Scientific Journal of Informatics
Vol. 9, No. 2, Nov 2022
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/.php/sji e-ISSN 2460-0040

A Comparative Analysis of Classification Algorithms for Cyberbullying Crime Detection: An Experimental Study of Twitter Social Media in Indonesia

Ari Muzakir[1*], Hadi Syaputra[2], Febriyanti Panjaitan[3]
[1,2,3]Department of Informatics Engineering, Faculty of Computer Science,
Universitas Bina Darma, Indonesia

Abstract.
Purpose: This research aims to identify content that contains cyberbullying on Twitter. We also conducted a comparative study of several classification algorithms, namely NB, DT, LR, and SVM. The dataset that we use comes from Twitter data which is then manually labeled and validated by language experts. This study used 1065 data with a label distribution, namely 638 data with a non-bullying label and 427 data with a bullying label.
Methods: The weighting process for each word uses the Bag Of Word (BOW) method, which uses three weighting features. The three-word vector weighting features used include unigram, bigram, and trigram. The experiment was conducted with two scenarios, namely testing to find the best accuracy value with the three features. The following scenario looks at the overall comparison of the algorithm's performance against all the features used.
Result: The experimental results show that for the measurement of accuracy weighting based on features and algorithms, the SVM classification algorithm outperformed other algorithms with a percentage of 76%. Then for the weighting based on the average recall, the DT classification algorithm outperformed the other algorithms by an average of 76%. Another test for measuring overall performance (F-measure) based on accuracy and precision, the SVM classification algorithm, managed to outperform other algorithms with an F-measure of 82%.
Value: Based on several experiments conducted, the SVM classification algorithm can detect words containing cyberbullying content on social media.

Keywords: Cyberbullying, Model Comparison, Machine Learning, Bag of Word, Classification
Received March 2022 / **Revised** March 2022 / **Accepted** October 2022

INTRODUCTION

Social media has become a tool for conveying ideas, freedom of expression, sharing information and knowledge freely on network media. Indonesia is a country with the most users in the world. The Ministry of Communication and Information (Kemenkominfo) noted that around 90% are active social media users. However, freedom of expression on social media is an opportunity for acts of abuse that are offensive, blasphemous, and other actions that harm other people or specific groups. Twitter is a microblog application widely used by social media activists in Indonesia. Every day, millions of information are shared freely and quickly on social networks. Negative actions often occur on social networks and are uncontrolled in number, such as fraud, hoaxes, opinions that contain hate, to acts of bullying by social media users themselves.

In a survey conducted by the Indonesian Internet Service User Association (APJII), active social media users were 196.71 million from 266.91 million users in 2020 [1]. The average time users spend surfing in cyberspace per day is more than 5 hours. When viewed from the number of internets and social media users in Indonesia, the data is fantastic. However, behind the increasing penetration of internet users, many facts related to crime and harassment occur between social users themselves. Indonesia ranks third in terms of cyberbullying after Japan and South Korea [2]. Cyberbullying is the most experienced by school-age children in 40 countries, including Indonesia. Nearly 90% of Indonesian citizens stated that they had experienced bullying on social media [3].

*Corresponding author.
Email addresses: arimuzakir@binadarma.ac.id (Muzakir), hadisyaputra@binadarma.ac.id (Syaputra), febriyanti_panjaitan@binadarma.ac.id (Panjaitan)
DOI: 10.15294/sji.v9i2.35149

Sentiment analysis studies have recently become a new trend and increase, especially regarding text classification. Several studies that discuss hate speech, abusive, trolling, aggression, and cyberbullying are classified based on Twitter social media tweets. Hidayatullah et al. [4] classified tweets into two classes and compared the performance of the NBC [5] and SVM algorithms in classifying [6]. Fadli et al. [7] identified cyberbullying on Twitter using the deep learning method to see the best accuracy for cyberbullying detection. Based on the results of previous research, it is necessary to take proper and immediate action to prevent early cyberbullying practices on social media.

