Advances in Social Science, Education and Humanities Research, volume 512
Proceedings of the 1st International Conference on Folklore, Language, Education and Exhibition (ICOFLEX 2019)

Comparison of the Performance of the k-Nearest Neighbor, Naïve Bayes Classifier and Support Vector Machine Algorithm With SMOTE for Classification of Bully Behavior on the WhatsApp Messenger Application

Irwansyah Saputra[1*], Puput Irfansyah[2], Erlando Doni Sirait[3], Dwi Dani Apriyani[4], Michael Sonny[5]

[1]*Computer Science STMIK Nusa Mandiri, Jakarta, Indonesia*
[2]*Department of Informatic Unniversitas Indraprasta PGRI, Jakarta, Indonesia*
[3]*Department of Informatic Unniversitas Indraprasta PGRI, Jakarta, Indonesia*
[4]*Department of Informatic Unniversitas Indraprasta PGRI, Jakarta, Indonesia*
[5]*Department of Informatic Unniversitas Indraprasta PGRI, Jakarta, Indonesia*
[*]*Corresponding author. Email: 14002085@nusamandiri.ac.id*

ABSTRACT
WhatsApp is the most popular messaging application in Indonesia. This causes the emergence of cyberbullying behavior by its users. Cyberbullying is a dangerous problem because it has a very serious impact on the victim's psyche such as feelings of hurt and disappointment. This study aims to classify WhatsApp chat into two classes, namely, bully and not bully. The classification algorithms used are k-NN, NBC, and SVM with SMOTE. The results show that the SVM algorithm with SMOTE is better at solving this case with an accuracy of 83,57%.

Keywords: *Cyberbullying, WhatsApp, k-NN, NBC, SVM, SMOTE*

1. INTRODUCTION

Internet users in Indonesia are more than 132 million people in 2017 [1]. Messaging applications on smartphones such as WhatsApp Messenger also experienced a significant increase in users, namely 35.8 million users, and became the most popular messaging application in Indonesia. The increased use of these applications creates a new form of attack known as cyberbullying. Cyberbullying is a behavior of aggression which refers to bullying behavior carried out by a person through social media such as the web, SMS, social networks, chat rooms, etc. [2].

Cyberbullying is a dangerous problem because it has a very serious impact on the victim's psyche such as feelings of hurt and disappointment. Cyberbullying on WhatsApp is prone to do because the increasing number of users of the application are of various ages [3]. Previous research has detected aggression and bullying on Twitter based on content features and profile network embeds. This research focuses on phenomena that occur in social networking media Twitter. This social networking media has several barriers to detecting negative behavior due to short tweets, lots of spam, and complicated grammar which makes it more difficult for machine processing to process natural language, extract text-based attributes, and characterize Twitter users' interactions. In this study, researchers explored the characteristics of Twitter users concerning content and network embeds such as following & followers and utilized attributes with machine learning classifications to automatically detect Twitter aggressors and bullies. [4].

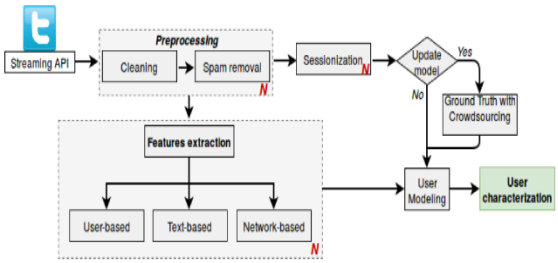

Figure 1 Problem Solving Method

The methods used in data analysis, labeling, and classification of up to millions of tweets, and the modeling algorithm built by the Random Forest classifier can distinguish between normal, aggressive, and bullying users with high accuracy up to> 91%[4].

On the other hand, there is research conducted regarding the introduction of sarcasm statements on the WhatsApp Group in Indonesian [5]. Sarcasm is an expression of annoyance, criticism, ridicule using harsh words that are meant to offend someone or something. Example of sarcasm statement: "I never forget someone's face, but in your case, I would love to make an exception" [5]. Sarcastic statements can cause difficulties for many Natural Language Processing (NLP) based systems. The text of the conversation data is retrieved from the WhatsApp group using a smartphone, from the Settings menu-Email Chat-(select No Media) because this research will not discuss anything related to the media, then select the email address you want to send the data to. The data provided by WhatsApp are text files. The data retrieved has a certain time interval, depending on the number of conversation lines and the number of group members as a sample set. Next, a feature vector is built for each statement example in the sample set and uses it to build a classification model. The features are extracted using various components of words and produce several kinds of sarcasm.

