Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.031848
Article

Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning

Khalid M. O. Nahar[1,*], Mohammad Alauthman[2], Saud Yonbawi[3] and Ammar Almomani[4,5]

[1]Computer Science Department, Yarmouk University, Irbid, Jordan
[2]Department of Information Security, Faculty of Information Technology, University of Petra, Amman, Jordan
[3]Software Engineering Department, University of Jeddah, Jeddah, KSA
[4]Research and Innovation Department, Skyline University College, P.O. Box 1797, Sharjah, UAE
[5]IT Department-Al-Huson University College, Al-Balqa Applied University, P. O. Box 50, Irbid, Jordan

*Corresponding Author: Khalid M. O. Nahar. Email: khalids@yu.edu.jo
Received: 28 April 2022; Accepted: 15 September 2022

Abstract: Social media networks are becoming essential to our daily activities, and many issues are due to this great involvement in our lives. Cyberbullying is a social media network issue, a global crisis affecting the victims and society as a whole. It results from a misunderstanding regarding freedom of speech. In this work, we proposed a methodology for detecting such behaviors (bullying, harassment, and hate-related texts) using supervised machine learning algorithms (SVM, Naïve Bayes, Logistic regression, and random forest) and for predicting a topic associated with these text data using unsupervised natural language processing, such as latent Dirichlet allocation. In addition, we used accuracy, precision, recall, and F1 score to assess prior classifiers. Results show that the use of logistic regression, support vector machine, random forest model, and Naïve Bayes has 95%, 94.97%, 94.66%, and 93.1% accuracy, respectively.

Keywords: Cyberbullying; social media; naïve bayes; support vector machine; natural language processing; LDA

1 Introduction

Technology and Internet use and integration into our daily lives for communication have made remarkable progress in recent years. However, its danger lies in its availability anytime, anywhere, and amongst all ages. In some cases, age restrictions are not placed on social media sites. Many children and teens are not mature enough to know the consequences of what they send or post on the Internet [1], and a serious negative effect is cyberbullying.

Cyberbullying can be defined as using the Internet and social networking sites to harm, abuse, threaten or provoke a person or group. Given the ability to access social media, posts, comments,

and any content shared by an individual on these platforms by acquaintances or strangers can easily be viewed. Whether personal, negative, or offensive content, what individuals share via social media creates an accessible public record known as an online reputation. Cyberbullying can harm anyone involved, not just the individual who has been bullied but also those who practice or even participate in it [2]. The study in [3] claims that teens who experienced cyberbullying thought about suicide at an average of 11.5, whereas those who experienced verbal bullying thought about suicide at an average of 8.4.

Based on the aforementioned consequences, early detection of cyberbullying on social platforms becomes essential to reduce its negative effects on individuals, particularly teens who decide to take their own lives because of undetected bullying.

The World Health Organization [4], United Nations International Children's Emergency Fund, and United Nations Organization for Cultural, Scientific, and Educational Aims have released a report on violence against children and the End Violence Partnership. The report shows that one billion children are physically, sexually, or psychologically abused yearly, and countries have not done enough to stop it. For a complete and accurate picture of the scope and nature of internet crimes against children, Baines [5] argues in his paper that international collaboration in research is essential. Children may easily access cyberbullying, a cybercrime that utilizes smartphones, because of its accessibility.

In this research, we proposed a methodology for detecting such behaviors (bullying, harassment, and hate-related texts) using supervised machine learning algorithms (SVM, Naïve Bayes, Logistic regression, and random forest) and for predicting the topic associated with these text data using unsupervised natural language processing, such as latent Dirichlet allocation (LDA). Consequently, the significant outcomes of this research are as follows:

- To improve performance, we developed an automated detection approach that includes feature extraction into the classifiers.
- Several commonly used supervised machine learning classifiers for cyberbullying detection were evaluated.
- Unsupervised LDA was used to identify the text's topic.
- We tested the usability and accuracy of the classifiers investigated in this work on a large, generic dataset.

