Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

A Survey on Cyberbullying Classification and Detection

S. Venkatesh Perumal[1], J.C. Miraclin Joyce Pamila[2]

[1]PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, India

[2]Department of Computer Science and Engineering, Government College of Technology, Coimbatore, India

E-mail: [1]venk.71772177107@gct.ac.in, [2]miraclin@gct.ac.in

Abstract

Social media platforms have seen an increase in the prevalence of cyberbullying. Protecting social media platforms against cyberbullying is essential as social media is extensively used among people of all ages. Events of cyberbullying have been rising, especially among young individuals who spend most of their time switching between various social media sites. This study gives an overview of the existing research on the categorization and detection of cyberbullying using several methods from the Deep learning and Machine learning field like Convolutional Neural Network, Recurrent Neural Network, Long Short -Term Memory, Gated Recurrent Unit, Bi-GRU-Attention-CapsNet, Support Vector Machine, Random Forest, Naive Bayes, and k-Nearest Neighbor, along with the study that examines the effects of various feature extraction techniques like Term frequency and Inverse Document Frequency, Information Gain, Dolphin Echolocation Algorithm, and Improved Dolphin Echolocation Algorithm.

Keywords: Social media, Cyberbullying detection, Machine Learning, Deep Learning

1. Introduction

In general, lives are made easier by digital technology, which also supports government, business, and many other areas. Cyberbullying is a poor outcome of internet use, as is true for most things. But, with everything good, comes something bad. Cyberbullying is bullying that

DOI: https://doi.org/10.36548/jitdw.2023.2.001
Received: 25-03-2023. received in revised form: 25-04-2023, accepted: 09-05-2023, published: 19-05-2023

takes place in a digital environment, such as on a computer, a tablet, a smartphone, social media, online games, and other places. Bullying is still a concern for the majority of people in 2023. 73 percent of students believe they have experienced bullying, and 44 percent report that it has occurred within the past 30 days [1]. About 60% of teenagers have been subjected to cyberbullying [2]. Because it can occur at any time, one could argue that cyberbullying is more harmful than conventional in-person bullying. Cyberbullying is harder to detect since the words bullies use are in the digital sphere, which means it frequently goes unreported.

Text preprocessing, feature engineering, and classification are the next three steps in the methodology for classifying text. Text preprocessing is the process of converting text into a comprehensible and consistent format before feeding it into a model for additional analysis and learning. Some of the preprocessing techniques include lower case conversion, punctuation removal, number removal, stop word removal, stemming and tokenization etc.

In order to make raw data useful for Deep Learning and Machine Learning models, relevant information must be extracted through the process of feature engineering. Machine learning algorithms for classification use manual feature engineering techniques that are a part of machine learning approaches. The second is an illustration of one of the more recent Deep Learning techniques, which makes use of neural networks to automatically pick up multiple layers of abstract features from the input data. These features are then used for classification by the deep learning architectures.

This survey focuses on the earlier approaches used for the cyberbullying classification and detection. The following is how the paper is structured. The literature review on the classification and detection of cyberbullying is described in Section II. Section III gives the comparative analysis of the existing methods. Finally, section IV concludes this survey and offers a proposed idea to implement in the future

2. Literature review

The review of detection in cyberbullying and classification on multiple Twitter datasets, as well as numerous feature engineering techniques employed by earlier researchers, are the main topics of this section.

R. Zhao et al., [3] developed a representation learning framework specifically for detecting cyberbullying and used word embeddings to expand a predetermined set of insulting

terms and varying weights to produce bullying features, which are then merged with latent semantic features. Before being input into a linear Support Vector Machine (SVM) classifier, the final representation is created using a bag-of-words. This model gives 76.8% precision, 79.4% recall, and 78% F-score for Twitter dataset.

S. Minaee et al., [4] built a sentiment analysis framework using a combination of Long Short -Term Memory (LSTM) and Convolutional Neural Network (CNN) models. Two different datasets such as IMDB dataset (Movie review data) and Stanford Sentiment Treebank2 dataset (SST2) were used. Three deep learning models employed include individual methods such as LSTM, CNN as well as ensemble methods such as CNN+LSTM. Glove embedding was used to represent each word in reviews. Among all, CNN+LSTM outperformed well when compared to others with 90% accuracy for IMDB dataset and 80.5% accuracy for SST2.

V. Banerjee et al., [5] used CNN for the detection of cyberbullying and used the Twitter dataset. Glove was utilized to test several word embedding techniques. Several different layers such as Input, Word Embedding, Convolution, Max pooling, fully connected, Dropout, Softmax, and Classification layer in CNN were used. This method attained 93.97% accuracy.

T. Anuprathibha et al., [6] constructed a framework for automated sentiment analysis of tweets utilizing effective feature selection and classification methods. Three machine learning models were employed for classification, including k-Nearest Neighbour (KNN), Naive Bayes (NB), and Support Vector Machine, as well as two alternative feature selection methods, including the Dolphin Echolocation Algorithm (DEA) and Improved Dolphin Echolocation Algorithm (IDEA), which were performed over Cancer and Drugs dataset. Among all, IDEA-SVM outperformed with 96.58% accuracy.