The set of words and statements executed for the experiment were manually checked and labeled. The method proposed is pattern-based. This method uses several feature sets to classify sentences (statements) into two parts, positive words and negative words. The second feature uses punctuation patterns, several exclamation marks, question marks, dots, all capital words, quotation marks, and the number of vowels repeated more than two times. Question marks and exclamation marks are considered possible indicators of sarcasm. But punctuation serves as the weakest predictor because the number of signs is only found in a very small number. The third feature is synthetic and semantic features. This feature calculates the value of words that are not common (including English or other language words except Indonesian), uncommon words such as the word "heeey" in statement number 3, presence of common sarcastic expressions, number of interjections and expressions. laugh. These features produced weak predictors as well, as not many signs were found. The accuracy results obtained were inadequate because this study only used data sampling and manual analysis experiments [5].

2. METHOD

The research method used in this study uses the KDD (Knowledge Discovery in Database) model. KDD is a method proposed by Fayyad in 1996. KDD is the process of extracting new information and knowledge from large databases [6]. The KDD process uses data mining methods to extract what is considered knowledge following measurement specifications and thresholds, the use of a database is required to perform data pre-processing subsampling and transformation of the database. Habibullah stated that in general, the KDD method proposed by Fayyad has 5 process stages to arrive at its goal, namely Knowledge Discovery Goals, Integration, Pre-processing Data mining, Interpretation, or Evaluation [7].

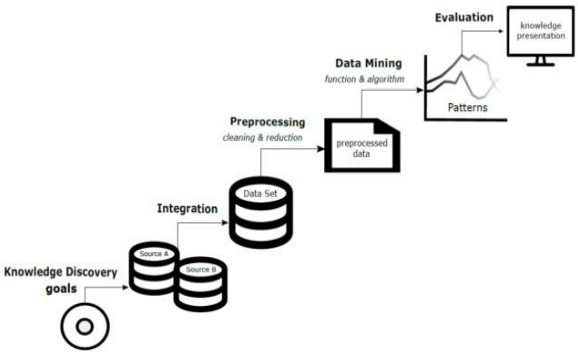

Figure 2 KDD Method Process Flow

2.1. Knowledge Discovery Goals

The first stage in working on a data mining activity is knowing the goals of the organization based on the problems it has. This study aims to classify the test data into predetermined classes, namely bully and not being bullied. The testing process is carried out using three classification algorithms, namely k-NN, NBC, and SVM.

2.2. Data Integration

By the objectives of data mining activities, the origin of data sources will be determined, collected, and combined into target data. If the application domain is large enough, then the target data can be in the form of a data warehouse or data mart. Usually, not all data attributes will be used so that the data can be selected based on only the relevant attribute subset so that it becomes a dataset.

The dataset used in this study is in the form of text chat in the WhatsApp group "Postgraduate 2017". The dataset consists of text and class attributes which will be labeled using the crowdsourcing technique with 21 participants/respondents. Labels consist of two types, namely bully and non-bully.

2.3. Data Preprocessing

The resulting dataset is often raw and lacks quality, for example, there are missing values, incorrect input values, and inconsistencies. As a result, it is necessary to pre-process data first. The cleaning process includes eliminating duplicate data, filling in / disposing of missing data, correcting inconsistent data, and correcting typos.

The dataset is cleaned using the Text Preprocessing technique which consists of tokenization (removing symbols, the punctuation or special characters or non-letters), Normalization Indonesian Slang (converting non-standard words into standard), stemming (removing affixes), stop words (removing meaningless words, discarding missing data, correcting inconsistent data, and correcting typos) and transforming the word “no / no” to

eliminate ambiguous words such as not bad, not greedy that have positive meanings.

2.4. Data mining

This process is the core of the knowledge extraction process from data. The algorithm will be chosen according to the objectives of the data mining activities that have been determined in the first stage.

