Cyberbullying Detection Modelling at Twitter Social Networking

Ika Yunida Anggraini1), Sucipto2), and Rini Indriati3)

1, 2,3)*Information System, University of Nusantara PGRI Kediri*

1ikayunida123@gmail.com
2sucipto@unpkediri.ac.id
3rini.indriati@unpkediri.ac.id

***Abstract*— Cybercrimes often happened in social networking sites. Cyber-bullying is a form of cybercrime that recently trended in one of popular social networking sites, Twitter. The practice of cyber-bullying on teenager can cause depression, murderer or suicidal thoughts and it needs a preventing action so it will not harmful to the victim. To prevent cyber-bullying a text mining modelling can be done to classify tweets on Twitter into two classes, bullying class and not bullying class. On this research we use Naïve Bayes Classifier with five stages of pre-processing : replace tokens, transform case, tokenization, filter stopwords and n-grams. The validation process on this research used 10-Fold Cross Validation. To evaluate the performance of the model a Confusion Matrix table is used. The model on 10-Fold Cross Validation phase works well with 77,88% of precision , 94,75% of recall and 82,50% of accuracy with +/-5,12% of standard deviation.**

***Keywords*— Modelling, Cyberbullying, Twitter.**

I. INTRODUCTION

A very rapid website development causes users of social networking sites like Facebook, Twitter, Instagram and Youtube increasing from year to year. Data from the Ministry of Communication and Information of the Republic of Indonesia (Kominfo) said users of social networking sites such as Twitter almost reached 20 million active users and ranked the top 5 Twitter users in the world [1]. Millennials spend a lot of their time on social networking sites and often spreading their personal information with friends so it can be seen by public. This causes a lot of crime on social networking sites. One type of crime on social media that often occurs is cyber bullying [2]. UNICEF (United Nations Children's Fund) revealed in Indonesia itself that in 2016 as many as 41-50 % of adolescents in the age range of 13-15 years had experienced cyber bullying [3].

Cyber-bullying or cyberbullying is an act of attacking, humiliating, or harming others intentionally and repeatedly on social media, messages, or other online means [4]. Cyber bullying is a public concern because the traditional and cyber bullying practices among teenagers can cause depression, suicide and attempted murder [5]. With the dangers of the effects of cyber siege, it is necessary to take precautionary measures so as not to cause harm to the victims. To detect cyber acts on Twitter, modelling can be done using text mining. In previous studies, sentiment analysis can be used to classify tweets containing abuse or bullying content into negative, neutral and positive sentiments [2]. Besides that, association rule algorithms like Apriory can be used to find patterns of bullying words in Indonesia [6].

In this study, text mining modelling was carried out using a classification algorithm, namely Naïve Bayes Classifier on the Rapid Miner Studio Community Edition software version 8.1.001. In classifying data tweets to detect cyber-abuse on Twitter social networking sites, data is classified into two classes, namely "Bullying" and "Not Bullying" classes. The "Bullying" class is a class of tweets that contain cyber bullying action, while the "Not Bullying" class is a class of tweets that is not a cyber-bullying action. This document is a template. An electronic copy can be downloaded from the journal website. For questions on paper guidelines, please contact the editor of journal as indicated on the journal website. Each paper Information about final paper submission is available from the conference website.

II. RESEARCH METHODS

The first process in this study is data collection. The data collection technique used is crawling data. Crawling data on Twitter is a process to retrieve or download data from a Twitter server with the help of Twitter's Application Programming Interface (API) in the form of user data and tweet data [7]. Next labeling for the result of crawling data will be done. Labeled or classed data is imported into the Rapid Miner environment. After that, an example filter is used to filter attributes with missing values and also remove duplicates to delete duplicates in the data. The next data is through five preprocessing stages, namely replace tokens, transform cases, tokenize, stopwords and n-gram filters. The replace tokens process is a process that is performed to replace the substring in each token that is specified using the Regular Expressions (RegEx) at the replace dictionary using the operator Replace Tokens [8]. Transform case is a preprocessing process that converts all letters to the data as desired, like all capital letters to become Latin, or vice versa [9].