The problem of diversity in language, different definitions of a word that leads to hate speech, and limited data available for training and testing the system became a challenge [8]. There is still much content that is hateful, offensive, and causes unpleasantness [9], [10]. Of course, this has to do with the influence of the language used in content posted by users that is unpleasant [11]. The automated hate speech detection process faces challenges in identifying hateful content, such as non-standard variations in spelling and grammar. It is not enough to use general data analysis to identify because it will be time-consuming and complex. However, it is necessary to automatically identify with a more in-depth and accurate classification technique. This study aims to identify content that contains cyberbullying on Twitter. We also conducted a comparative study of several classification algorithms, namely Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), and Support Vector Machine (SVM). This study classifies data to detect cyberbullying on Twitter social media into two classes, namely bullying and non-bullying. The bullying class contains tweets in the form of words containing cyberbullying elements. In contrast, the non-bullying class contains tweets in words that do not contain elements of cyberbullying. The benefits of this research can help parents, governments, and countries to protect young people from online bullying (cyberbullying) and reduce the number of cyberbullies. There are two contributions to our research, namely: (1) We modified the new dataset to detect and identify cyberbullying content on social media, and (2) We experiment with the dataset to measure the best performance of machine learning methods and model comparisons from classification algorithms including Naive Bayes, Decision Tree, Logistic Regression, Support Vector Machine to find the best model.

This paper will explain into four sessions. Session 2 discussed the methodology and data used. Discussion of the experiments and test results is in session 3. In session 4, we conclude our work and plan for further research.

METHODS

This section discusses how to create the dataset and the methodology in conducting cyberbullying using the machine learning classification approach.

Data Preparation

Data preparation, also known as data pre-processing, is the initial process of obtaining and preparing data so that the initial raw data can be processed and used at a later stage. Based on Figure 1, three processes are carried out at the data preparation stage: data collection, data cleaning, and data pre-processing. The first stage is preparing the dataset to obtain and process the data in this study, and then we modify the addition of the data. Based on the results of the literature and previous research, search for keywords on Twitter to find words that contain cyberbullying, namely physical appearance, sex, culture, politics, cornering words (such as idiots, losers, and others). From the scraping results obtained valid data as much as 1870 data. Furthermore, the data went through a cleaning and labeling process to obtain as many as 1065 data. The labeling process was carried out manually with the assistance of several students from universities in Palembang. Furthermore, the annotation process was carried out by three linguists from a private university in Palembang to validate the labeling suitability.

For the classification process, labeling is determined for each data, using 1 for *Bullying* and 0 for *Non-Bullying*. The data distribution is 638 data with a Non-Bullying label and 427 with a Bullying label (Figure 2). After the data is clean, then proceed to the preprocessing data stage. Data preprocessing consists of 6 stages: case folding, filtering, stemming, tokenization, padding, and vectorization. The data generated from the preprocessing stage is data in the form of word vectors, resulting from the last preprocessing stage, namely vectorization. The next stage is a model classification that aims to produce a model to classify cyberbullying. This stage has four stages: data splitting, model training using machine learning models, model testing and evaluation, and determining the best parameters to see the best model capable of detecting and identifying cyberbullying content on Twitter.

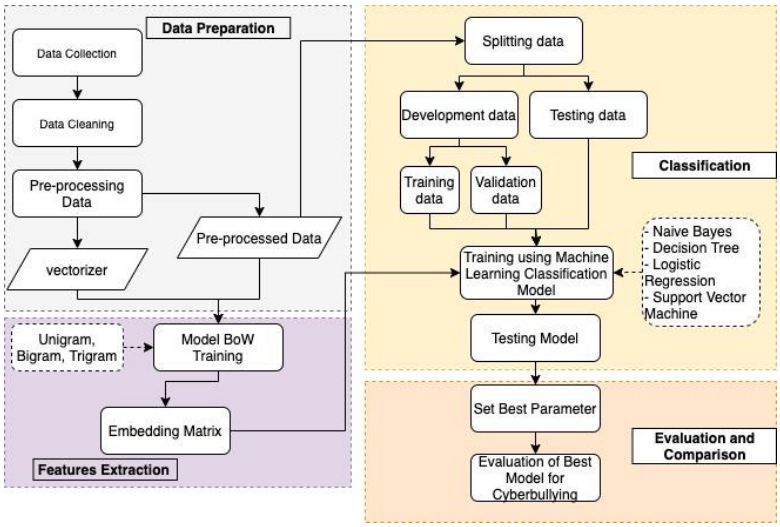

Figure 1. The proposed architecture model

The data splitting stage aims to divide the data to be used to train and test classification models such as NB, DT, LR, and SVM [12], [13]. The data distribution is 60 % training data and 40 % testing data. The evaluation phase aims to measure the performance of the resulting model based on the best parameters determined at the testing and evaluation phase.

