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

Z.Sunitha Bai * , Sreelatha Malempati

lathamoturi@rediffmail.com

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

zsunithabai@gmail.com *Corresponding author

Department of Computer Science and Engineering, R.V.R. and J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

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 cyber bullying on social media messages. Cyber-bullying is the type of abusing someone with the insulting language. Personal abusing, sexual harassment, other types of abusing come under cyber-bullying. Several existing systems are developed to detect the bullying words based on their situation in the social networking sites (SNS). SNS becomes platform for bully 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 on synthetic dataset collected from various data a source which consists of 5k messages belongs to bullying and non-bullying. The performance is analyzed by showing Precision, Recall, F1-Score and Accuracy.

Keywords:

Enhanced Text Mining Approach, Ensemble Algorithm, Detecting Cyber Bullying

1. Introduction

Web 2.0 is most widely used to improve the user-created platforms for social networking sites (SNS) users [1]. TM is sub-domain in data mining (DM) to mine the accurate patterns. TM is most widely used to extract the patterns from the 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 the interesting and required information that is useful in various applications. Every day huge data is generated in social media. Social networking sites (SNS) are most widely used to communicate with various types of users. In [2], various ML algorithms are discussed about the challenges that are facing in the detection of cyber bullying. By using ML algorithms the prediction becomes a more accurate behaviour of the human [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. Cyber bullying becomes more complicated in SNS because personal abuse becomes more complicated on social media platforms. Detecting the bullying messages and preventing this message is more complex to the SNS developers.

In this paper, An Enhanced text mining approach is developed by using Ensemble Algorithm (ETMA) to analyze the various messages that are collected from Twitter data. This dataset is a synthetic dataset that consists of 16k Twitter messages with 7 attributes. The ETMA follows the powerful pre-processing after initializing the dataset. Training, feature extraction, and applying the algorithm on the dataset. The performance is analyzed by showing the parameters such as precision, recall, accuracy, F1-measure, and duration.

Manuscript received May 5, 2023
Manuscript revised May 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.5.1

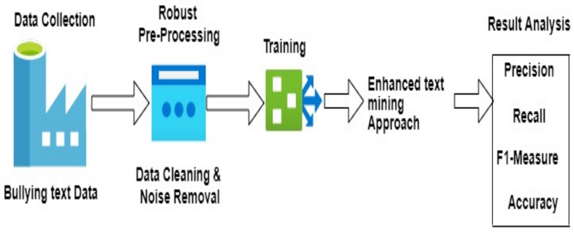

Figure 1: Processing of Enhanced Text Mining Approach

2. Literature Survey

G. M. Abaido et al., [4] proposed the approach that detects cyber-bullying among Arab community students. The data is collected from 200 students belonging to UAE. 91% of analysis shows that cyber-bullying 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 the 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 by using significant properties. The proposed approach is most widely used in various applications such as Twitter, Ask.fm, Instagram data, etc.

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

Monirah A et al., [9], it was 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, word introduction was delivered for each word using (GloVe) procedure. They came up with a game plan of word embedding that was subsequently supported by CNN computation for portrayal.

Batoul H et al., [10] discussed several approaches that are used to detect cyber-bullying messages in the Arabic language. The authors in this paper made their research on multilingual cyber-bullying detection and finally proposed the solution for the issue of Arabic cyber-bullying.

Xiang Z et al., [11] proposed the novel pronunciation-based convolutional neural network (PCNN) to solve the issues in cyber-bullying. The proposed approach corrects the spelling mistakes and pronunciation of the given bullying text by solving the noise and bullying data sparsity. To solve these issues various integrated approaches are included in this and achieve better results in this.

3. Dataset Description

The dataset consists of 16k 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.

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

TF-IDF [12] is one of the mathematical approaches that counts the relevant word in specific document and also used in many document. Based on the two parameters, the count of words that occur in document and IDF counts the number of relevant words in group of documents. Here every document is considered as message in dataset. TF is used to measure the high term frequency value based on the more 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 for the better classification of text [13]. IDF is extracted from document frequency (DF) which shows the total number of term occur in total number of documents [14].

$$\mathrm{IDF}\ (\mathrm{t}, \mathrm{d}, \mathrm{D}) = \log \frac{|\mathrm{D}|}{\mathrm{DF}(\mathrm{t}, \mathrm{D})} \qquad (1)$$

In (2), DF shows the term‘t’ in corpus ‘D’. Symbol in Equ-2 represents the overall tweets in corpus D. To prevent the abnormal cases, the formula is given as:

$$\mathrm{IDF}\ (\mathrm{t}, \mathrm{d}, \mathrm{D}) = \log \frac{|\mathrm{D}|+1}{\mathrm{DF}(\mathrm{t},\mathrm{D})+1} \qquad (2)$$

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

$$\mathrm{TF} - \mathrm{IDF}(\mathrm{t}, \mathrm{d}, \mathrm{D}) = \mathrm{TF}(\mathrm{t}, \mathrm{d}) * \mathrm{IDF}(\mathrm{t}, \mathrm{d}, \mathrm{D}) \qquad (3)$$

In equation (3), initializes the t as tweet d in corpus D, and TF value of term t is present in document d. Finally, this approach measures the total number of stop words, bad words and total number of word count.