C. Iwendi et al., [7] proposed the Deep learning algorithms' performance and efficacy in identifying insults in Social Commentary which were determined empirically. Wikipedia Detox dataset which includes an annotated dataset of 100,000 comments on Wikipedia articles was utilized.

A. Agarwal et al., [8] presented Bi-directional Long Short -Term Memory (Bi-LSTM), a Recurrent Neural Network (RNN) -based method, to identify and categorise the

bullying posts. The Tomek Link under-sampling method was used to lessen the data imbalance in the Wikipedia dataset utilised. Word embeddings from two separate sources were used to initialise the model, and the sparseness of the data representation in an embedding layer was reduced via max pooling. This method achieved 89% precision, 86% recall, and 88% F-score.

Ni Made Gita Dwi Purnamasari et al., [9] performed the cyberbullying classification on the twitter data. 300 tweets were used, of which 150 featured bullying and 150 did not. These tweets were manually categorized by professionals and divided into two categories. Support Vector Machine with Information Gain (IG) was used for classifying data using feature selection. This method achieved 75% accuracy, 70.27% precision, 86.66% recall, and 77.61% F-score.

Alotaibi M et al., [10] presented a way for automatically detecting aggressive behaviours in cyberbullying using a consolidated deep learning model. The Transformer block, BiGRU and CNN deep learning models were combined into the multichannel technique, and the hate speech dataset was utilized. This model attained 88% accuracy.

A. Kumar et al., [11] performed the bullying content classification using the Bi-GRU-Attention-CapsNet (Bi-GAC) model. Two different Benchmark datasets such as Formspring.me and MySpace were utilized, and to create a sequence context feature vector, an ELMo embedding-trained Bi-GRU encoder was used. This feature vector has problems since it contains unnecessary and useless features. This model attained 94.03% and 93.89% F1-score with the respective datasets.

K. S. Alam et al., [12] applied several n-gram analyses, machine learning algorithms, feature extraction and ensemble models over the twitter datasets. In the research, it was found that the best individual classifiers for detecting cyberbullying, Logistic Regression and Bagging ensemble models, were outperformed by the suggested Voting Classifiers for the "Single Level Ensemble (SLE) and Double Level Ensemble (DLE) models". When K-Fold cross-validation was combined with TF- IDF (Unigram) feature extraction, the suggested SLE and DLE models achieved 96% accuracy.

A. Al-Hassan et al., [13] analyzed Arabic hate speech detection in twitter platform using four deep learning models including individual methods such as LSTM, GRU as well as ensemble methods such as CNN+LSTM and CNN+GRU. Self-made corpus categories of Arabic tweets categorized into five different categories: none, religious, racist, sexist, or

general hate, were used. It includes 11000 labeled Arabic Tweets out of 37000 retrieved tweets from twitter API. An SVM model served as the baseline for comparison with 4 deep learning models. The findings demonstrate that in terms of identifying hateful tweets, all four deep learning algorithms beat the SVM model. The deep learning models have an average recall of 75%, whereas the SVM attained a recall of 74%. Among all, CNN+LSTM outperformed well with 72% precision, 75% recall and 73% F1score.

R. R. Dalvi et al., [14] presented the software to detect the bully tweets, posts, etc. Support Vector Machine, Naive Bayes, as well as Term Frequency and Inverse Document Frequency, are the different machine learning techniques that were utilised for categorization. The Twitter API was utilised to retrieve tweets from a place and identify whether or not they contain bullying. SVM achieved better accuracy of 71.25% than the NB.

3. Comparative Analysis

Table 1. Comparative analysis on the cyberbullying classification and detection methods

Ref. No.	Year	Dataset used	Methodology used	Result
[3]	2016	Twitter dataset	Support Vector Machine	78% f1-score
[4]	2019	IMDB dataset, SST2 dataset	Ensemble of CNN and LSTM	IMDB - 90% accuracy, SST2 – 80.5% accuracy
[5]	2019	Twitter dataset	Convolution Neural Network	93.97%-accuracy
[6]	2019	Cancer and Drugs Dataset	SVM with Improved Dolphin Echolocation Algorithm as feature selection technique	96.58%-accuracy
[7]	2020	Wikipedia Detox Dataset	Bidirectional Long Short - Term Memory	82.18%-accuracy
[8]	2020	Wikipedia dataset	Bi-directional long short - term memory	88% F-score.
[9]	2020	Twitter dataset	SVM with Information Gain as feature selection technique	75% -accuracy
[10]	2021	Hate speech dataset	Transformer block, BiGRU and CNN	88%-accuracy

[11]	2021	Formspring.me and MySpace dataset	Bi-GRU-Attention-CapsNet	Formspring.me - 94.03% F1-score, MySpace - 93.89% F1-score
[12]	2021	Benchmark dataset	Single Level Ensemble model and Double Level Ensemble model voting classifiers	96%-accuracy
[13]	2021	Self-made corpus	Ensemble of CNN and LSTM	73% F1-score
[14]	2021	Twitter dataset	Support Vector Machine	71.25%-accuracy

Table 1 shows the various datasets used to detect and classify the cyberbullying comments on the social media, and also shows the several machine learning, deep learning models for classification with the respective accuracy obtained by the models.

