Cyber-bullying Detection based on Machine Learning Method (Case Study: Instagram Comment Section)

Eri Eli lavindi[1], Edi Jaya Kusuma[2], Suko Tyas Pernanda[3], Roni Apriantoro[4]

[1,3,] Informatics Engineering, Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang, Indonesia
[4]Telecommunication Engineering, Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang, Indonesia
[2] Department of medical record and health information, Faculty of Health Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Abstract— Social media is popular communication platform for last decade. Social Platform such as Facebook, Instagram, and Twitter provide real-time and efficient way of communication overseas. The ease of using social media does not only provide positive benefits, but can also have a negative impact on its users. One of social media negative impact is cyber-bullying which define as a type of harassment through online media. The effect of cyber-bullying to the victim particularly is mental health disorder. Usually, being the victim of cyber-bullying can increase the stress and anxiety level, lower self-esteem, loneliness, sadness, and disappointment. This study evaluates the comment on Instagram post of Indonesia influencer to determine whether it classified as bullying or non-bullying. This study utilizes count vectorizer as feature extraction and compare several machine learning methods such as Naïve Bayes, SVM, and Random Forest. The evaluation result show that both Naïve Bayes and Random Forest are able achieve 77% accuracy. Moreover, Naïve Bayes method also generate higher percentage compared to other methods. This result indicate that Naïve Bayes are capable in detecting cyber-bullying comment in social media platform.

Keywords—Cyber-bullying, Machine Learning, Social media,Text Mining

1. Introduction

Survey conducted by the Association of Indonesian Internet Service Providers in 2021-2022 shows the number of internet active user in Indonesia reaches 77.02% of the total population [1]. Currently, internet became primary needs in daily activity as telecommunication facility. The development of internet also leads the emergence of various type of social media platforms [2]. These social media platforms such as Facebook, Instagram, Twitter, etc provide real-time and efficient way of communication overseas [3]. Recently, the used of social media platforms also provide useful environment as promotion media in supporting their users business.

The ease of using social media does not only provide positive benefits, but can also have a negative impact on its users. One of social media negative impact is cyber-bullying. Cyber-bullying is definition of harassment through online media [4]. The effect of cyber-bullying to the victim particularly is mental health disorder. Usually, being the victim of cyber-bullying can increase the stress and anxiety level, lower self-esteem, loneliness, sadness, and disappointment [5]. In some case, it leads the suicidal behaviour.

Recently, the developer of the social media platform releases a comment filter in order to prevent the potential of cyber-bullying. This comment filter works by detecting the comment that has intention to bullying, then remove or block it from user. This technique used technology approach such as data analyst and machine learning methods. There are several studies proposed the comment detection that contain cyber-bullying. Muhammed Ali Al-Garadi et al [4] provide the literature review regarding detection cyber-bullying studies by comparing the used of several machine learning methods such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, Association Rule Mining, and Rule-Based Algorithm. The conclusion of this study shows SVM method often used for cyber-bullying detection. Another study proposed by Andrea Pereraa et al [6] present the detection and prevention of cyber-bullying in Twitter platform using SVM and feature extraction Term Frequency-Inverse Document Frequency (TF-IDF). The 1000 textual data was collected from Twitter. The evaluation result shows the accuracy of proposed method is 74.50%, then the precision, recall, and F1 Score 74% respectively. Samar Almutiry et al [7] proposed cyber-bullying detection in Arabic using Arabic Sentiment Analysis (ASA) where this method combined with SVM. The dataset was compiled from twitter platform. The result indicates that using Light Stemmer can achieve 85.49% efficiency. However, when using Arabic Stemmer Khoja, the result show 85.38% efficiency. From these studies, it can be seen that the resource of dataset was gathered from Twitter. However, based on Digital Indonesia survey in 2022 [8], the percentages of internet user in Indonesia from ages of 16 to 64 who mostly used Instagram is 84.8% each month. Moreover, 22.9% of internet user from ages of 16 to 64 declare that Instagram is their favourite social media platform.

Therefore, this study proposed cyber-bullying detection in Instagram platform. The data conversion used Count Vectorizer method to normalize the form of data. Then, the result will be evaluated in several machine learning methods in order to find suitable method for cyber-bullying detection.

2. Data Set

This research utilize dataset provided by Kaggle.com. This dataset contains 650 data in form of textual collected from comment section in Instagram post of public influencer. This data has several attributes consists of Instagram name, comment, category, and posting date. Below can be seen the sample of data which used in this study.