2 Literature Reviews

We have identified many types of research about cyberbullying detection. Researchers have provided a fertile environment related to cyberbullying based on text mining paradigms. Hence, this research inspired researchers after theme, making it easier to start but difficult to innovate and add something new.

Dinakar et al. [6] designed predictive models for automatic cyberbullying detection, including visual features to complement textual features. They have used Linguistic Inquiry and Word Count to analyse the text for cyberbullying and used Microsoft's Oxford to extract visual features from images. For classification, they used the Bagging Algorithm. The accuracy for the textual approach was 77.6%, whereas that of the visual approach was 70.5%, and that of the combined approach was 81.4%.

In 2021, Desai et al. [7] proposed a transformer-based cyberbullying detection model based on the bidirectional encoder from transformer (BERT) architecture. The authors improved the BERT architecture by adding a task-specific layer. These models had 71.25% and 52.70% accuracy rates when using SVM and Naive Bayes, respectively; the authors claimed that their model had an accuracy rate

of 91.90% when using the same dataset, which is more than the accuracy rate of the typical machine learning model.

Muneer et al. [8] used several classification algorithms to investigate pooled Twitter datasets, including RNN, CNN, and transformer-based classifiers. Decoupled weight decay optimizer (AdamW) was used to improve BERTweet's F1 score by up to 8.4%, resulting in a 64.8% macro F1 score for BER Tweet. The ensemble model was further improved using particle swarm optimization. BER Tweet's standalone model outperformed the ensemble by 0.53%, resulting in 65.33% macro F1 for the TweetEval dataset and the combined datasets by 0.55%, resulting in 68.1% macro F1.

A worldwide dataset of 37,373 unique tweets from Twitter was collected by Singh et al. to examine cyberbullying tweets [9]. Several machine learning techniques were used (logistic regression, light gradient boosting machine, stochastic gradient descent [SGD], random forest, AdaBoost Naive Bayes, and support vector machine [SVM]). In addition, various performance indicators were used to gauge how well each classifier performed on the global dataset, including accuracy, precision, recall, and the F1 score. LR outperformed all other methods, with a median accuracy of 90.57%. Logistic regression achieved an F1 score of 0.928; SGD achieved a precision of 0.968, and SVM achieved a recall of 0.964 and 1.00.

Different approaches were applied [10] to obtain the desired results. Incorporating user information, such as gender and age, would increase the accuracy of cyberbullying detection. Precision, recall, and f-measures were calculated to classify the dataset using SVM and 10-fold cross-validation, as shown in Table 1.

Table 1: Accuracy measurement for paper [10]

The feature used in the classifier	Precision	Recall	F-measure
Female specific (34% corpus)	0.40	0.05	0.08
Male specific (66% corpus)	0.44	0.21	0.28

In 2014, Nahar et al. [11] designed a novel framework for automatic cyberbullying detection and streaming data with insufficient labels. The framework can extract valid positive and negative examples by utilizing enhanced training methods based on the confidence voting function. User context, language understanding, and baseline keywords were used to construct the enhanced feature sets in the proposed method. They also suggested a fuzzy SVM technique for effectively detecting cyberbullying by this team of researchers. Streaming data are dynamic and complex, and the proposed technique efficiently addresses these issues. Moreover, they have varying levels of accuracy and recall depending on the situation.

Nahar and colleagues created an automatic cyberbullying detection framework in 2014 [11]. The framework uses enhanced training methods based on the confidence voting function to extract reliable positive and negative examples. User context, linguistic understanding, and baseline keywords generated the enriched feature sets. The researchers also proposed a fuzzy SVM algorithm for cyberbullying detection. The proposed method effectively addresses the dynamic and complicated nature of streaming data. They have different precision and recall results for each scenario.

In 2020, Singh et al. [12] combined different machine learning models such as Naïve Bayes with term frequency-inverse document frequency algorithm (TF-IDF) word vectors, and the SVM using

their custom word vectors to improve their model evaluation matrices had a 0.96 accuracy rate with 0.88 precision and 0.94 recall.

