13] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, "BullyNet: Unmasking Cyberbullies on Social Networks," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 332–344, 2021, doi:10.1109/TCSS.2021.3049232.

[14] R. Zhao and K. Mao, "Cyberbullying Detection Based on Semantic Enhanced Marginalized Denoising Auto-Encoder," *IEEE*

Transactions on Affective Computing, vol. 8, no. 3, pp. 328–339, 2017, doi:10.1109/TAFFC.2016.2531682.

[15] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro and L. Coheur, "Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks," *2018 IEEE International Conference on Fuzzy Systems (FUZZIEEE)*, pp. 1-7, 2018. doi: 10.1109/FUZZIEEE.2018.8491557.

[16] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, Et Al., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," *Proc. 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 740-745, Dec. 2016.

[17] Johnson, Elizabeth K, and Michael D Tyler, "Testing the Limits of Statistical Learning for Word Segmentation," *Developmental science*, vol. 13, no.2 , pp.339-45, 2010. doi:10.1111/j.1467-7687.2009.00886.x

[18] Gunisetti Tirupathi Rao, "Dr. Rajendra Gupta, " An approach of Clustering and Analysis of Unstructured Data," *SSRG International Journal of Computer Science and Engineering*, vol. 6, no. 11, pp. 64-69, 2019. Crossref, https://doi.org/10.14445/23488387/IJCSE-V6I11P114

[19] I. Nazar, D.-S. Zois and M. Yao, "A Hierarchical Approach for Timely Cyberbullying Detection," *Proc. IEEE Data Science.Workshop (DSW)*, pp. 190-195, 2019.

[20] Y. Win, "Classification Using Support Vector Machine to Detect Cyberbullying in Social Media for Myanmar Language," *2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pp. 122-125, 2019. doi: 10.1109/ICCE-Asia46551.2019.8942212.

[21] Lin L., Linlong X., Nanzhi W., GuocaiY, "Text Classification Method Based on Convolution Neural Network," *in 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp . 1985-1989, 2017.

[22] Sang-Uk Jung, Jungho Byun, Seongyeol Bae, Donghwi Song, "Predicting Neologisms for Marketing: a Text Mining Approach," *SSRG International Journal of Economics and Management Studies*, vol. 7, no. 7, pp. 5-9, 2020. Crossref, https://doi.org/10.14445/23939125/IJEMS-V7I7P102.

[23] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, 2008.

[24] R. Johnson and T. Zhang, "Effective Use of Word Order for Text Categorization With Convolutional Neural Networks," arXivPrepr. arXiv1412.1058, 2014.

[25] X. Zhang, J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification," I*n Advances in Neural Information Processing Systems*, 2015, pp. 649–657.

[26] Dr.K.Anuradha, Dr.M.Vamsi Krishna, Dr.Banitamani Mallik, Prof.B.P.Mishra. Festus, "A Survey Paper on Sentiment Analysis: Approaches, Methods Challenges," *International Journal of Computer Trends and Technology*, vol. 67, no.10, pp.25-34, 2010. DOI: 10.14445/22312803/IJCTT-V67I10P105.

[27] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a Large-Scale Corpus for Evaluating Event Detection on Twitter," *In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 2013, pp. 409–418.

[28] S. Salawu, Y. He and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," *IEEE Transactions on Affective Computing* , vol. 11, no. 1, pp. 3-24, Jan. 2020.

[29] A. Onan, "Sentiment Analysis on Product Reviews Based on Weighted Word Embeddings and Deep Neural Networks," *Concurrency and Computation: Practice and Experience*, pp. e5909, 2020.