2017 1st International Conference on Informatics and Computational Sciences (ICICoS)

Cyberbullying Classification using Text Mining

Noviantho, Sani Muhamad Isa, Livia Ashianti
Computer Science Department,
BINUS Graduate Program - Master of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
noviantho.hong@gmail.com, sani.m.isa@binus.ac.id, liviaashianti@gmail.com

***Abstract*—Cyberbully is a misuse of technology advantage to bully a person. Cyberbully and its impact have occurred around the world and now the number of cases are increasing. Cyberbullying detection is very important because the online information is too large so it is not possible to be tracked by humans. The purpose of this research is to construct a classification model with optimal accuracy in identifying cyberbully conversation using Naive Bayes method and Support Vector Machine (SVM) then applying n-gram 1 to 5 for the number of class 2, 4, and 11 for each method. Naive Bayes yields an average accuracy of 92.81%, SVM with a poly kernel yields an average accuracy of 97.11%. It can be concluded that SVM with poly kernel yields higher accuracy than SVM with other kernels, Naive Bayes, and Kelly Reynolds research method of decision tree (J48) and k-NN.**

Keywords—cyberbully, naive bayes, support vector machine

I. INTRODUCTION

Cyberbullying can be defined as the use of technological advances through mobile phones, e-mail, chat rooms or social networking platforms such as Twitter or Facebook to embarrass or threaten others [1]. The word cyberbullying appeared not long ago, but lately the number of occurrences is increasing. People who do cyberbullying do not need a strong physic because they just need an access to a cell phone or computer with the desire to terrorize, embarrass or threaten others. Confidentiality of identity in cyberspace is a key of cyberbullying and also the factor that makes the people who never did any bullying in real life becomes a cyberbully [2].

According to research from The University of British Columbia, cyberbullying is a bigger problem than traditional bullying. There were surveys of 733 adolescents, stating that 25-30% of them had been involved in cyberbullying, while only 12% of them had been involved in traditional bullying. 95% of them declared using mocks in internet only as a joke, and the rest is meant to insult or hurt someone. It states that teenagers greatly underestimate the danger of cyberbullying [3].

The detection of cyberbullying in the use of online platforms becomes very important. Due to too much information that is not possible to track by humans, automatic detection is needed that can identify threatening situations and hazardous content. This enables large-scale monitoring of social media [7].

II. BACKGROUND AND RELATED WORK

In recent years there are several studies related to cyberbullying analysis and detection using text mining by classifying conversations or posting. Yin, Xue and Hong [4] use supervised learning, labeling using N-grams and weighting using TF-IDF. Dinakar, Reichart and Lieberman [5] conducted a supervised machine learning approach, they collected youtube comments, labeled them manually and implemented various binary and multiclass classifications. Kelly Reynolds [6] used the decision tree (J48) and k-nearest neighbor (k = 1 and k = 3), labeling using Amazon Mechanical Turk. Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes and Bart Desmet [7] use Support Vector Machines (SVM) as classification algorithms because they are well proven in classification. In their research, when the preprocessing step happens, they apply tokenization, PoS-tagging and lemmatization to the data by using LeTs Preprocess Toolkit.

Based on those facts, this research will be done to classify cyberbullying on text conversations using text mining method by developing previous research from Kelly Reynolds (2012). The research was conducted by identifying the characteristics of cyberbullying on the conversation as well as classification using SVM and Naive Bayes method as comparison with Kelly Reynolds method of decision tree (J48) and k-NN (k = 1 and k = 3). Support Vector Machines (SVM) as the classification algorithm, since they have been proven to work well for high-skew text classification tasks. Naïve Bayes requires little data for training, can produce maximum results. Besides classification of 2 classes conducted Kelly Reynolds, in this research will be classified 4 classes and 11 classes for using recommendations according to results of classification.

III. METHODOLOGY

The methodology used in this research shows in Fig 1.

A. Data Collection

The data used to create a data set is a textual conversation taken from the Kaggle (www.kaggle.com) which provides 1,600 conversations in Formspring.me. Question, Answer and Severity are the fields used as label in this research. Each conversation is a combination of Question and Answer fields. The combined results of Question and Answer from excel files are made into files with txt extensions and grouped in folders 0 through 10 according to the severity level used as labels. After data collection, data is imported into Rapid Miner to continue

the process of Preprocessing, Extraction, Classification and Evaluation.