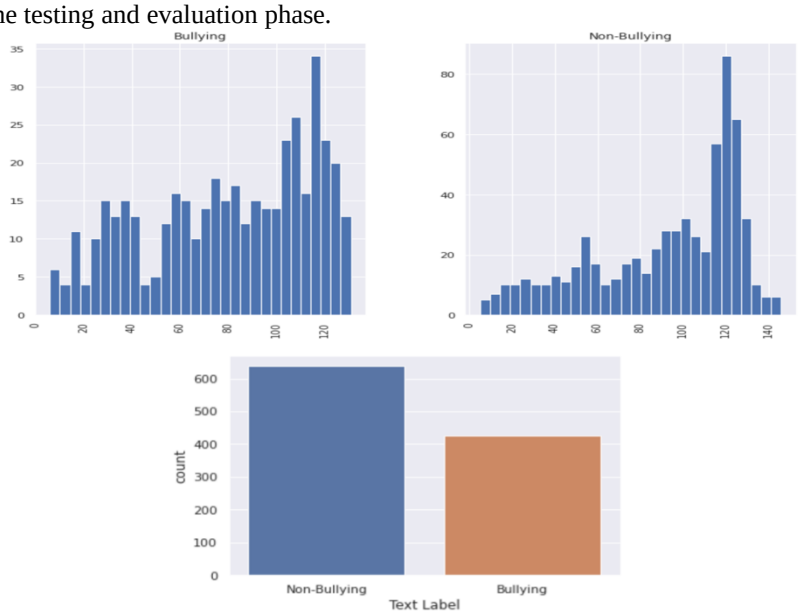

Figure 2. Data distribution process

Another goal of our research is to compare machine learning features and algorithms to find out which combination of features and algorithms has the best performance. Based on Figure 1, we carried out several stages: 1) pre-processing the data; 2) feature extraction; and 3) classification and evaluation. Then we adopted the pre-processing method used by [8] with a slight modification (Table 1). There are six steps in the general data pre-processing stage, namely: 1) deletion of retweets; 2) text cleaning; 3) lowercase; 4) spelling correction; 5) negation handling; and 6) removal of stopwords.

Table 1. The examples of pre-processing data

Before	After
@pengguna 1 @pengguna 2 idihh mau jadi artis dadakan cuyyy…	mau jadi artis dadakan
@pengguna 3 yahhh… gk penting amat lu sok kecantikan dikampung..	Tidak penting sekali kamu kencantikan di kampung

Extraction Features

After going through the pre-processing stage, the extraction process is carried out to find attributes that have an important influence in a sentence and reduce the dimensions of the data. This technique is known as word weighting. We use the Bag Of Words (BOW) model [14] to represent this study's text. In general, in the BOW model there are 3 features, namely bigram, unigram, and trigram to do the weighting for each word vector in each sentence.

Classification and Evaluation

Our research is based on supervised learning to identify and classify bullying content on Twitter. We use four classification algorithms for comparison to see which algorithm is the best, namely NB, DT, LR, and SVM. We used the Google Colab platform based on Python programming to do this research. The evaluation in this study used a 10-fold cross-validation method. We only use the average percentage weighting of accuracy and recall for each word vector weighted in a sentence based on features. To measure the performance of cyberbullying detection, we use the average percentage weighting for *accuracy, precision, recall* [15].

Accuracy measurement aims to see the predictions (positive and negative) ratio to the overall data (see Equation (1) below).

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

Precision measurement shows the true positive predictions ratio compared to the positive results (see Equation (2) below).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall measurement (sensitivity) aims to see the ratio of true positive predictions compared to the comprehensive data that are true positive (see Equation (3) below).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Where *TP* is a true positive, namely the number of positive data that is classified correctly by the system. *TN* is true negative, which is the number of negative data that is classified correctly by the system. *FN* is a false negative, that is, the number of negative data but is classified incorrectly by the system. While the *FP* is a false positive, that is, the number of positive data but is classified incorrectly by the system.

RESULT AND DISCUSSION

In this section, we discuss the experiment results and analysis.

Experiment Result

We carried out three experimental scenarios in this study. First, we compare each combination of features and algorithms used. Second, we compare the performance of each feature of the BOW model, namely bigram, unigram, and trigram. Next, we tested the combination of the three features based on the classification algorithm used. Third, we compare the performance of each algorithm when using the metric score measurement methods simultaneously, namely accuracy, precision, recall, and f-measure.

