An Ensemble Approach for Cyber Bullying Text messages and Images

Zarapala Sunitha Bai[1], Sreelatha Malempati[2],
[1] Department of Computer Science and Engineering, R.V.R. and J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India, Research Scholar,Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
[2] Department of Computer Science and Engineering, R.V.R. and J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India
Corresponding Author: zsunithabai@gmail.com

Abstract
Text mining (TM) is most widely used to find patterns from various text documents. Cyber-bullying is the term that is used to abuse a person online or offline platform. Nowadays cyber-bullying becomes more dangerous to people who are using social networking sites (SNS). Cyber-bullying is of many types such as text messaging, morphed images, morphed videos, etc. It is a very difficult task to prevent this type of abuse of the person in online SNS. Finding accurate text mining patterns gives better results in detecting cyber-bullying on any platform. Cyber-bullying is developed with the online SNS to send defamatory statements or orally bully other persons or by using the online platform to abuse in front of SNS users. Deep Learning (DL) is one of the significant domains which are used to extract and learn the quality features dynamically from the low-level text inclusions. In this scenario, Convolutional neural networks (CNN) are used for training the text data, images, and videos. CNN is a very powerful approach to training on these types of data and achieved better text classification. In this paper, an Ensemble model is introduced with the integration of Term Frequency (TF)-Inverse document frequency (IDF) and Deep Neural Network (DNN) with advanced feature-extracting techniques to classify the bullying text, images, and videos. The proposed approach also focused on reducing the training time and memory usage which helps the classification improvement.

Keywords:
Convolutional neural networks (CNN), Text mining (TM), Term Frequency (TF)-Inverse document frequency (IDF), Deep Neural Network (DNN).

1. Introduction

The usage of social networking sites (SNS) is increasing rapidly every day. SNS is a platform that gives huge opportunities and communication to people belonging to several fields. People may discuss various issues that are more popular using this platform. In the SNS platform, cyber-bullying is one of the significant issues in the present situation. Cyber-bullying is increasing day by day by using several types of messages and images. In 2021, 77.96% of SNS users feel bad about cyber-bullying [1]. 95% of people are accepted that they are witnesses of some kind of cyber-bullying occurring online. So, this is the time to stop cyber-bullying [2]. Cyber-bullying is of many types such as abusing the person using an SNS platform with comments, personal messages; morphed images, etc are used to abuse the person in SNS and also in other platforms. This has become a more complicated issue and creates a lot of issues in my personal life. Many SNS providers are trying to solve this issue by blocking users based on their behaviour. Still, this is an unsolved issue in SNS. Text classification is a domain that belongs to various fields. This can be used to solve the various misclassification issues present in this domain.

Text classification is mainly based on the extraction of features by removing the noise from the given text inputs called words, sentences, phrases, etc [3]. It is very important to find the patterns that belong to a specific language such as English. To classify the different types of text messages various feature extraction methods are used. Classifying the bullying messages gives a huge impact on using these feature extraction methods. This paper mainly focused on finding the bullying content from text messages and images present in the SNS. Sentiment analysis (SA) is one of the significant tasks in finding the sentiments from the user messages or tweets in online SNS. Various feature extraction models are used to extract the text and image features to analyze the sentiments [4]. These techniques improve the classification of sentiments present in the dataset. Online SNS are platforms for attackers to attack the victim with message bullying and image bullying. Machine Learning (ML) is most widely used to detect language and Images automatically and prevents these attacks [5]. Many researchers are trying to develop an automated cyber-bullying model to detect and prevent this type of message [6]. Parts of Speech (POS) is most widely used to find the features that belong to polarity [7]. In [8], the author

Manuscript received November 5, 2023
Manuscript revised November 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.11.7

developed cyber-bully detection (CBD) by using SA and emojis. Emoji is an expression-based image used to express a person's emotions.