Testing the dataset in this study was carried out using three classification algorithms, namely k-NN, NBC, and SVM which will be implemented using the RapidMiner version 9.1.0 tool.

2.5. Knowledge Evaluation

This study compares the accuracy level between the k-NN, NBC, and SVM algorithm approaches which are evaluated using precision, recall, and F-measure. The explanation of the evaluation results can be described as follows [8]

2.5.1. *Precision*

Precision is the level of accuracy between the information requested by the user and the answers given by the system. The precision formula is as follows:

$$p = \frac{tp}{tp+fp} \quad \text{(1)}$$

2.5.2. *Recall*

A recall is the success rate of the system in recovering information. The Recall formula is as follows:

$$r = \frac{tp}{tp+fn} \quad \text{(2)}$$

2.5.3. F-measure

F-measure is one of the evaluation calculations in information retrieval that combines recall and precision.

$$f = 1/\left(a\frac{1}{p} + (1-a)\frac{1}{r}\right) \quad \text{(3)}$$

3. RESULTS

3.1. Knowledge Discovery Goals

This study aims to classify the test data into predetermined classes, namely bully and not being bullied. Cyberbullying is a behavior of aggression which refers to bullying behavior carried out by someone through social media such as the web, SMS, social networks, chat rooms, and others. Cyberbullying is a dangerous problem because it has a very serious impact on the victim's psyche such as feelings of hurt and disappointment. Cyberbullying activities include various types of negative activities such as flaming (sending angry, rude, and vulgar messages), harassment (repeatedly sending offensive messages), cyberstalking (repeatedly sending harmful threats or messages that are very intimidating), denigration (posting statements untrue or cruel), impersonation (pretending to be someone else to make the person look bad or in danger), outing and trickery (posting things that contain personal or sensitive information about other people or forwarding private messages and or engaging in tricks to collect embarrassing information and spread it), exclusion (intentionally removing someone from an online group). Although cyberbullying does not involve personal contact between the perpetrator and the victim, it can be psychologically and emotionally damaging to the victim.

This explanation reveals the importance of detecting cyberbullying in messaging applications such as WhatsApp to minimize the number of cyberbullying victims. The cyberbullying detection process is carried out using three classification algorithms, namely k-NN, NBC, and SVM.

3.2. Data Integration

This stage describes the source from which the dataset can be collected. The raw data that will be used as a dataset has a chat text type. The data was collected from the WhatsApp group "Postgraduate 2017". There are 18 active members in the group. All group members are postgraduate students of STMIK Nusa Mandiri class 2018-2019. The dataset consists of text and class attributes with a total of 1000 records that will be labeled using the crowdsourcing technique with 21 participants. The label consists of two parts, namely Bully and Not being bullied. After experiencing data cleaning, the remaining 380 records consisted of 204 texts that were bullied, and 176 were recorded with no category. All records that have been labeled will be grouped into one dataset in the form of a comma-separated value (* .csv).

3.3. Data Preprocessing

This stage aims to clean data from noise. The dataset is cleaned using the Text Preprocessing technique which is carried out using the Gata Framework [9] and use the Text Preprocessing feature provided by Rapid Miner. The text preprocessing stages used can be seen in Figure 3.

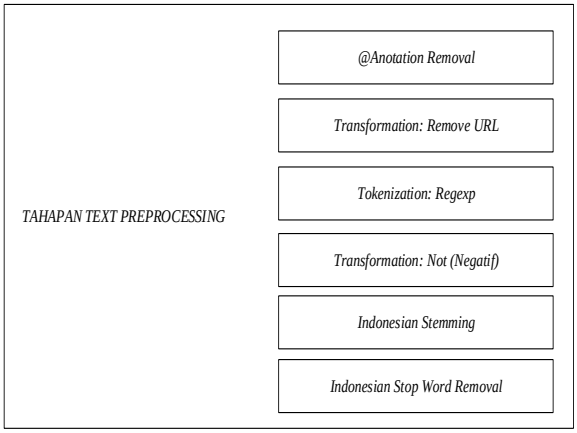

Figure 3 Stages of Text Preprocessing

3.4. Tokenization

The tokenization stage is a process carried out to eliminate symbols, the punctuation of special characters, non-letter characters from the dataset.