B. Preprocessing

Conversation Text on each set of data is later preprocessed in order to facilitate the processing of text conversations at the next stage:

1) Data Cleaning & Data Balancing: The amount of data obtained from *www.kaggle.com* is 12,729 data, including 11,661 data given non-cyberbullying label and 1068 data labeled cyberbully. Data cleaning is done with Microsoft excel by eliminating conversations that have total characters under 15 letters, deleting meaningless words like "haha", "hehe", "wkwk", "emm", "umm". For the purposes of data balancing on the classification of 2 classes （cyberbully, non-cyberbully), 4 classes (non-cyberbully, cyberbully level severity low, cyberbully level severity middle, cyberbully level severity high), and 11 classes (non-cyberbully, cyberbully level severity 1 – 10), then the data used amounted to 1.600 for balancing data (800 labeled cyberbully and 800 labeled non-cyberbully)with the following allocation:

a) 2 Class: each class amounts to 800 data
- Class No : 800 data with label severity 0
- Class Yes : 800 data with label severity 1-10

b) 4 Class: each class amounts to 240 data
- Class No : 240 data with label severity 0
- Class Low : 240 data with label severity 1 – 3
- Class Middle : 240 data with label severity 4 – 7
- Class High : 240 data with label severity 8 - 10

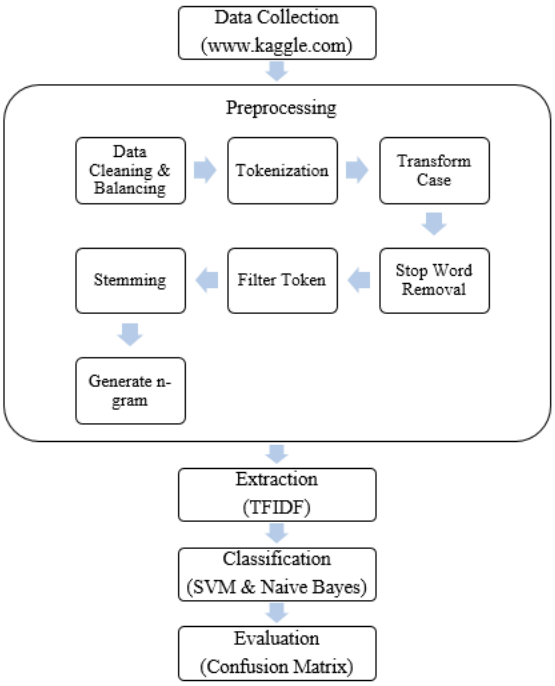

Fig. 1. Research Method

c) 11 Class: each class amounts to 80 data
- Class 0 : 80 data with label severity 0
- Class 1 : 80 data with label severity 1
- Class 2 : 80 data with label severity 2
- Class 3 : 80 data with label severity 3
- Class 4 : 80 data with label severity 4
- Class 5 : 80 data with label severity 5
- Class 6 : 80 data with label severity 6
- Class 7 : 80 data with label severity 7
- Class 8 : 80 data with label severity 8
- Class 9 : 80 data with label severity 9
- Class 10 : 80 data with label severity 10

2) Tokenization: tokenization is the process of cutting or separating each word that compiles a document or conversation. In general, every word is identified or separated by other words by a space character, single quoting character ('), dot (.), semicolon (;), colon (:), so the tokenizing process uses non-letters mode to perform word separation.

3) Transform case: Transformation into the lower case to facilitate the next process with purpose not to distinguish between capital letters and lowercase letters.

4) Stop Word Removal: Delete unnecessary words on every text conversation in accordance with English vocabulary by using Stop Word Filter (English).

5) Filter Token: The token filter is selecting the word that the number of characters between 3-25, because below 3 characters word is stopword and above 25 are character is rarely used words.

6) Stemming: The words on the text conversation are transformed into a basic word using the Porter Stemmer algorithm.

7) Generate n-grams: The process of generating n-grams is to form a set of words from a paratable and graph, usually by moving one word forward, in this research a n-gram of 2 to 5, because the experiments have been done n-gram over 5 is stable (the result is the same as n-gram 5).

C. Extraction

The preprocessing text conversations will be transformed into a vector space model where text conversations are represented with a vector of extracted features. Features resulting from the extraction are words or combinations of words to form a list of words and the calculation of the weight with TF-IDF

D. Classification

In this stage the classification will use the Naïve Bayes & SVM method with linear, poly, RBF, and sigmoid kernels. Each conversation in the form of questions and answers is combined into one text conversation. The collected text conversations are randomly divided into sets of training and test data.