Table 2. Accuracy for each feature and algorithm

Feature	Weighted Avarage Accuracy (%)			
	NB	DT	LR	SVM
Unigram	0.65	0.71	0.70	**0.75**
Bigram	0.62	0.71	0.70	**0.76**
Trigram	0.59	0.71	0.70	**0.76**
Combination (unigrams, bigrams, and trigrams)	0.68	0.71	0.70	**0.76**

Table 2 shows the classification process based on the combination of the machine learning algorithms used in this study. The classification algorithm and word representation features are measured based on the average word vector weighting using word unigrams, bigrams, and trigrams. The best accuracy based on experiments was obtained through the SVM classification algorithm with an almost stable mean value for each feature, namely 75% (unigram) and 76% (bigram, trigram, and combination), followed by the DT

algorithm (70%) and LR (71%). The measurement of the average weighting based on the recall percentage in Table 3 shows that the DT algorithm gets the best F-measure score with a value for each feature between 75% to 77%. The score has improved on the combination features, both for the DT algorithm itself and for other algorithms such as SVM, which is slightly below it with a score of 76%, and NB with a score of 72%.

Table 3. The recall for each feature and algorithm

Feature	Weighted Avarage Recall (%)			
	NB	DT	LR	SVM
Unigram	0.52	**0.75**	0.62	0.62
Bigram	0.65	**0.76**	0.72	0.62
Trigram	0.58	**0.76**	0.63	0.63
Combination (unigrams, bigrams, and trigrams)	0.72	**0.77**	0.62	0.76

For the overall measurement of the classification algorithm based on performance using all features, the F-measure score of the SVM algorithm is obtained, which is superior to other algorithms (Table 4). Table 4 compares the average F-measures between the four algorithms when using all of these features. SVM is 82% superior, followed by DT by 80%, while NB always gets the lowest score in each measurement. Interestingly, there is a high spike in the recall score of DT, which was initially only at 76 %, but the average F-measure performance was at a score of 84%.

Table 4. The measure performance using all features and algorithms

	NB	DT	LR	SVM
Accuracy	0.66	0.71	0.70	**0.76**
Precision	0.76	0.70	0.73	**0.78**
Recall	0.66	**0.84**	0.83	0.80
F-measure	**0.71**	**0.80**	**0.78**	**0.82**

Discussion and Limitation

Based on several experiments conducted in the previous section, the SVM classification algorithm can detect words containing cyberbullying content with an F-measure level of 82 % better than other algorithms such as NB, LR, and DT. Although the experimental results show, the score value is not too different from the other algorithms. This ability is very influential from the correct data pre-processing technique [16], [17]. Suppose the pre-processing technique accomplishes appropriately and precisely. In that case, it will impact some features that lose meaning after the word segmentation process at the text pre-processing stage. The impact on the original semantics of information is not fully revealed [18]. For this reason, the selection and process of data preparation must be made carefully without neglecting the proper data cleaning process. If the dataset is not good, it will impact the model, prone to overfitting and bias in transferring the dataset to the model [19].

Machine learning can solve problems in sentiment classification and improve classification results by changing multi-dimensional word vectors containing n-grams to find out the classification features. This method is not only simple but also practical. However, the problem of the semantic relationship between texts and the relationship between words cannot reflect [20]. So the deep-based approach needs to be further investigated to be used and overcome some problems in engineering machine learning. Traditional Approaches to Classification NB, SVM, NBSVM, LDA, ANN, CNN, RecNNs, RNNs, and others that come out showed promising results. However, it cannot segment text in languages written with special characters (Chinese, Indian, Arabic, and others) written without spaces. The impact is less than optimal performance [21].

CONCLUSION

This paper has experimented with a machine learning method to detect cyberbullying text from Twitter automatically. Several textual features and machine learning algorithms are used to build classification models such as NB, DT, LR, and SVM. Experiments on collected tweets evaluate the quality of cyberbullying automatic detection, and the model shows that the SVM algorithm gets the best score compared to other algorithms, namely the average F-measure value of 82% and for the four criteria: with four criteria: accuracy, precision, recall, and F-measure. This study identified content on the Twitter social media network with 1065 data for processing test data and training data. From 1065 data, after distribution analysis, there are 638 data with non-bullying labels and 427 data with bullying labels. Then pre-processing techniques are used to improve the structure of the language. When the process sends to the classification model, the algorithm can identify and detect it correctly. The word insertion process uses the BOW model

with three features, namely unigram, bigram, and trigram. Because the text used in this study is limited, we plan to collect more texts by scraping technique and use more essential data. Thus, the learning process and training data can be maximized for the cyberbully detection process.