In this paper, the deep neural networks (DNN) model Deep Belief Networks (DBN) is used to analyze the tweets data and image data for the classification of cyber-bullying. The proposed model is the integration of pre-processing techniques, tokenization, feature extraction techniques, and DBN. DBN is applied to two real-time datasets to analyze the performance. Figure 1 shows the overall system architecture.

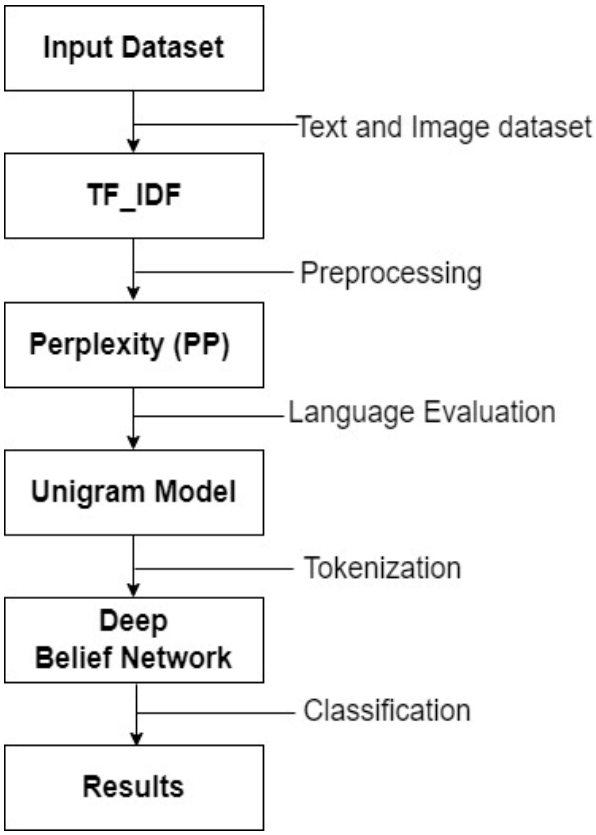


Figure 1: Architecture Diagram

This paper is organized as follows. The next section describes the methodology we followed int his paper. Section 3 discusses the companies' survey results. In section 4, we discuss the results of the students' survey. Section 5 summarizes the results and gives some recommendations. Finally, in section 6, we give some concluding remarks.

2. Literature Survey

Krizhevsky et al. [9] developed a DCNN model to find the objects in images. Object recognition in images is done by using the ImageNet dataset. This approach extracts better features from object detection. M. Anand et al., [10] proposed the long short term memory cell (LSTM) with and without intgeration of word GloVe embeddings to find the abused comments and store the websites that are circulating this types of messages and prevent this websites and improves the safety discussions in online platforms. In this paper, the kaggle dataset is used to find the several types of toxic comments such as toxic, severe toxic, obscene, threat, insult, and identity hate. By using the Dl trained models the comments are classified.

Y. Li et al., [11] proposed the text mining approach which is used to classify the text messages by using the term-based technique. In the existing methods various issues are identified such as polysemy and synonymy. From the past many years the pattern based methods perform better than the term-based methods. These methods cannot work on large datasets and these remains as the huge issue in text mining.

Chikashi Nobata et al., [12] developed the ML-based approach which is utilized to detect the hate speech collected from user comments present online by using two domains. A dynamic corpus consists of user comments annotated for abusive language. Yoon Kim [13] proposed the CNN model is used to train the vectors for the classification of sentence-level tasks. The combination of several DL, models gives extreme outputs on multiple datasets. M.O.Ibrohim et al., [14] proposed the integrated model by using the word embedding (word2vec) feature, and its combinations with part of speech and/or emoji were used to identify hate speech and abusive language on Twitter in the Indonesian language. A few integrated models that are unigram with part of speech and/or emojis were also utilized during the experiment and the results were studied. The classification algorithms used in this study were Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). The combination of unigram features, part of speech, and emoji obtained the highest accuracy value of 79.85% with an F-Measure of 87.51%. Z. Waseem et al., [15] introduced the method that finds