The tokenization process is the process of cutting an item, both schematic elements (attributes) and attribute values, into atomic words (single words) that are done using delimiter [10]. In the tokenization process for word vector formation, term frequency technique is used. Term frequency is a method used to indicate the frequency of a term or word that appears in a document [11]. In Rapid Miner, term frequency is calculated from the number of frequency words in a document divided by the number of words. Then the normalized end word vector is calculated from term frequency divided by the root of the sum of all term frequencies [12]. Then the stopwords filter process is carried out. Stopwords are words that often appear to form a sentence but do not show information from a document. Examples are the words "are", "which" or other [10]. In the last preprocessing, the n-gram process is used to determine the probability of a word sequence (sequences of words) [13]. In this study 2-n or bigram was chosen because it was considered to fit the tweets data type which was limited to 240 characters.

After preprocessing, the next data is through the validation process using Cross Validation (K-Fold Validation). Cross Validation used in this study is 10-Fold Cross Validation which divides the data into 10 folds of the same size and in each fold will be tested with 9 subsets as training subsets and 1 subset as validation subset [11]. The next process is the formation of a model using the Naïve Bayes Classifier. Naïve Bayes Classifier is a data mining algorithm that uses statistical classifiers. This algorithm can predict the probability of membership in a class. The classification of Bayes applies the Bayes theorem. Bayes theorem was discovered by Thomas Bayes in the early 18th century. Bayes's theorem is formulated as eq. 1.

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \text{ .... (1)}$$

Where X is data with an unknown class. H is the class hypothesis of data X. P (H | X) is the probability of H based on condition X (posterior probability). P (H) is the probability of H (prior probability). P (X | H) is the probability of X based on condition H. And P (X) is the probability of X. In this study another bayesian classification approach is used, namely the Gaussian Naïve Bayes Classifier that uses a Gaussian distribution or normal distribution. The probability measure in normal distribution is presented in eq. 2.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \text{ .... (2)}$$

Where μ is the mean of the distribution, σ is the standard deviation of the mean, while σ^2 is a variant of the mean [11]. To avoid zero probability of words that have never appeared in the document, a smoothing process is carried out. Smoothing is used to harmonize probability estimates to produce a more accurate probability (eq. 3).

$$Pr_{add}(t_i \mid d_j) = \frac{f_{ij}+\lambda}{|d_j|+|V|\lambda} \text{ .... (3)}$$

Traditional additive smoothing can be stated as follows:

Where fij is the value in the attribute, dj is the number of words in the token and V is the number of classes. And if λ = 1, the smoothing is a Laplace smoothing or Laplace correction type [10]. The next process is the assessment of modeling performance using Confusion Matrix. Confusion Matrix or also called error matrix is a table that describes the performance or performance of an algorithm (Table I). Each column of the matrix represents the predicted class and the actual class. Evaluation in Confusion Matrix in this study uses the parameters of precision, recall and accuracy [14][15].

TABLE I
CONFUSION MATRIX

		Actual Class	
		Condition Positive	**Condition Negative**
Predicted Condition	**Predicted Condition Positive**	True Positive (TP)	False Positive (FP)
	Predicted Condition Negative	False Negative (FN)	True Negative (TN)

The results of precision are obtained from the calculation of the number of positive values classified correctly (True Positive) divided by the value of positive values that are classified correctly (True Positive) and the number of negative values that are incorrectly classified as positive (False Positive). Recall results are calculated from the number of positive values classified correctly (True Positive) divided by the number of positive values that are classified correctly (True Positive) and the false positive values are classified as negative (False Negative). While the accuracy result is calculated from the number of values classified correctly (True Positive and True Negative) divided by the number of all data [14].

III. RESULTS AND DISCUSSION

The validation process is a process of evaluating the performance of a model. The validation process in this study was done using Cross Validation with k-10 or also called 10-Fold Cross Validation (Fig. 1). The 10-Fold Cross Validation process will divide the data into 10 subsets. Each subset will experience iteration ten times so that each subset has the opportunity to become a training subset or validation subset. And the results of precision, recall and accuracy in the 10-Fold Cross Validation process are calculated from the average precision, recall and accuracy in each iteration performed.