In 2020, Ali et al. [13] discussed how to detect cyberbullying using text datasets that were publicly available on the Internet. The authors split the datasets into 80% training and 20% testing sets and then applied classification algorithms. They used SVM, logistic regression, Naïve Bayes, random forest, and ensemble approaches.

These classification methods on each dataset had an average accuracy for each classifier, as shown in Table 2.

Table 2: Accuracy measurement for paper [13]

Algorithm	Accuracy
Random Forest	78%
Naïve Bayes	76%
SVM	80%
Logistic Regression	78.6%
Ensemble	79.5%

The aforementioned papers produced a decent job related to cyberbullying detection, but cyberbullying has several types and forms. We found the importance of estimating the type of textual data and determining whether or not such data are related to bullying, which was proposed in this paper.

Amongst Arab scholars, Abaido [14] examined the prevalence of cyberbullying, its venues, its character, and their attitude toward reporting cyberbullying to combat residual silence. More than 200 academics in the United Arab Emirates have contributed data. More than three-quarters (78%) mentioned that the most popular platforms for cyberbullying were Instagram (55.5%) and Facebook (38%). Furthermore, calls for smartphone applications are among the options being considered.

Cyberbullying detection on Instagram has been improved with the introduction of CONcISE by Yao and colleagues [15] for each comment classification. They suggested a formula that uses fewer features while maintaining better classification accuracy.

Ozel et al. [16] used ML approaches such as SVM, C4.5, NB multinomial, and KNN classifiers to identify cyberbullying in Turkish tweets and Instagram posts. Chi-square and data gain methods are also used to improve classification accuracy.

An ML technique proposed by Das et al. [17] was used to identify people involved in harassment-based bullying and a new word indicator of bullying. Infers and social structures in which the user prefers to bully and victimize are also considered. The learning strategy requires less monitoring to deal with the complex character of cyberbullying.

Yadav et al. [18] presented a new method for identifying cyberbullying on social media networks. An additional linear neural network classification layer is utilised with the BERT model's single linear neural network layer. Form spring, and Wikipedia datasets are used to train and evaluate the model. A considerable improvement in the performance accuracy of the suggested model is observed in the Form spring and Wikipedia datasets compared with the performance accuracy produced by Previous models. The proposed model does not require oversampling because of the vast size of the Wikipedia dataset to obtain better results. On the contrary, the Form spring dataset requires oversampling.

Relevant research comparison results are summarised and shown in Table 3.

Table 3: Comparative analysis of related research

Researchers	Classifier	Main results	Dataset
Rafiq and colleagues [19], 2018	AdaBoost and LR	NA	Vine
Galán-García and colleagues [20], 2016	NB, KNN, RF, J48 and SMO	- SMO: 0.684% - J48: 0.658% - RF: 0.664% - NB: 0.339% - KNN: 0.0597%	Twitter
Al-garadi and colleagues [21], 2016	SVM and NB	- Naïve Bayes: 0.71% - SVM: 0.78%	Twitter
Salminen and colleagues [22], 2020	LR, NB, SVM and XGBoost	-LR: 0.768% - NB: 0.606% - SVM: 0.648% - XGBoost: 0.774%	Wikipedia, YouTube, Reddit and Twitter
Van Hee and colleagues [23], 2018	LSVM	F1 score (English): 0.64%. F1 score (Dutch) 0.61%	ASKfm in Dutch and English

Harassment and bullying have become a familiar and easy societal phenomena in a vast space like social media. Thus, this paper aims to detect and estimate the type of such actions and prevent or at least decry these incidents on social media. Having someone to watch over teenagers is common, but they cannot detect everything those teenagers write and send. A system that provides help in detecting suspicious activity would help the victims and the bully.

3 Methodology and Experiment

In this paper, we provided a thorough description of the dataset used to detect cyberbullying on Twitter, its visualization, and the methods presented for performing sentiment analysis on the dataset selected and addressing the assessment metrics of each classifier. A model of cyberbullying is depicted in Fig. 1, including pre-processing, feature extraction, classification, and evaluation, and we have covered each phase in great detail.