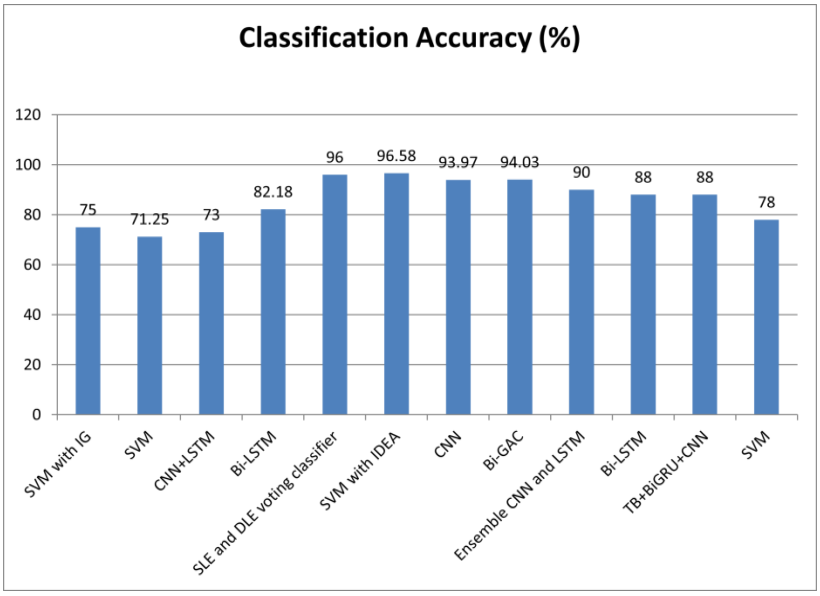

Figure 1. Performance of Various Classifiers

Figure 1 shows the accuracy of various classifiers such as Support vector machine, Convolution Neural Network, Long Short-Term Memory, and Bidirectional Long Short-Term Memory for the cyberbullying classification and detection in social media. From this figure, it is identified that the SLE and DLE voting classifier and the SVM with Improved Dolphin

Echolocation Algorithm attain similar accuracy. But compared with others, SVM with Improved Dolphin Echolocation Algorithm achieves better performance.

4. Conclusion

In this study, a thorough review on the existing methods that are already present in the classification and detection of using various deep learning and machine learning classification algorithms, cyberbullying, as well as the effect of word embedding methods on the classification task have been provided. From this study, it is observed that ensemble methods outperform well when compared with individual algorithms by attaining better results for the classification tasks. In addition, word embedding techniques influence more when combined with the algorithms for classification. In the future work, classification may be applied with ensemble methods for the classification and detection of cyberbullying for a better efficiency.

References

[1]All the Latest Cyberbullying Statistics for 2023(online). Available: https://www.broadbandsearch.net/blog/cyber-bullying-statistics.

[2]Summary of Our Cyberbullying Research (2007-2021) (online). Available: https://www.cyberbullying.org/summary-of-our-cyberbullying-research.

[3]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.

[4]S. Minaee, E. Azimi, and A. Abdolrashidi, “Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models”, arXiv Prepr. arXiv1904.04206, 2019.

[5]V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, ‘‘Detection of cyberbullying using deep neural network’’ in Proc. 5th Int. Conf.Adv. Comput. Commun. Syst. (ICACCS), Mar. 2019, pp. 604–607.

[6]T. Anuprathibha and C. S. Kanimozhiselvi, ‘‘Enhanced medical tweet opinion mining using improved dolphin echolocation algorithm based feature selection’’ Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 10, pp. 2049–2055, 2019.

[7]C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, ‘‘Cyberbullying detection solutions based on deep learning architectures’’Multimedia Syst., 2020.

[8]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 classweighting’’ in Neural Information Processing (Communications in Computer and Information Science), vol. 1333, Springer, 2020, pp. 113–120.

[9]N. M. G. D. Purnamasari, M. A. Fauzi, Indriati, and L. S. Dewi, “Cyberbullying identification in Twitter using support vector machine and information gain based feature selection’’ Indones. J. Electr. Eng. Comput. Sci., vol. 18, no. 3, pp. 1494–1500, 2020.

[10] Alotaibi M, Alotaibi B, Razaque A. A multichannel deep learning framework for cyberbullying detection on social media. Electronics. 2021; 10(21):2664.

[11] Kumar and N. Sachdeva, ‘‘A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media’’ World Wide Web, Jul. 2021.

[12] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, ‘‘Cyberbullying detection: An ensemble based machine learning approach’’ in Proc. 3rd Int.Conf.Intell. CommunTechnol. Virtual Mobile Netw. (ICICV), Feb.2021, pp. 710–715.

[13] Al-Hassan and H. Al-Dossari, ‘‘Detection of hate speech in Arabic tweets using deep learning’’ Multimedia Syst., Jan. 2021.

[14] 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.