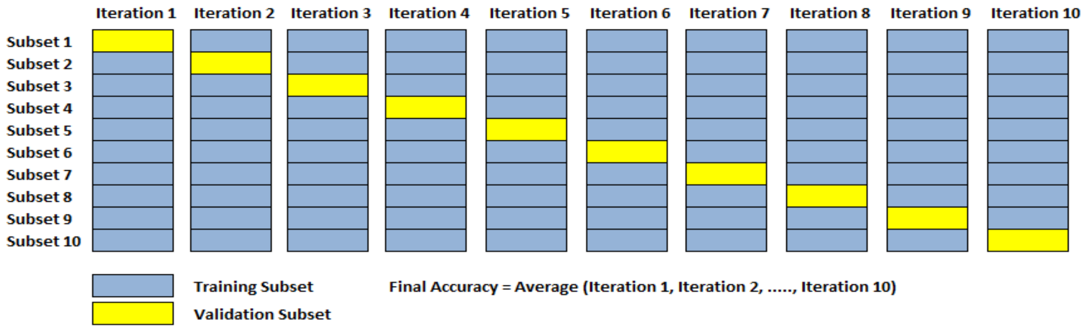

Fig.1 Cross Validation Process

In this study, 200 row datasets from crawling data will be used as research data (Table II). The dataset will be divided into ten subsets with the same amount of data as 20 rows in each subset as in table 2. The 10-Fold Cross Validation process will divide the data into a training subset of 180 rows and a validation subset of 20 rows.

TABLE II
RESEARCH DATASET

Dataset	**Data Total**
Subset1	20
Subset2	20
Subset3	20
Subset4	20
Subset5	20
Subset6	20
Subset7	20
Subset8	20
Subset9	20
Subset10	20

The validation process in this study was carried out using Cross Validation operators in RapidMiner software. The 10 k-folds parameter is selected to perform modeling validation using 10-Fold Cross Validation. The validation process with 10-Fold Cross Validation models in the RapidMiner software as in the Fig. 2.

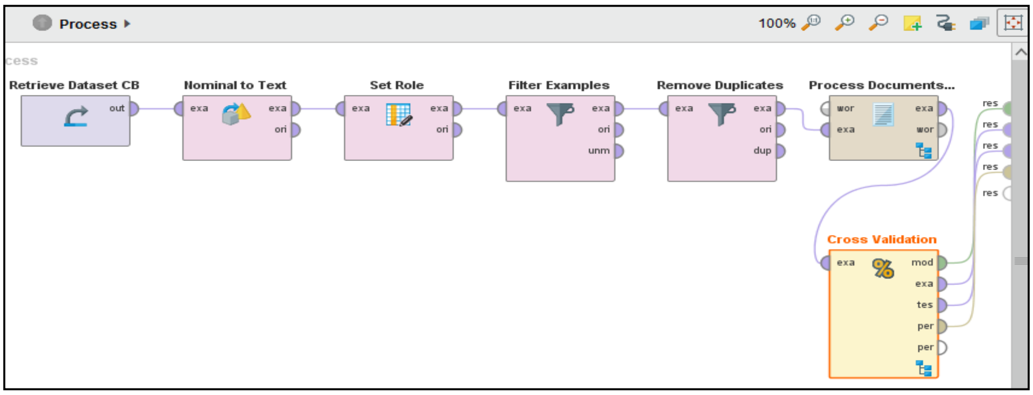

Fig.2 Cross Validation Process in RapidMiner

The 10-Fold Cross Validation process in the RapidMiner software has two sub-processes, namely the training sub-process and the testing sub-process (Fig. 3). The training sub-process functions to conduct the model training process. In the training sub-process, the Naïve Bayes operator is used to do the modeling. Furthermore, the Laplace correction parameter option is added to avoid zero probability of attribute values that have never appeared before. Besides that Laplace's correction is used so that the classification of tweets using Naïve Bayes becomes more accurate.