Table 1 Comparison during the Tokenization Process

Process	**Texts**
Text before tokenization	Ya bagaimana baiknya, kalau ada masukan ya silahkan. Kalau mau rubah jadwal ya monggoh. Saya minggu tidak bisa kalau harus kuliah, karena minggu pekerjaan rumah sudah menunggu.
Text after tokenization	Ya bagaimana baiknya kalau ada masukan ya silahkan Kalau mau rubah jadwal ya monggoh Saya minggu tidak bisa kalau harus kuliah karena minggu pekerjaan rumah sudah menunggu

3.4.1. Normalization Indonesian Slang

This stage serves to convert non-standard words into standard ones. Like the word "ngeselin", the real meaning is to make you upset.

Table 2 Comparison when the Indonesian Slang Normalization Process

Process	Texts
Text before Normalization Indonesia Slang	Utk jurnal bisa dimana saja kampusnya Klo mau cari yg ga terlalu mahal coba di kampus swasta banyak tuh Utk periode februari terbit
Text after Normalization of Indonesia Slang	untuk membuat jurnal dapat memilih kampus apa saja Jika ingin mencari yang tidak terlalu mahal dapat mencoba di kampus-kampus swasta disana lebih banyak Untuk diterbitkan pada periode bulan februari

3.4.2. Transformation Not (negative)

This stage aims to eliminate ambiguous words such as words that are not bad, not greedy which have positive meanings.

Table 3 Comparison during the Not Transformation Process (negative)

Process	Texts
Text before Transformation Not (Negative)	ya sudah kalau tidak sempat revisi Cek kembali format penulisannya Kirim kembali kalau sudah pasti Ditunggu
Text after Not Transformation (Negative)	ya sudah kalau tidak_sempat revisi Cek kembali format penulisannya Kirim kembali kalau sudah pasti Ditunggu

3.4.3. Stopword Removal

Stopword removal is a step to eliminate irrelevant or meaningless words. The data dictionary becomes a reference for a word including the stop word or not. The data dictionary is in a web-based program that can be accessed online on the Gata Framework website [9].

Table 4 Comparison when the Stopword Process

Process	Texts
Text before Stopword	mungkin pemerintah harus datang ke lapangan untuk mengetahui seperti apa kehidupan di lapangan jangan_diam saja di kursi empuk hanya bisa mengatur sana dan sini dan membuat peraturan yang tidak_sesuai dengan kenyataan

Text after Stopword	pemerintah lapangan kehidupan lapangan Jangan diam kursi empuk mengatur peraturan tidak_sesuai kenyataan

3.4.4. Stemming

Stemming is a step to restore a word from its root (root) in other terms removing the affixes contained in the word.

Table 5 Comparison of Text during the Stemming process

Process	Texts
Text before the stemming process	semangat Masih seminggu dinilai diperbaiki kurangnya Kita buatkan buku Semuanya semangat lulus
Text after the stemming process	semangat masih minggu nilai baik kurang kita buat buku semua semangat lulus

3.4.5. Data mining

This stage involves a classification algorithm that is built according to the initial objectives of the study. Testing the dataset to classify bully and non-bullying classes in this study was carried out using three classification algorithms, namely k-NN, NBC, and SVM which were implemented with the RapidMiner version 9.1.0 tool. The proposed Rapid Miner design is as follows,

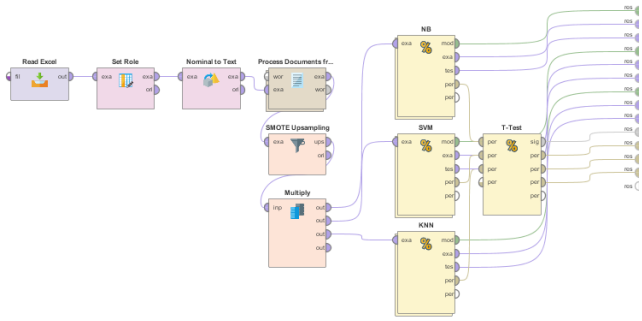

Figure 4. Model design of the KNN, NBC, and SVM

3.4.6. Knowledge Evaluation

After the data mining model is obtained, the next process is to compare the accuracy level between the k-NN, NBC, and SVM algorithm approaches which are evaluated using 10-fold cross-validation.