Table 1. Comments Sample of Dataset in Both Categories (Non-Bullying and Bullying)

Comment	Category
"Kaka tidur yaa, udah pagi, gaboleh capek2"	Non-bullying
"makan nasi padang aja begini badannya"	Non-bullying
"Hai kak Isyana aku ngefans banget sama kak Isyana.aku paling suka lagu kak Isyana itu lagu tetap didalam jiwa"	Non-bullying
"Makin jelek aja anaknya, padahal ibu ayahnya cakep!"	Bullying
"Muka anak nya ko tua banget yaa.. GK ngegemesin GK ada lucu2nya"	Bullying
"Muka nya muka kolot wkwk bukan muka bayi2 lucu gt"	Bullying

3. Proposed Method

The proposed method of this study will consist of two phases which are preprocessing process and evaluation process. The detail information of each process can be seen in Figure 1.

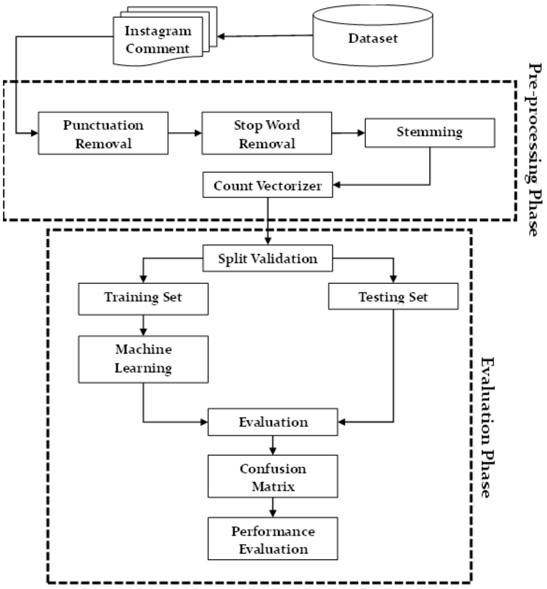

Figure. 1. Proposed Method of The Study

1.1 Pre-processing Phase

This phase describes the process of text processing which the purpose is to convert the form of dataset (textual form) into matrix dataset that possible to feed into machine learning algorithm [9]. There are several steps needed in order to convert this dataset which are:

1.1.1 Punctuation Removal

The purpose of punctuation removal is to eliminate the punctuation in the sentences of the comment which this character can be assumed by feature extraction algorithm as important feature [10]. The punctuation can increase the ambiguity or misunderstanding when conducting data training on the model.

1.1.2 Stop Word Removal

This step will remove several words which appear across all comment sentences in the dataset [11]. Commonly, the stop words can be classified as articles, conjunction, and pronouns.

1.1.3 Stemming

This process will convert the word in comment sentences into their root or base word [12]. Stemming is conducted to reduce the number of words indexed in the matrix dataset. Therefore, the number of indexed words in matrix dataset will be efficient.

1.1.4 Count Vectorizer

The count vectorizer will transform the data from textual form into vector based on the frequency of each words appear in the whole text [13]. The result of this step is in matrix form that ready to be evaluate in machine learning environment.

1.2 Evaluation Phase

1.2.1 Split Validation

Split validation is performed in order to split the dataset into training set and testing set, where training set will be used as data reference for training the model. Meanwhile, the testing set will be used as testing reference for the model.

1.2.2 Machine Learning Algorithm

In this study, there are several machine learning algorithms that use as classifier to detect the cyber-bullying comment, such as multinominal Naïve Bayes [14], SVM [15], and Random Forest [3]. Each algorithm will be evaluated in the same environment and compare their result to find which algorithm which has optimum performance.

1.2.3 Evaluation

The evaluation is performed in form of confusion matrix. The testing result of each models will be depicted into table form. From the confusion matrix table, it can be produced the parameters that indicate the performance of each model. These parameters consist of accuracy, precision, recall, and F1-score.

4. Experiment Result and Discussion

This research was conducted in order to determine which machine learning algorithm suitable in classifying cyber-bullying in social media dataset. This study used 650 data collected from comment section on Instagram post of Indonesia influencer. This dataset consists of two classes which are bullying and non-bullying. The preprocessing step was performed to transform the raw data (textual form) into executable data (matrix form). Then, this executable data was separated into two sets consist of training and testing set where the ratio of 75% and 25% respectively. After that, the training set was fed to the machine learning methods which in this study used Naïve Bayes, SVM, and Random Forest. The models obtained from training process was evaluate using confusion matrix to generate performance parameters such as accuracy, precision, recall, and F1-score[16][17].