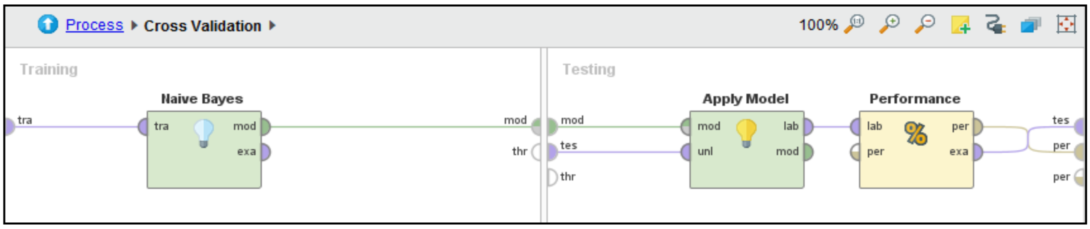

Fig. 3 Sub-process of Cross Validation in RapidMiner

The training model that has been formed in the training sub-process will be applied to the testing sub-process using the Apply Model operator. The operator will divide the data into a training subset and a validation subset. Furthermore, the performance of the model is assessed by the Performance (Binominal Classification) operator in the testing sub-process. The operator is used to assess performance in two class classification models or binominal classification. After doing the 10-Fold Cross Validation process, we get the results of precision, recall and accuracy for each subset as in Table III.

TABLE III
TABLE OF PRECISION, RECALL AND ACCURACY RESULTS IN 10-FOLD VALIDATION

Iteration	**Data Total**		***Precision***	***Recall***	***Accuracy***
	Learning	***Testing***			
1	180	20	66.67%	100%	80%
2	180	20	64.29%	90%	70%
3	180	20	84.62%	100%	90%
4	180	20	78.57%	100%	85%
5	180	20	84.62%	91.67%	85%
6	180	20	73.33%	100%	80%

7	180	20	80%	88.89%	85%
8	180	20	66.67%	100%	80%
9	180	20	100%	76.92%	85%
10	180	20	80%	100%	85%

Table III above shows the results of precision in 10-Fold Cross Validation which varies in the range of 64.29% to 100%. The biggest precision is on iteration 9 with 100% result, while the lowest precision is on iteration 2 with 64.29%. The results of the precision of the ten iterations resulted in an average precision of 77.88% and a standard deviation of +/- 10.23%.

Recall results also show figures that vary in the range of 76.92 %% to 100%. The lowest recall is in iteration 9 with a result of 76.92%. While the largest recall is in 6 iterations with recall results of 100%, namely on iterations 1, 3, 4, 6, 8 and 10. The recall results from the ten iterations resulted in an average recall of 94.75% of the standard and standard deviation of +/- 7.41%. Furthermore, the accuracy results in 10-Fold Cross Validation have a value range of 70% to 90%. The biggest accuracy is on iteration 3 with 90% result, while the lowest accuracy is on iteration 2 with 70% result. The result of accuracy of the ten iterations results in an average accuracy of 82.50% and a standard deviation of +/- 5.12% (Fig. 4).

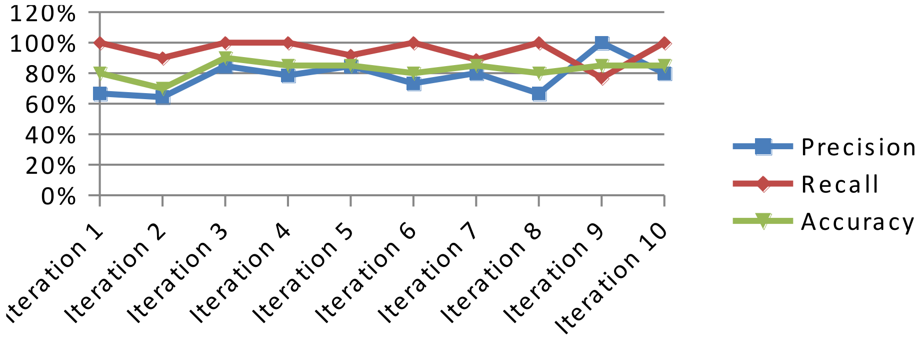

Fig. 4 Graphic of Each Iteration in 10-Fold Cross Validation

The average precision value is 77.88%, recall of 94.75% and accuracy of 82.50% which is a result that can be said to be high enough to show that the Naïve Bayes Classifier can work well on modeling detection of cyber bullying on Twitter social networks. Whereas the low standard deviation value in precision, recall and accuracy of the ten iterations in 10-Fold Validation in this study shows that the cyber detection model is a stable model. This is evidenced by the results of a range or a small range of values from the best and worst cases in 10-fold which shows the quality of predictions[16].