3.1 Data Collection

In this research, we have used the Twitter sentiment analysis dataset provided by Analytics Vidhya [24]. This dataset consists of 32k tweets labeled as hatred or non-hatred-related tweets. A sample of the dataset we have used is shown in Table 4**: Sample of the dataset**.

3.2 Data Pre-Processing

For the detection of cyberbullying, the pre-processing stage is essential. In addition, removing spam and tidying up the text (for example, by removing stop words and punctuation marks) are necessary [25]. Unwanted noise in text detection was cleaned up using this model. For example, repetitive words, special characters, and stop words were all omitted from the final product. Given

this pre-processing, the remaining words can be stemmed back to their original roots, resulting in an entirely clean dataset for the prediction and training functions of the proposed model.

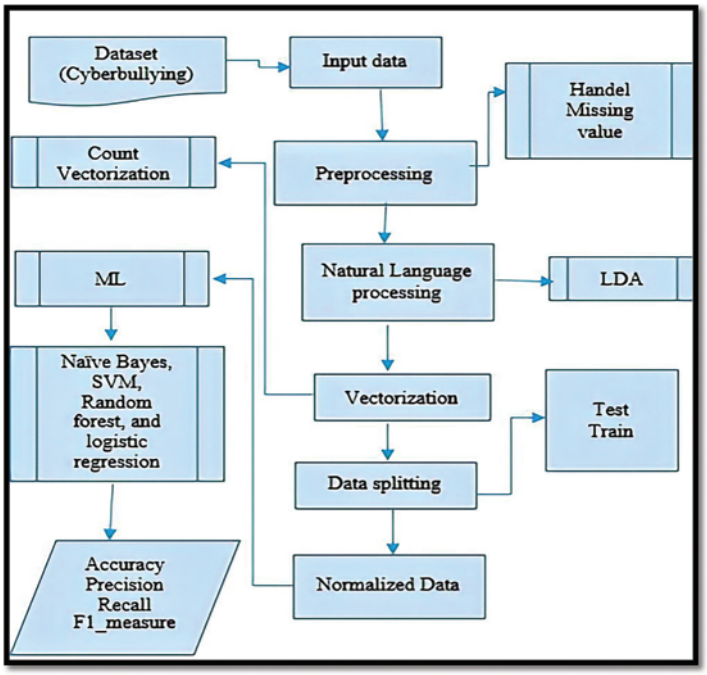

Figure 1: Research methodology

Table 4: Sample of the dataset

Id	Label	Tweet
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	1	@user proves that small dicks make small minds #deploraball #whitesheet #inbred
4	1	@user tonight in #cologne. proudly presented by the government, police + media! it's a big staging going on there!

3.3 Estimating the Topics

An essential step in detecting cyberbullying in our work is to estimate the type of text data we have and handle this matter. We have used the topic modeling technique, a branch of Unsupervised Natural Language Processing, and the probabilistic modeling approach, namely, LDA, in our model.

We defined the following words to comprehensively understand LDA:

- **Latent:** This refers to anything concealed in the data. We are unaware of a priori [26].
- **Dirichlet:** Dirichlet analysis is used in topic modeling to analyze the distribution of topics and words within a topic [26].
- **Allocation:** After LDA, we will allocate themes to the papers and words in the document to topics [26].

We have used two matrices to understand the mathematics behind LDA:

1. Θtd = P (t | d) is the probability distribution of topics in documents, where Θ (theta) is the distribution of subjects over the content of a document.
2. Φwt = P (w | t) the distribution of words in the topic's text, where Φ (phi) represents a distribution of words over topics

Furthermore, the probability of a word given in a document, that is P (w | d), is presented as follows:

$$\sum_{t \in T} P(w|t,d)\, p(t,d) \tag{1}$$

where T is the total number of topics; hence, P (w | d) is presented as follows:

$$\sum_{t=1}^{T} p(w|t)p(t|d) \tag{2}$$

The probability of W indicates the number of terms in our lexicon for all the texts. Fig. 2***: LDA Matrix*** represents the equations in the form of a matrix.