3.5. The Results of Test on k-NN, NBC dan SVM Methods

The design of the testing process for the k-NN method model used can be seen in Figure 5.

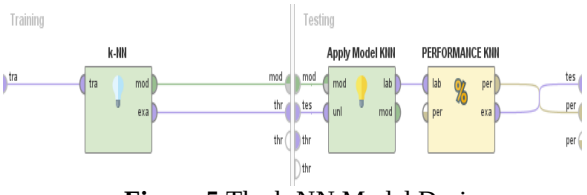

Figure 5 The k-NN Model Design

The design of the NBC model testing process used can be seen in Figure 6.

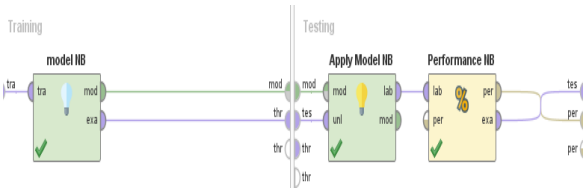

Figure 6 NBC Model Design

The design of the testing process for the SVM model used can be seen in Figure 7.

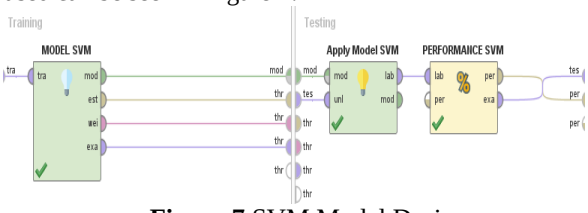

Figure 7 SVM Model Design

Figures 5, 6, and 7 are a detailed explanation of the process design shown in Figure 4. The data used for the validation process consists of two types, namely data with a bully class of 204 records and data with a non-bullying class of 176 records. Before use, all data has gone through the preprocessing process so that the data is clean from noise and suitable for use. The number of k used in the k-NN algorithm is k = 3.

Furthermore, there is a Confusion matrix table which contains information about the actual classification and predictions made by the classification system. The system performance is generally evaluated using data in a matrix. The following table shows the Confusion matrix for classifications that have positive and negative classes [10]

Table 6 *Confusion matrix*

		Actual	
		Negative	***Positive***
Predicted	*Negative*	**a**	**b**
	Positive	**c**	**d**

Information:

a is the number of correct predictions that an instance is negative;
b is the number of incorrect predictions that an instance is positive;
c is the number of incorrect predictions for which an example is negative; and
d is the number of correct predictions that an instance will be positive.

3.6. Algoritm Evaluation of k-NN, NBC and SVM

3.6.1. *Confusion matrix*

In addition to the ROC curve, accuracy measurements are also carried out using a Confusion matrix. The following is the calculation of accuracy for the three algorithms with SMOTE. The accuracy of the k-NN algorithm can be seen in table 7.

Table 7 Confusion matrix k-NN algorithm

Accuracy: 78,91% +/- 4,69%			
	True Tidak	**True *Bully***	**Class *Precision***
Pred. Tidak	204	86	70,34%
Pred. Bully	0	118	100,00%
Class Recall	100,00%	57,84%	

The accuracy obtained is 78.91% of the 204 bully class data and 176 of the non-bullying class data in the "Postgraduate 2017" group chat text on the WhatsApp Messenger application.

Bully class data under predictions is 169 records. No bullied data included in prediction "not bully", namely 35 records. The non-bullied data included in the bully prediction, namely 36 records and the non-bullied data which matched the prediction of 140 records.

Furthermore, the results of the accuracy calculation for the NBC algorithm can be seen in table 8.

Table 8 Confusion matrix NBC algorithm

Accuracy: 82,59% +/- 2,81%			
	True Not Bully	**True *Bully***	**Class *Precision***
Pred. Not Bully	160	27	85,56%
Pred. *Bully*	44	177	80,09%
Class *Recall*	78,43%	86,76%	

The accuracy obtained is 82.59% of the 204 bully class data and 176 the non-bullying data on the group chat text "Postgraduate 2017" on the WhatsApp Messenger application.