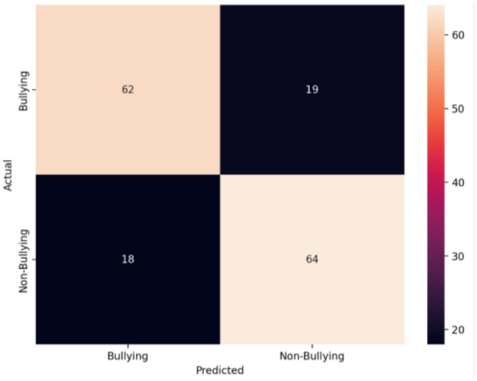

Figure. 2. The Confusion Matrix of Naïve Bayes Method

The confusion matrix result of Naïve Bayes model can be seen in Figure 2. The result identify that Naïve Bayes model can predict the actual data of bullying class as 62 data, then 64 data of non-bullying class are correctly predicted.

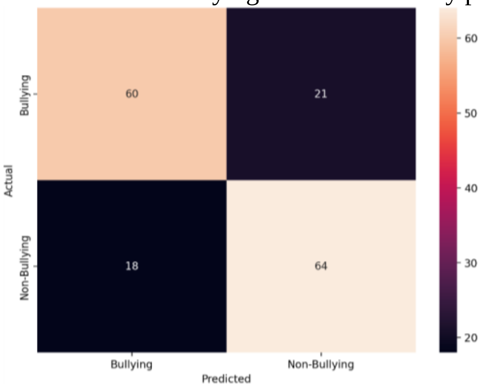

Figure. 3. The Confusion Matrix of SVM Method

Meanwhile, the evaluation result of SVM method described in Figure 3. The result show that the method can precisely predict the class bullying and non-bullying with 60 and 64 correct data respectively. However, the number of miss prediction is higher than Naïve Bayes method especially in Bullying class, which denoted that the Naïve Bayes has better sensitivity in predicting Bullying class. Moreover, the result of Random Forest method can be seen in Figure 4.

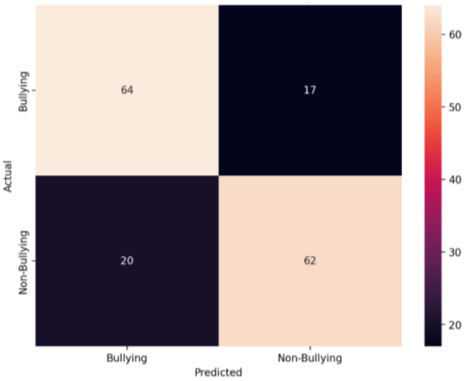

Figure. 4. The Confusion Matrix of Random Forest Method

The Random Forest method can achieve 64 and 62 data that predicted correctly. The prediction result of Random Forest has similarity to the Naïve Bayes method, even though there is slightly different result, particularly in non-bullying prediction result. Random Forest has smaller number in false prediction of bullying data, this mean that Random Forest has best sensitivity compare to other methods. However, in non-bullying class, Naïve Bayes and SVM outperform Random Forest in term of sensitivity.

Table 2.Comments Sample of Dataset in Both Categories (Non-Bullying and Bullying)

No	Model	Accuracy	Precision	Recall	F1-Score
1	Naïve Bayes	77%	77.5%	77.5%	77.5%
2	Support Vector Machine	76%	76.0%	76.0%	76.0%
3	Random Forest	77%	77.0%	77.5%	77.5%

In Table 2, it can be seen the performance parameter of each proposed model. The result show both Naïve bayes and Random Forest have similar result especially in accuracy, recall, and F1-score. The accuracy score indicates the capability of the model in predicting the correct label of both classes [18]. Then, the recall defines the percentage of correct predictions in non-bullying class within entire data predicted as non-bullying [19]. F1-score is the average of both precision and recall which have been weighted [20]. The different between Naïve Bayes and Random Forest result is the precision value where Naïve Bayes has slightly higher percentage than Random Forest. The higher precision result indicate that Naïve Bayes has more percentages of correct prediction data within entire data predicted as bullying class compared to the other methods.

5. Conclusion

Cyber-bullying is form of harassment conducted through electronic or online media. The effect of cyber-bullying to the victim particularly is mental health disorder. Usually, being the victim of cyber-bullying can increase the stress and anxiety level, lower self-esteem, loneliness, sadness, and disappointment. This study used Instagram comment dataset which consist of 650 data to determine whether it classified as bullying or non-bullying. This study utilizes count vectorizer as feature extraction and compare several machine learning methods such as Naïve Bayes, SVM, and Random Forest. The evaluation result denoted that both Naïve bayes and Random Forest have similar result especially in accuracy, recall, and F1-score. The different between Naïve Bayes and Random Forest result is the precision value where Naïve Bayes has slightly higher percentage than Random Forest. The higher precision result indicate that Naïve Bayes has more percentages of correct prediction data within entire data predicted as bullying class compared to the other methods.