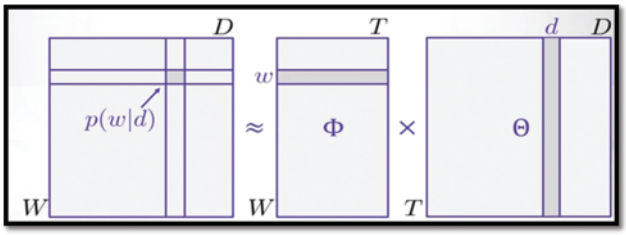

Figure 2: LDA Matrix

Finally, we obtain a conditional probability equation for a single word in a document that is about a certain subject:

$$p\left(z_{d,n} = K|\vec{z}_{\_d,n}, \vec{w}, \alpha, \gamma\right) = \frac{n_{d,k} + \alpha_k \; V_{k,W_{d,n}} + \gamma_{Wd,n}}{\sum_i^k n_{d,i} + \alpha_i \; \sum_i V_k + \gamma_i} \tag{3}$$

where $n\ (d,\ k)$ indicates instances in which document D refers to topic K; $V\ (k,\ w)$ indicates topic k's frequency of use of the given word; αk indicates the Dirichlet distribution parameter for distributing documents to topics, and λw indicates the topic-to-word distribution Dirichlet parameter.

The above-mentioned equation has two components. The frequency with which a topic appears in a text may estimate its importance in the document.

3.4 Dataset Splitting

After the dataset has been pre-processed and normalized, we split the dataset into 70% training set and 30% testing set.

3.5 Data Vectorisation

We must convert the text to numerical form and vectorize the tweets to provide a simple approach to tokenizing the collection of texts and building a vocabulary of recognized terms because the model cannot understand human language. A key component of cyberbullying categorization is extracting features from the text. We utilized TF-IDF to extract features for the suggested model. Text feature extraction is based on word statistics in the TF-IDF, a hybrid of TF and IDF algorithms [27].

3.6 Data Classification

Different classifiers were used in this study to classify tweets as cyberbullying or not. SVM, Nave Bayes, random forest, and logistic regression are classifier models developed during data classification. F1 score and accuracy rate were computed for each model separately.

4 Experimental Results

Several evaluation measures were used in this study to assess the model's ability to distinguish cyberbullying from non-cyber bullying. Four machine learning algorithms have been used in this study: SVM, Naive Bayes, random forest, and logistic regression. A study of accepted research measures is essential for assessing the relative merits of various models under investigation. SM platforms (such as Twitter) with cyberbullying classifiers are typically evaluated using the following criteria:

- Accuracy is a metric used to assess the accuracy of cyberbullying prediction models by comparing the number of actual cases with the total number of cases. Consequently, the following value may be calculated as follows:

$$Accuracy = \frac{(\mathrm{tp} + \mathrm{tn})}{(\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn})} \tag{4}$$

The terms 'true positive,' 'true negative,' 'false positive' and 'false negative' are used interchangeably.

- **Precision** measures the percentage of relevant tweets in a given group, which are true positives (tp) and false positives (fp).

$$\mathrm{Precision} = \mathrm{tp}\,(\mathrm{tp} + \mathrm{fp}) \tag{5}$$

- **Recall** determines the percentage of relevant tweets that have been retrieved from the total amount of relevant tweets.

$$\mathrm{Recall} = \mathrm{tp}\,(\mathrm{tp} + \mathrm{fn}) \tag{6}$$

- Using the F-measure, precision and recall can be measured separately.

$$\mathrm{F\,measure} = 2 \times \mathrm{precision} \times \mathrm{recall}\,\mathrm{precision} + \mathrm{recall} \tag{7}$$

The primary topics in the data were identified using LDA after pre-processing and normalizing the dataset. Using LDA, a machine learning technique, a set of documents may be analyzed to locate clusters of words. Unstructured data can be analyzed using word count and comparable word

patterns [28]. By detecting patterns amongst text data, such as word frequency and distance between similar words and expressions that appear most often, we estimated what each text set is about. The results are shown in Fig. 3***: Token Distribution***.