hate speech from the publicly available corpus consisting of 16k tweets. To improve hate speech detection the extra-linguistic features with the integration of character n-grams detect accurate hate speech. F. D. Vigna et al., [16] proposed the alert-based approach which detects hate speech in SNS. This approach mainly focused on finding personal abuse, caste abuse, and religious abuse based on the text. This approach combined with the SVM and LSTM to classify the hate speech words and also by speech recognition. The two classification approaches give accurate hate speech recognition. H. Yenala et al., [17] proposed the novel DL approach that automatically finds the irrelevant language. This is used to solve several issues in finding irrelevant language. Irrelevant language means spelling mistakes and variations present in the language. The proposed approach is called Convolutional Bi-Directional LSTM (C-BiLSTM) which is combined with CNN and BLSTM. BLSTM is used to filter the irrelevant language and CNN is used to extract the significant features present in the given dataset. Thus the C-BiLSTM obtained better accuracy compared with the existing models.

M. M. Islam et al., [18] developed an effective approach that is used to detect bullying messages online. This approach is merged with NLP and ML approaches. This combination of BoW and TF-IDF achieved better accuracy compared with existing ML algorithms. A. Shekhar et al., [19] proposed a novel technique that is used to detect cyberbullying with the help of the Bag-of-Phonetic-Codes model. Based on the pronunciation, the wrong-spelled and abused words are removed. The proposed approach used the BoW model to extract the textual features. The Soundex algorithm is used for the phonetic code creation in this model. Experiments show that the proposed approach obtained the accurate detection of cyber-bullying detection. C. Sharma et al., [20] detect cyberbullying based on the meaning of the text. This is also used to reduce the spreading of hurtful messages over the internet. Adopting the features of NLP with ML gives better performance. A. Wadhwani et al., [21] developed a new model that solves various issues such as mismatched bullying and irrelevant content detection. This paper mainly focused on detecting the Injurious comments that trouble online users in SNS. The proposed DNN model finds the patterns of the input message and analyses the type of the messages based on the metrics such as toxic, hate, serious toxic, threat, etc. C. S. Wu et al., [22] introduced the classification based on the videos that are normal or hate videos. By using the online crawler the video dataset is collected. The audio is extracted from the videos and converted the audio into textual format by using an online Speech-to-Text converter to get the transcript of the videos.

Murshed et al., [23] proposed the DEA_RNN model that detects cyber-bullying messages in online SNS. The approach applied to 10k tweets data to analyze the cyber-bullying text. The proposed approach DEA_RNN obtained the good results compared with existing models such as Bi-LSTM, RNN, SVM and MND, RF. The accuracy is up to 91.54% and 90.67% precision, 89.89% recall, 90.21% F1-score, and 91.84% specificity. Zarapala Sunitha Bai et al., [24] proposed the ETMA approach used to detect cyber-bullying messages in real-time applications such as Twitter. This classifies the text into bullying and non-bullying messages available in datasets. This is the combination of the TF-IDF and CNN model. This gives accurate results for the classification of cyber-bullying messages.

3. Methodology

This research mainly focused on detecting cyber bullying using images and videos. In the literature survey, a lot of existing methods are discussed to detect cyber bullying messages and images. A lot of drawbacks are identified with the existing approaches. In this paper, the methodology is combined with various techniques such as data pre-processing with Term Frequency-Inverse Document Frequency (TF-IDF), and Deep Neural Network (DNN) is used for the detection of cyberbullying words. By applying the confusion matrix the performance of the DNN is shown in this paper.

4. Data Pre-processing with Term Frequency - Inverse Document Frequency (TF-IDF)

Team frequency (TF) and Inverse document frequency (IDF) is the feature extraction technique that is used to extract the features present in the

dataset. TF-IDF is the statistical technique in NLP and information retrieval (IF). TF-IDF is used to extract the bullying words present in the dataset. This calculates the importance of bullying terms within every review and dataset. Text vectorization is the process utilized to transform the words in the given review. Of the many vectorization techniques, TF-IDF is one of the significant approaches.