REFERENCES

[1] W. Uriawan, A. Wahana, D. Wulandari, W. Darmalaksana, and A. Rosihon, "Pearson Correlation Method and Web Scraping for Analysis of Islamic Content on Instagram Videos," in *2020 6th Int. Conf. Wirel. Telemat. (ICWT)*, 2020, pp. 1–6.

[2] M. A. Maulana, "The Effects of Anonymity, Psychological Needs and Cyber Victimization Toward Cyberbullying Behavior Among Adolescents in Cirebon City," in *3rd ASEAN Conf. Psychol. Couns. Humanit.*, 2017, pp. 42–51.

[3] H. Margono, X. Yi, and G. K. Raikundalia, "Mining Indonesian Cyber Bullying Patterns in Social Networks," in *Proc. Thirty-Seventh Australas. Comput. Sci. Conf.*, 2014, pp. 115–124.

[4] A. F. Hidayatullah and M. R. Ma'Arif, "Pre-processing tasks in Indonesian Twitter messages," in *J. Phys.: Conf. Ser.*, 2017, vol. 801, no. 1, p. 12072.

[5] A. Muzakir and R. A. Wulandari, "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," *Sci. J. Informatics*, vol. 3, no. 1, pp.19-26, 2016.

[6] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis," *Sci. J. Informatics*, vol. 6, no. 1, pp. 138–149, 2019.

[7] H. F. Fadli and A. F. Hidayatullah, "Identifikasi Cyberbullying Pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM," *AUTOMATA*, vol. 2, no. 1, 2021.

[8] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate Speech Detection: Challenges and Solutions," *PLoS One*, vol. 14, no. 8, p. e0221152, 2019.

[9] J. Seglow, "Hate Speech, Dignity and Self-Respect," *Ethical Theory Moral Pract.*, vol. 19, no. 5, pp. 1103–1116, 2016.

[10] K. Sreelakshmi, B. Premjith, and K. P. Soman, "Amrita CEN at HASOC 2019: Hate speech detection in roman and devanagiri scripted text," in *CEUR Workshop Proc.*, 2019, vol. 2517, pp. 366–369.

[11] S. Assimakopoulos, F. H. Baider, and S. Millar, *Online hate speech in the European Union: A discourse-analytic perspective*. Springer Nat., 2017.

[12] A. Muzakir and U. Ependi, "Model for Identification and Prediction of Leaf Patterns: Preliminary Study for Improvement," *Sci. J. Informatics*, vol. 8, no. 2, pp. 244–250, 2021.

[13] Y. N. Ifriza and M. Sam'an, "Performance Comparison Of Support Vector Machine and Gaussian Naive Bayes Classifier for Youtube Spam Comment Detection," *J. Soft Comput. Explor.*, vol. 2, no. 2, pp. 93–98, 2021.

[14] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding Bag-Of-Words Model: A Statistical Framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1, pp. 43–52, 2010.

[15] A. Sethy and B. Ramabhadran, "Bag-of-word Normalized N-Gram Models," 2008.

[16] U. Naseem, I. Razzak, and P. W. Eklund, "A Survey of Pre-Processing Techniques to Improve Short-Text Quality: A Case Study on Hate Speech Detection on Twitter," *Multimed. Tools Appl.*, 2020.

[17] Y. N. Ifriza and M. Sam'an, "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *J. Soft Comput. Explor.*, vol. 2, no. 2, pp. 93–98, 2021.

[18] J. Xie, B. Chen, X. Gu, F. Liang, and X. Xu, "Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification," *IEEE Access*, vol. 7, pp. 180558–180570, 2019.

[19] W. Yin and A. Zubiaga, "Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions," *PeerJ Comput. Sci.*, vol. 7, pp. 1–38, 2021.

[20] P. Wang, B. Xu, J. Xu, G. Tian, C. L. Liu, and H. Hao, "Semantic Expansion Using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification," *Neurocomputing*, vol. 174, pp. 806–814, 2016.

[21] W. Cui *et al.*, "Short Text Analysis Based on Dual Semantic Extension and Deep Hashing in Microblog," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 4, 2019.