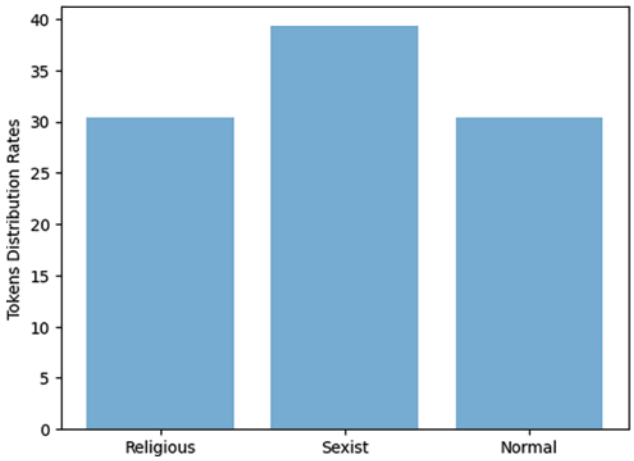

Figure 3: Token distribution

We then started the classification process to differentiate non-bullying from bullying or hatred-related data.

As mentioned earlier, the data were split into 70% training and 30% testing sets. We used supervised learning algorithms to detect cyberbullying text data for each classifier. We calculated the precision, recall, and F1 measure alongside the accuracy rate. The results are shown in Fig. 4: Random *forest confusion matrix*, Fig. 5: Naive Bayes *confusion matrix*, Fig. 6: SVM *confusion matrix*, and Fig. 7: Logistic *regression*.