5. 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 convolutional neural network (CNN) algorithm, which is called ETMA, and this can solve the complex issues. The main aim of this approach is to develop the efficient detection of cyber-bullying based on accurate meanings and reduce the computational time and cost. By using the CNN, efficient classification of bullying words is done [15] [16] [17]. 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 which is followed by feature selection. The proposed approach in this paper starts with robust pre-processing and effective training with VGG-16 which is called a pre-trained model to get better results.

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

Example: realdonaldtrump you are the man donald trump donâ€™t listen to anyone else ever follow your own instincts and godgiven ability thank god for donald trump much love.

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

6. CNN

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 layer and dense layer are is applied and it is called as a final dense layer.

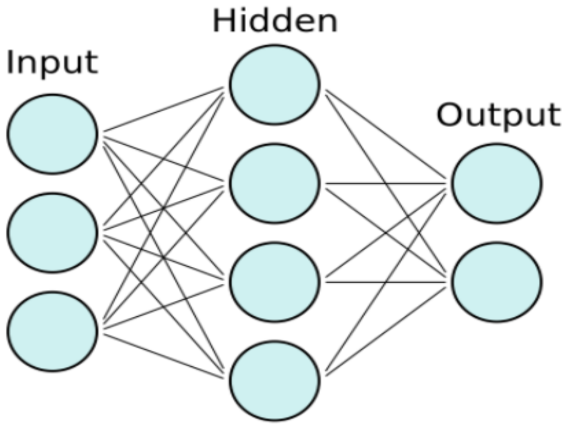

Figure 2: CNN Architecture

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. In this paper, the shape is one-dimensional and consists 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.

Convolutional Layer: This layer mainly focused on performing the convolution operation to extract the new matrix with convolved features. In matrix the filter is slid (convolve operation) over 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 at the time of training. With this operation a new matrix is created that consists of convolved features and these are sent to the next layer for analyzing.

Fully Connected Layer - In this layer, the main function is fully_con which takes the previous layer output as the input and predict the number of classes such as Aggression, Attack, Badwords, Racism, Sexism, Toxicity and the activation function is called as softmax function. In this layer, a regression layer is present, and all the previous layers are combined and measure the classification sentences.

$$\mathrm{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \qquad (4)$$

This is considered as 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, Toxicity. Based on the given tweet the softmax function measures the scores of the tweet and assigns the values for each type of tweet and it is considered as bully text. The overall results are analyzed by using a confusion matrix and analyzing the performance of the proposed approach.

7. Experimental Results

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

Precision: This is one of the significant metric that shows the correct positive predictions based on the type of twitter.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (5)$$

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

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

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

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

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

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

Table 1: 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

Table 1 shows the comparison between SVM, CNN and TF-IDF-CNN. Among all the approaches the proposed approach TF-IDF+CNN achieved the high performance for classifying the twitter data.

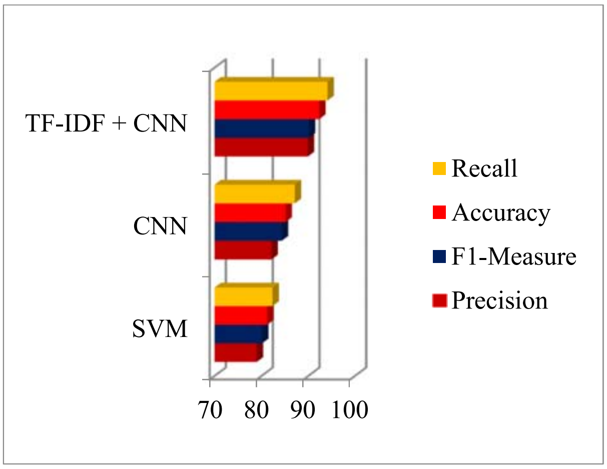

Figure 2: Comparison Graph between Existing and Proposed Algorithms

8. Conclusion

Cyber-bullying becomes more complicated for OSNS users. Cyber-bullying is also considered as attacks personally which is mainly occur in OSNS by 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, Toxicity. The text messages that are given by various types of users indicate whether cyber-bullying is occurring or not. The ETMA in this paper showed the huge performance compared with existing approaches such as SVM and CNN. The proposed approach ESTM is the combination of TF-IDF and CNN achieved the performance of Precision-89.89%, F1-Measure-90.12%, Accuracy-92.32%, Recall-94.12%, Duration (Sec)-58.34. The performance is increased by 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", Comput. Hum. Behav., vol. 63, pp. 433-443, Oct. 2016.

[4] G. M. Abaido, "Cyberbullying on social media platforms among university students in the united arab emirates", Int. J. Adolescence Youth, vol. 25, no. 1, pp. 407-420, Dec. 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 Conf. Web Sci. (WebSci), pp. 13-22, 2017.

[6] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning", Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, pp. 409-416, Jul. 2017.

[7] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models", Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. 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] 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.

[12] 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.

[13] A. Onan, “Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks,” Concurrency and Computation: Practice and Experience, p. e5909, 2020.

[14] 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.

[15] R. Johnson and T. Zhang, “Effective use of word order for text categorization with convolutional neural networks,” arXivPrepr. arXiv1412.1058, 2014.

[16] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in Advances in neural information processing systems, 2015, pp. 649–657.

[17] 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.