Term Frequency (TF): The TF of word in document. By using several steps the raw count of the word (bullying words) is obtained in the document. Here document means dataset. Thus the frequency is adjusted based on the dataset instances or the frequency of words in the dataset.

Inverse Document Frequency (IDF) is measured for a particular word which is equal to overall messages present in the dataset, divided by overall reviews that consist of particular work. IDF is measured as:
The scores of TF-IDF are measured by multiplying.

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}*\mathrm{IDF} \quad (1)$$

$$\mathrm{TF}(\mathrm{t},\mathrm{d})=\log\big(1+\mathrm{freq}(\mathrm{t},\mathrm{dt})\big) \quad (2)$$

$$\mathrm{IDF}(\mathrm{t},\mathrm{D})=\log\left(\frac{\mathrm{N}}{\mathrm{count}\,(\mathrm{d}\in \mathrm{D}:\mathrm{t}\in \mathrm{d})}\right) \quad (3)$$

Perplexity (PP): In this paper, to analyze the given text dataset the perplexity is used to evaluate the language model perplexity. This approach gives the inverse probability of test set, the no of words are normalized.

$$\mathrm{PP}(\mathrm{Words})=\mathrm{P}(\mathrm{w}_1,\mathrm{w}_2\ldots\ldots\mathrm{w}_\mathrm{N})^{-\frac{1}{\mathrm{N}}} \quad (4)$$

$$=\sqrt[N]{\frac{1}{P(w_1,w_2\ldots.w_N)}}$$

4.1 Unigram model

In this model, tokenization is used to process the raw text into small words. The input sentences break the text into words that are called tokens. These tokens are used to help the scenario and meaning of the sentences. In tokenization, the unigram model is used to consider every token. The probability of token X is given as the past scenario is the probability of token X. If the unigram model generates the text, this will always predict the general tokens.

$$\mathrm{P}(\mathrm{a})=\prod_{\mathrm{x}=1}^{\mathrm{M}}\mathrm{p}(\mathrm{a}_\mathrm{x}) \quad (5)$$

$$\forall \mathrm{x}\ \mathrm{a}_\mathrm{x}\epsilon\,\mathcal{V},\sum_{\mathrm{a}\epsilon\mathrm{v}}\mathrm{p}(\mathrm{a})=1 \quad (6)$$

a: Sentence
a: sub-word forming sentence
V: Vocabulary

4.2 Deep Neural Network (DNN)

In this paper, the DNN model is introduced to find the accurate cyber-bullying in online SNS and other online images. The proposed approach is applied to two datasets a Twitter dataset and an online synthetic dataset consisting of online bullying images. In DNN various connected components are called nodes. In DNN, nodes are very small and act as the neurons in the human brain. The neuron starts the process of the signal received by the neuron. Based on the input received the signal is transferred from one neuron to another neuron. From this feedback, a complex network is created. Here, the input is a text message (reviews or tweets) or bullying images and nodes will process these data. In DNN this is the general process for every dataset. Deep Belief Networks (DBN) is the other algorithm which is used in this paper for the processing of complex datasets such as twitter data and image data. DBN is one of the best algorithms in DNN which is used to process the datasets.

4.3 Deep Belief Networks (DBN)

DBN is creative algorithm consists of stacked RBMs. DBN follows the hierarchical representation of input text dataset and image dataset. Hinton et al., [25] introduced the DBN algorithm that trains the single layer at a time. Every layer processes the input text data and image data. In this x is the visible component and ℓ is the hidden layer with joint distribution [26].

$$\mathrm{p}(\mathrm{x},\mathrm{h}^1,\ldots.\mathrm{h}^\ell)$$

$$=\mathrm{p}\big(\mathrm{h}^{\ell-1},\mathrm{h}^\ell\big)\Big(\prod_{\mathrm{k}=1}^{\ell-2}\mathrm{p}\big(\mathrm{h}^\mathrm{k}\big|\mathrm{h}^{\mathrm{k}+1}\big)\mathrm{p}(\mathrm{x}|\mathrm{h}^1) \quad (7)$$