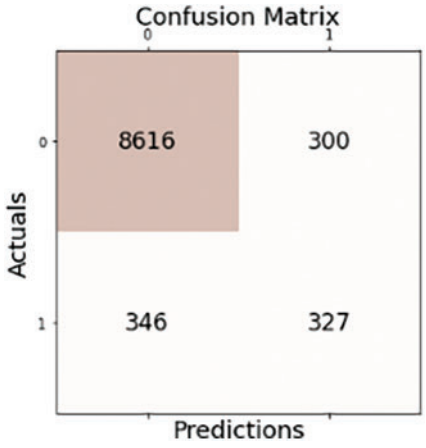

Figure 4: Random forest confusion matrix

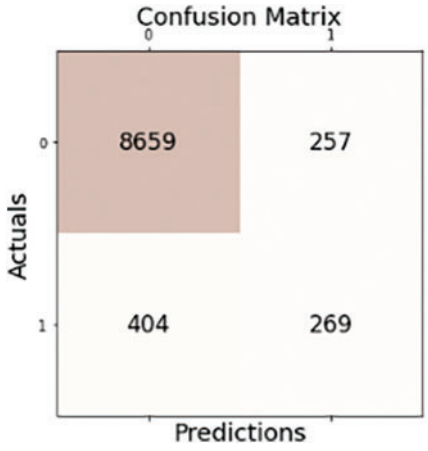

Figure 5: Naive Bayes confusion matrix

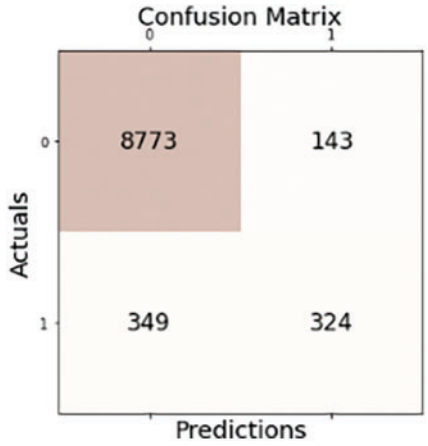

Figure 6: SVM confusion matrix

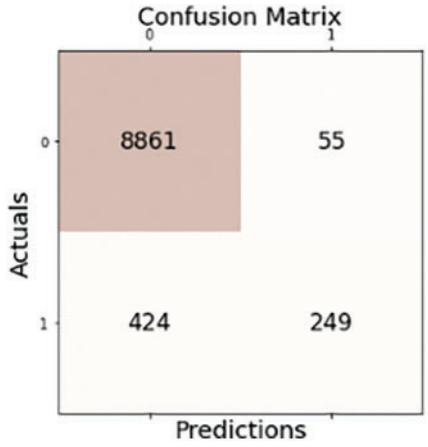

Figure 7: Logistic regression

As shown in the above figures, there are a diversity of the confusion matrix results, such as the results of logistic regression and random forest. Logistic regression is often used to solve large-scale industrial problems because it is easy to use. Usually, it does not give a single output but the probability of each output. The logistic regression algorithm is not too affected by small multicollinearity and can handle small amounts of noise in the data. Conversely, the Random Forest Classifier works better with more categorical data than numeric data and logistic regression, which is a little confusing regarding categorical data. Logistic regression works better when the number of noise variables is less than or equal to the number of explanatory variables. As the number of explanatory variables in a dataset grows, the random forest has a higher rate of both true and false positives.

These confusion matrix results and the accuracy rate of classifiers are summarised in Fig. 8: *Accuracy* measurement *of* supervised machine learning *classifiers* for cyberbullying detection.

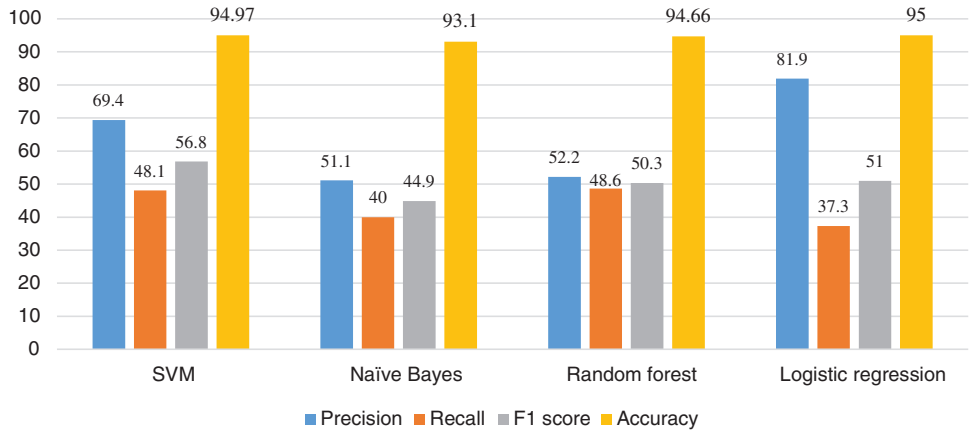

Figure 8: Accuracy measurement of supervised machine learning classifiers for cyberbullying detection

5 Conclusion and Future Work

This work introduced cyberbullying detection, an important problem in this era. We proposed a methodology by which we can detect and find the type of textual dataset. Kaggle provided the method used to investigate real-world tweets. The results show that our dataset has three main topics: religious tweets with 30.4% token distribution, 39.3% of tokens related to sexist tweets, and 30.3% of tokens related to not known topics, which we labeled as normal topics. Finally, we were satisfied with the accuracy rate of our learning model. Our SVM, Naïve Bayes, random forest, and logistic regression reached an accuracy rate of 94.9%, 93.1%, 94.66%, and 95%, respectively. **Future work:** The estimation method used for the dataset is valid and applicable, but it is not accurate enough because of the estimation criteria for the distribution of words related to each cluster. Therefore, identifying different datasets related to text cyberbullying and a vector matrix of the words in each new dataset is necessary to calculate the cosine similarity between these matrices and clusters we obtained after applying LDA to our original dataset.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Olweus, "Cyberbullying: An overrated phenomenon?" *European Journal of Developmental Psychology*, vol. 9, no. 5, pp. 520–538, 2012.