IV. CONCLUSIONS

Based on the results of research conducted by researchers, it can be concluded that detection of cyber bullying on Twitter social networks can be done with several techniques. First, data is collected through a crawling data process. Second, the data selection process, data cleaning and preprocessing are carried out to prepare the data in the mining process. Third, classification is done using the Naïve Bayes Classifier. The results of the modeling process of cyber detection in the 10-Fold Cross Validation process have an average precision of 77.88%, a recall of 94.75% and an accuracy of 82.50% with a standard deviation of accuracy of +/- 5.12 % that shows the model is a stable model. In this study there are still many shortcomings so that it needs to be developed in the future by using other text mining algorithms and various features to find the best model in modeling cyber detection.

REFERENCES

[1] www.kominfo.go.id, “Kominfo : Pengguna Internet di Indonesia 63 Juta Orang,” 2017. [Daring]. Tersedia pada: http://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker#.VR_pF2OY6tE. [Diakses: 01-Jul-2018].

[2] H. Sanchez dan K. Shreyas, “Twitter Bullying Detection,” California, 2011.

[3] T. Viva, “1 dari 4 Remaja Pernah Alami Perundungan di Dunia Maya,” *Gaya Hidup VIVA*, 2017. [Daring]. Tersedia pada: https://www.viva.co.id/gaya-hidup/parenting/963329-1-dari-4-remaja-pernah-alami-perundungan-di-dunia-maya. [Diakses: 01-Jul-2018].

[4] J. W. Patchin dan S. Hinduja, *Words Wound : Delete Cyberbullying and Make Kindness Go Viral*. Minneapolis: Free Spirit Publishing, 2014.

[5] S. Hinduja dan J. W. Patchin, “Cyberbullying and Suicide,” 2010.

[6] H. Margono, X. Yi, dan G. K. Raikundalia, “Mining Indonesian Cyber Bullying Patterns in Social Networks,” in *Proceedings of Thirty-Seventh Australasian Computer Science Conference (ACSC 2014)*, 2014, no. ACSC, hal. 115–124.

[7] J. E. Sembodo, E. B. Setiawan, dan Z. K. A. Baizal, “Data Crawling Otomatis pada Twitter,” in *IND. SYMPOSIUM ON COMPUTING*, 2016, no. September, hal. 11–16.

[8] RapidMiner, “Replace (Dictionary),” 2018. [Daring]. Tersedia pada: https://docs.rapidminer.com/latest/studio/operators/blending/values/replace_dictionary.html. [Diakses: 01-Jul-2018].

[9] A. Go, R. Bhayani, dan L. Huang, “Twitter Sentiment Classification using Distant Supervision,” USA, 2009.

[10] B. Liu, *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*, Second Edi. Berlin: Springer, 2011.

[11] J. Han dan M. Kamber, *Data Mining : Concepts and Techniques*, 2nd Editio. San Francisco: Morgan Kaufmann Publishers, 2006.

[12] S. Genzer, “Term Frequencies and TF-IDF: How are these calculated?,” *RapidMiner Knowledge Base*, 2018. [Daring]. Tersedia pada: community.rapidminer.com/t5/RapidMiner-Text-Analytics-Web/Term-Frequencies-and-TF-IDF-How-are-these-calculated/ta-p/46333. [Diakses: 01-Jul-2018].

[13] D. Jurafsky dan J. H. Martin, *Speech and Language Processing*. New Jersey: Prentice Hall, 2009.

[14] F. Gorunescu, *Data Mining : Concepts, Models and Techniques*. Berlin: Springer, 2011.

[15] Sucipto, Kusrini, and E. L. Taufiq, “Classification method of multi-class on C4.5 algorithm for fish diseases,” in Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment, 2016, pp. 5–9.

[16] S. Sucipto, “Analisa Hasil Rekomendasi Pembimbing Menggunakan Multi-Attribute Dengan Metode Weighted Product,” Fountain Informatics J., vol. 2, no. 1, p. 27, May 2017.