Hence, every layer of DBN is created as RBM; DBN training is as same as RBM.
By using the DBN training, the classification is initialized for the given dataset. Two stage parallel training is started such as: 1) learning of stacked

RBM in layer based manner, 2) deep tuning classifier for supervised learning. An optimization issue is solved at every stage. Training the dataset $D = \{(a^{(1)}, b^{(1)}), \dots\dots, (a^{(|D|)}, b^{(|D|)})\}$ with a as input and b as label, the optimization issue is solved in pre-trained phase at every layer k,

$$\min_{\theta_k} \frac{1}{|D|} \sum_{i=1}^{D} \left[-\log p\left(a_k^{(i)}; \theta_k\right) \right] \qquad (8)$$

The parameters in RBM models represent the following metrics such as $\theta_k = (W_k, b_k, c_k)$. a_k^i is the visible layer k which is input $x^{(i)}$. The layers are updated in step by step wise and solve the ℓ issues from last to first hidden layer. For better filtering or tuning phase the following optimization problem is solved:

$$\min_{\phi} \frac{1}{|D|} \sum_{i=1}^{D} [\mathcal{L}(\emptyset; y^{(i)}, h(x^{(i)}))] \qquad (9)$$

Where $\mathcal{L}()$ represents loss function, at layer ℓ the hidden features are represented, $\emptyset$ represents the metrics of the classifier. This is written as $h(x^{(i)}) = h(x_l^{(i)})$. Thus this can be used to classify the text data and image data.

4. Dataset Description

The proposed approach performance is analyzed by using the two real time dataset such as twitter dataset and image based cyber-bullying datasets. The dataset consists of 17k bullying and non-bullying messages (Sentences). As per the dataset description there are 6135 are bullying messages, 7235 non-bullying messages and 2630 normal messages. The proposed algorithm is applied on these twitter dataset. 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.

The second Image based cyber-bullying dataset consists of 1500 training images and 1500 testing images collected from various online sources. The images contain several types of cyber-bullying images such as personal abuse, threatening, image morphing etc.

5. Dataset Description

The proposed approach performance is analyzed by using the two real time dataset such as twitter dataset and image based cyber-bullying datasets. The dataset consists of 17k bullying and non-bullying messages (Sentences). As per the dataset description there are 6135 are bullying messages, 7235 non-bullying messages and 2630 normal messages. The proposed algorithm is applied on these twitter dataset. 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.

The second Image based cyber-bullying dataset consists of 1500 training images and 1500 testing images collected from various online sources. The images contain several types of cyber-bullying images such as personal abuse, threatening, image morphing etc.

Table 1: Twitter dataset Description

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

Table 2: Image dataset

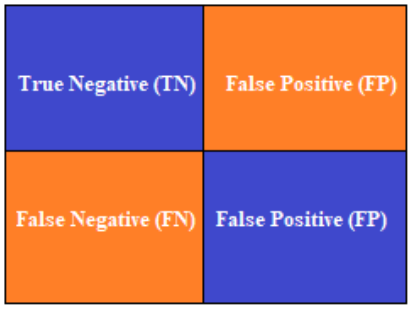
Message Type	Training	Testing
Bullying	1500	1500
Non-Bullying	500	500
Normal Messages	1k	1k

Table 1 and table 2 shows the total no of data belongs to training and testing.

6. Performance Metrics

The performance of the model is analyzed by using the confusion matrix. This will specify the performance of classification models for given test data. This will specify the values for test data that are known. This matrix is divided into two attributes

such as predicted values and original values along with an overall number of predictions.



True Negative (TN): The prediction value is false and actual value is also false.
True Positive (TP): The prediction value is true and actual value is false.
False Positive (FP): The predicted value is true and actual value is false.
False Negative (FN): The predicted value is false and actual value is true.