[2] W. Akram and R. Kumar, "A study on positive and negative effects of social media on society," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 10, pp. 351–354, 2017.

[3] P. K. Smith, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of Child Psychology*, vol. 49, no. 4, pp. 376–385, 2008.

[4] E. G. Krug, L. L. Dahlberg, J. A. Mercy, A. B. Zwi, and R. Lozano, "Global status report on preventing violence against children," *World Health Organization*, Switzerland, 2020.

[5] D. V. Baines, "Online child sexual abuse: The law enforcement response," in *Presented at the World Congress III Against the Sexual Exploitation of Children and Adolescents, s. ECPAT Int.*, Rio deJaneiro, Brazil, 2008.

[6] K. Dinakar, R. Reichart and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, California, United States, vol. 5, no. 3, pp. 11–17, 2011.

[7] A. Desai, S. Kalaskar, O. Kumbhar and R. Dhumal, "Cyber bullying detection on social media using machine learning," in *ITM Web of Conf.*, vol. 40, EDP Sciences, Mumbai, India, pp. 03038, 2021.

[8] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on twitter," *Future Internet*, vol. 12, no. 11, pp. 187, 2020.

[9] V. K. Singh, S. Ghosh and C. Jose, "Toward multimodal cyberbullying detection," in *Proc. of the 2017 CHI Conf. Extended Abstracts on Human Factors in Computing Systems*, pp. 2090–2099, United States, 2017.

[10] M. Dadvar, F. D. Jong, R. Ordelman and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, University of Ghent, Belgium, 2012.

[11] V. Nahar, S. Al-Maskari, X. Li and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Australasian Database Conf.*, Springer, Australia, pp. 160–171, 2014.

[12] S. Singh and S. K. Shakambhari, *Cyber-bullying Detection Using Machine Learning*, Bangalore: CMR Institute of Technology, 2020.

[13] A. Ali and A. M. Syed, "Cyberbullying detection using machine learning," *Pakistan Journal of Engineering Technology*, vol. 3, no. 2, pp. 45–50, 2020.

[14] 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.

[15] M. Yao, C. Chelmis and D. -S. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," in *The World Wide Web Conf.*, United States, pp. 3427–3433, 2019.

[16] S. A. Özel, E. Saraç, S. Akdemir and H. Aksu, "Detection of cyberbullying on social media messages in turkish," in *2017 Int. Conf. on Computer Science and Engineering (UBMK)*, Turkey, pp. 366–370, 2017.

[17] K. Das, S. Samanta and M. Pal, "Study on centrality measures in social networks: A survey," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–11, 2018.

[18] J. Yadav, D. Kumar and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in *2020 Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, India, pp. 1096–1100, 2020.

[19] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv and S. Mishra, "Scalable and timely detection of cyberbullying in online social networks," in *Proc. of the 33rd Annual ACM Symp. on Applied Computing*, United States, pp. 1738–1747, 2018.

[20] B. Irena and E. B. Setiawan, "Fake news (hoax) identification on social media twitter using decision tree c4. 5 method," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 4, no. 4, pp. 711–716, 2020.

[21] 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.

[22] J. Salminen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–34, 2020.

[23] C. Van Hee, "Automatic detection of cyberbullying in social media text," *PLoS One*, vol. 13, no. 10, pp. e0203794, 2018.
[24] A. Vidhya, "Twitter sentiment analysis," [Online]. 2022. Available: https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis
[25] J. -M. Xu, K. -S. Jun, X. Zhu and A. Bellmore, "Learning from bullying traces in social media," in *Proc. of the 2012 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Canada, pp. 656–666, 2012.
[26] H. Jelodar *et al.,* "Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.
[27] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Logic Journal of the IGPL*, vol. 24, no. 1, pp. 42–53, 2016.
[28] M. Mustak, J. Salminen, L. Plé and J. Wirtz, "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda," *Journal of Business Research*, vol. 124, pp. 389–404, 2021.