Bully class data under the predictions were 174 records. No bullied data included in the prediction No, that is 30 records. The non-bullied data included in the bully prediction, namely 50 records and the non-bullying data which matched the prediction of 126 records.

Finally, the results of the accuracy calculation for the SVM algorithm can be seen in table 9.

Table 9. SVM algorithm confusion matrix

Accuracy: 83,57% +/- 7,34%			
	True Tidak	**True *Bully***	**Class *Precision***
Pred. Tidak	175	38	82,16%
Pred. *Bully*	29	166	85,13%
Class *Recall*	85,78%	81,37%	

The accuracy obtained is 83.57% of the 204 bully class data and 176 the non-bullying data in the group chat text "Postgraduate 2017" on the WhatsApp Messenger application.

Bully class data under predictions is 185 records. No bullied data included in the prediction No, that is 19 records. The non-bullied data included in the bully prediction, namely 51 records and the non-bullied data which matched the predictions of 125 records.

A summary of the results of calculating the accuracy of the three algorithms can be seen in table 10.

Table 10. Summary of the results of the accuracy of k-NN, NBC, and SVM

Algorithm	**Accuracy**
k-NN + *SMOTE*	78,91%
NBC + *SMOTE*	82,59%
SVM + *SMOTE*	**83,57%**

4. CONCLUSIONS

The increased use of the WhatsApp application creates a new form of attack known as cyberbullying. This study uses a feature provided by WhatsApp to export group conversations. The dataset is the chat text of the "Postgraduate 2017" group. After the dataset is determined, the next process is to clean up the noise or noise data using Text Preprocessing, then labeling the bully and non-bullying classes using a crowdsourcing technique with 21 participants. Next, carry out the classification process for the dataset by comparing the performance of the three

algorithms, namely k-NN, NBC, and SVM in solving the problem of classification of bully behavior in chat texts on the WhatsApp Messenger application. The test results show that the accuracy possessed by the k-NN, NBC, and SVM algorithms is 81.32%, 78.95%, and 81.58%, respectively. This test shows that the SVM algorithm is better than the other two algorithms.

REFERENCES

[1] A. Prabowo, "Pengguna Ponsel Indonesia Mencapai 142% dari Populasi," 2017. [Online]. Available: https://databoks.katadata.co.id/datapublish/2017/08/29/pengguna-ponsel-indonesia-mencapai-142-dari-populasi.

[2] C. D. Marcum, G. E. Higgins, T. L. Freiburger, and M. L. Ricketts, "Battle of the sexes: An examination of male and female cyberbullying," *Int. J. Cyber Criminol.*, vol. 6, no. 1, pp. 904–911, 2012.

[3] A. Pingit, "WhatsApp Naikkan Batas Usia Pengguna Menjadi 16 Tahun," 2018. [Online]. Available: https://katadata.co.id/berita/2018/04/27/whatsapp-naikkan-batas-usia-pengguna-dari-menjadi-16-tahun.

[4] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Detecting Aggressors and Bullies on Twitter," in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 2017, pp. 767–768.

[5] R. Afiyati, E. Winarko, and A. Cherid, "Recognizing the sarcastic statement on WhatsApp Group with Indonesian language text," *2017 Int. Conf. Broadband Commun. Wirel. Sensors Powering, BCWSP 2017*, vol. 2018-Janua, no. May, pp. 1–6, 2018.

[6] F. S. Process, G. J. Williams, and Z. Huang, "Modelling the KDD Process," *Proc Centre for Software Reliability Conference on Measurement for Software Control and Assurance*, no. June. pp. 1–8, 1987.

[7] H. Akbar, "Ingin Terapkan Data Mining? Ini Tahapannya," 2017. [Online]. Available: https://mti.binus.ac.id/2017/12/05/ingin-terapkan-data-mining-ini-tahapannya/.

[8] M. Maragoudakis, N. Fakotakis, and G. Kokkinakis, "A Bayesian Model for Shallow Syntactic Parsing of Natural Language Texts," no. January 2016, 2016.

[9] Windu Gata, "Text Mining Program," 2018. [Online]. Available: http://www.gataframework.com/textmining.

[10] F. Provost and T. Fawcett, "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions," in *THE THIRD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, 1997, pp. 43–48.