Precision: This is specified that the total number of correct results obtained by the proposed model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

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

$$\text{F1 Measure} = 2 \times \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

Accuracy: This parameter plays the major role in showing the overall accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

Recall: This metric is mainly focused on reducing the false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{No. of TP} + \text{No. of FN}} \quad (8)$$

Table 3: Performance of Existing and Proposed Algorithms applied on twitter dataset

	SVM	CNN	TF-IDF+CNN	TF-IDF+DNN
Precision	78.98%	82.12%	89.89%	95.56%
F1-Measure	80.12%	84.32%	90.12%	96.56%
Accuracy	81.23%	85.12%	92.32%	96.12%
Recall	82.34%	87.12%	94.12%	96.67%

Table 3 shows the comparison between SVM, CNN, TF-IDF-CNN and TF-IDF-DNN. Among all the approaches the proposed approach TF-IDF+CNN achieved the high performance for classifying the twitter data.

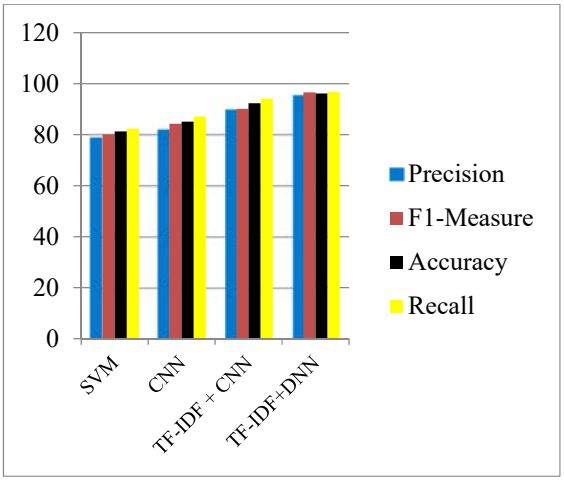


Figure 2: Comparison Graph between Existing and Proposed Algorithms

Table 4 shows the comparison between SVM, CNN, TF-IDF-CNN and TF-IDF-DNN. Among all the approaches the proposed approach TF-IDF+CNN achieved the high performance for classifying the Images data.

Table 4: Performance of Existing and Proposed Algorithms applied on Image dataset

	SVM	CNN	TF-IDF+CNN	TF-IDF+DNN
Precision	77.12%	83.23%	90.12%	96.87%
F1-Measure	81.52%	85.67%	91.23%	97.45%
Accuracy	82.53%	86.12%	91.32%	98.56%
Recall	83.44%	87.12%	94.34%	98.23%

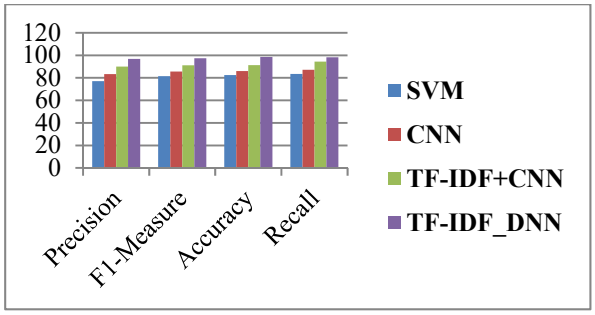


Figure 3: Comparison Graph between Existing and Proposed Algorithms

7. Conclusion

In this paper, the proposed DNN model gives the accurate classification of bullying words and images. The proposed approach TF-IDF+DNN works better on online SNS to detect and find cyber-bullying words and Images. These types of Images and text create a lot of issues for the victim. An efficient training model, pre-processing model, and word embedding methods make the system novel. The system proves to be useful for the analysis of the cyber-bullying rates on different social media platforms so that relative precautions and actions can be taken to decrease the cyber-bullying rate. The performance of proposed approach TF-IDF+DNN achieved a precision of 96.87%, F1-measure 97.45%, accuracy 98.56% and recall 98.23%.