References

[1] APJI, "Profil Internet Indonesia 2022," *Apji.or.Od*, no. June, 2022, [Online]. Available: apji.or.id

[2] P. D. Nugraheni, "The New Face of Cyberbullying in Indonesia: How can We Provide Justice to the Victims?," *Indones. J. Int. Clin. Leg. Educ.*, vol. 3, no. 1, pp. 57–76, 2021, doi: 10.15294/ijicle.v3i1.43153.

[3] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach," *J. Cyber Secur. Mobil.*, vol. 10, no. 4, pp. 745–774, 2021, doi: 10.13052/jcsm2245-1439.1046.

[4] T. Ige and S. Adewale, "AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naïve Bayes and Optimized Linear Support Vector Machine Interception of Cyberbully Contents in a Messaging System by Machine Learning Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 5–9, 2022, doi: 10.14569/IJACSA.2022.0130502.

[5] S. G. Handono, K. Laeheem, and R. Sittichai, "Factors related with cyberbullying among the youth of Jakarta, Indonesia," *Child. Youth Serv. Rev.*, vol. 99, no. August 2018, pp. 235–239, 2019, doi: 10.1016/j.childyouth.2019.02.012.

[6] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Comput. Sci.*, vol. 181, pp. 605–611, 2021, doi: 10.1016/j.procs.2021.01.207.

[7] S. Almutiry and M. Abdel Fattah, "Arabic CyberBullying Detection Using Arabic Sentiment Analysis," *Egypt. J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, 2021, doi: 10.21608/ejle.2021.50240.1017.

[8] S. Kemp, "Digital-2022-Indonesia-February-2022-v01_compressed.pdf." pp. 24–84, 2022. [Online]. Available: https://datareportal.com/reports/digital-2022-indonesia?msclkid=54849450ac3011eca46cf06ec644a888

[9] J. Cristian, V. Louise, and S. Koka, "Prediksi Keberhasilan Lamaran Pekerjaan Dengan Count Vectorizer dan Logistic Regression Abstak," vol. 4, pp. 16–25, 2022.

[10] R. Kustiawan, A. Adiwijaya, and M. D. Purbolaksono, "A Multi-label Classification on Topic of Hadith Verses in Indonesian Translation using CART and Bagging," *J. Media Inform. Budidarma*, vol. 6, no. 2, p. 868, 2022, doi: 10.30865/mib.v6i2.3787.

[11] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

[12] J. Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 1–46, 2016.

[13] A. Averina, H. Hadi, and J. Siswantoro, "Analisis Sentimen Multi-Kelas Untuk Film Berbasis Teks Ulasan Menggunakan Model Regresi Logistik," *Teknika*, vol. 11, no. 2, pp. 123–128, 2022, doi: 10.34148/teknika.v11i2.461.

[14] N. Rezaeian and G. Novikova, "Persian text classification using naive bayes algorithms and support vector machine algorithm," *Indones. J. Electr. Eng. Informatics*, vol. 8, no. 1, pp. 178–188, 2020, doi: 10.11591/ijeei.v8i1.1696.

[15] R. Kashef, "A boosted SVM classifier trained by incremental learning and decremental unlearning approach," *Expert Syst. Appl.*, vol. 167, p. 114154, 2021, doi: 10.1016/j.eswa.2020.114154.

[16] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," pp. 79–91, 2020, doi: 10.18653/v1/2020.eval4nlp-1.9.

[17] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," pp. 37–63, 2020, [Online]. Available: http://arxiv.org/abs/2010.16061

[18] J. A. Cottam, N. C. Heller, C. L. Ebsch, R. Deshmukh, P. MacKey, and G. Chin, "Evaluation of Alignment: Precision, Recall, Weighting and Limitations," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 2513–2519, 2020, doi: 10.1109/BigData50022.2020.9378064.

[19] E. Dritsas and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," *Big Data Cogn. Comput.*, vol. 6, no. 4, 2022, doi: 10.3390/bdcc6040139.

[20] R. Soleymani, E. Granger, and G. Fumera, “F-measure curves: A tool to visualize classifier performance under imbalance,” *Pattern Recognit.*, vol. 100, 2020, doi: 10.1016/j.patcog.2019.107146.