Acknowledgment

This work is done under the grant received (6/41) by Deanship of research at Islamic University of Madinah (IUM) for research that studies the economic effect of COvid-19 pandemic. We also give special thanks to the administration of IUM for their support in every aspect.

References

[1] Prusa, J.D., Khoshgoftaar, T.M. Improving deep neural network design with new text data representations. J Big Data 4, 7 (2017). https://doi.org/10.1186/s40537-017-0065-8.

[2] S. Ozcan, A. Homayounfard, C. Simms and J. Wasim, "Technology Roadmapping Using Text Mining: A Foresight Study for the Retail Industry," in IEEE Transactions on Engineering Management, vol. 69, no. 1, pp. 228-244, Feb. 2022, doi: 10.1109/TEM.2021.3068310.

[3] Zhang X, LeCun Y. Text understanding from scratch. 2015. arXiv preprint arXiv:1502.01710.

[4] Prusa JD, Khoshgoftaar TM, Dittman DJ. Impact of feature selection techniques for tweet sentiment classification. In: Proceedings of the 28th International FLAIRS Conference; 2015. p. 299–304.

[5] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer and Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning", International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.01005

[6] Prusa JD, Khoshgoftaar TM, Dittman DJ. Impact of feature selection techniques for tweet sentiment classification. In: Proceedings of the 28th International FLAIRS Conference; 2015. p. 299–304.

[7] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media. Association for Computational Linguistics; 2011. p. 30–8.

[8] K. Maity, S. Saha and P. Bhattacharyya, "Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish," in IEEE Transactions on Computational Social Systems, 2022, doi: 10.1109/TCSS.2022.3183046.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Info Proc Syst. 2012: 1097–105.

[10] M. Anand and R. Eswari, "Classification of Abusive Comments in Social Media using Deep Learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 974-977, doi: 10.1109/ICCMC.2019.8819734.

[11] Y. Li, A. Algarni, M. Albathan, Y. Shen and M. A. Bijaksana, "Relevance Feature Discovery for Text Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1656-1669, 1 June 2015, doi: 10.1109/TKDE.2014.2373357.

[12] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad and Yi Chang,

"Abusive Language Detection in Online User Content", Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, pp. 145-153, 2016.

[13] Yoon Kim, "Convolutional neural networks for sentence classification", EMNLP. Association for Computational Linguistics, pp. 1746-1751, 2014.

[14] M.O.Ibrohim, Muhammad Akbar Setiadi, and Indra Budi. 2019. Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. In Proceedings of the International Conference on Advanced Information Science and System (AISS '19). Association for Computing Machinery, New York, NY, USA, Article 18, 1–5.

[15] Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, 88--93.

[16] F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Esconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), 86--95.

[17] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal. 2017. Deep learning for detecting inappropriate content in text. In International Journal of Data Science and Analytics.

[18] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411601.

[19] A. Shekhar and M. Venkatesan, "A Bag-of-Phonetic-Codes Modelfor Cyber-Bullying Detection in Twitter," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-7.

[20] C. Sharma, R. Ramakrishnan, A. Pendse, P. Chimurkar and K. T. Talele, "Cyber-Bullying Detection Via Text Mining and Machine Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6.

[21] A. Wadhwani, P. Jain and S. Sahu, "Injurious Comment Detection and Removal utilizing Neural Network," 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), 2021, pp. 165-168.

[22] C. S. Wu and U. Bhandary, "Detection of Hate Speech in Videos Using Machine Learning," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020, pp. 585-590.

[23] Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," in IEEE Access, vol. 10, pp. 25857-25871, 2022.

[24] Zarapala Sunitha Bai, Sreelatha Malempati, "An Enhanced Text Mining Approach using Ensemble Algorithm for Detecting Cyber Bullying" International Journal of Engineering Trends and Technology, vol. 70, no. 9, pp. 393-399, 2022.

[25] Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[26] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets.,” Neural computation, vol. 18, no. 7, pp. 1527–54, 2006.