Each text conversation consisting of 1600 conversations is labeled according to the data set and text conversation status. The division of text conversations into data sets is done 10 times

TABLE I. CONFUSION MATRIX

		Prediction		**Total**
		-1 (Negative)	**+1 (Positive)**	
Actual	**-1 (Negative)**	*p*	*q*	*p+q*
	+1 (Positive)	*u*	*v*	*u+v*
Total		*p+u*	*q+v*	*m*

with cross validation. This research will be classified into 2 classes (800 data each class), 4 classes (240 data each class) and 11 classes (80 data each class).

E. Evaluation

To evaluate the classification model based on the accuracy can be measured from the accuracy of the model in classification with the method of confusion matrix. Confusion matrix is a matrix consisting of rows and columns as shown in Table I. The row corresponds to a predefined value while the column corresponds to the predicted value predefined by the classification model [8].

$$Accuracy\ rate = \frac{p+v}{p+q+u+v} = \frac{p+v}{m} \quad (1)$$

IV. RESULTS

In this section, we show the result of the classification using Naïve Bayes and SVM methods with Linear, Poly, RBF, and sigmoid kernels based on the number of classes and show average result number of n-grams (n-gram 1-5).

TABLE II. ACCURACY RATE CONFUSION MATRIX 2 CLASSES

2 Classes	**Average**
SVM-Linear	99.04%
SVM-Poly	99.41%
SVM-rbf	63.77%
SVM-Sigmoid	99.04%
Naïve Bayes	96.98%

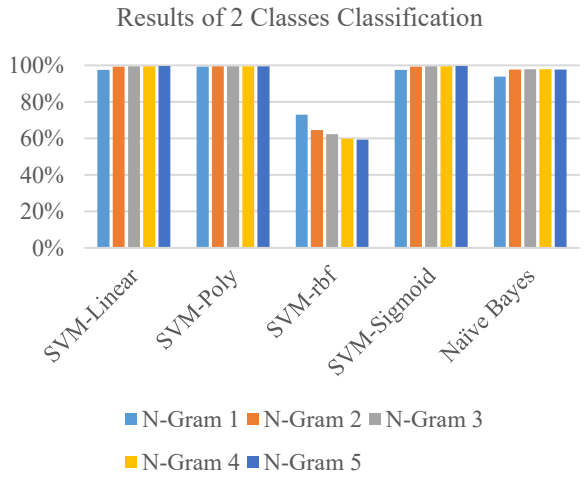

Fig. 2. Results of 2 classes classification

A. Results of 2 Classes Classification

Based on the table II and Fig.2, it can be seen that the SVM method with the kernel poly has the highest accuracy with an average accuracy of 99.41%. Naive bayes and SVM linear and sigmoid also have a high accuracy of above 95%, SVM RBF has the lowest accuracy of 63.77%.different from the others, svm RBF applied to the n-gram higher, the accuracy decreases.

B. Results of 4 Classes Classification

Based onTable III and Fig3 , it can be seen that the SVM method with the kernel poly has the highest accuracy with an average accuracy of 97.81%.SVM method with the kernel RBF has the lowest accuracy with an average accuracy of 81.90%.

TABLE III. ACCURACY RATE CONFUSION MATRIX 4 CLASSES

4 Classes	**Average**
SVM-Linear	95.21%
SVM-Poly	97.81%
SVM-rbf	81.90%
SVM-Sigmoid	95.21%
Naïve Bayes	92.37%

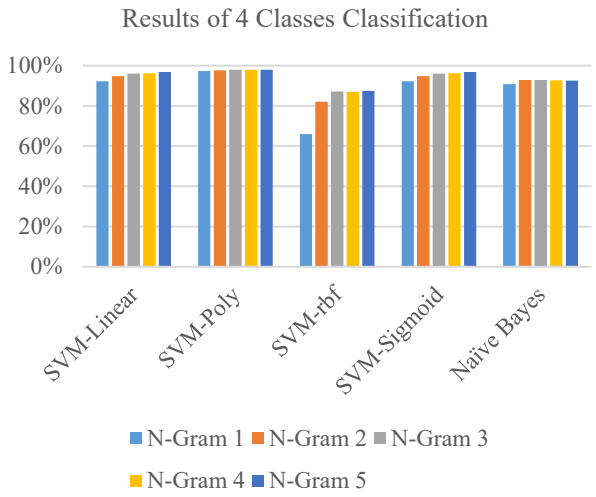

Fig. 3. Results of 4 classes classification

TABLE IV. ACCURACY RATE CONFUSION MATRIX 11 CLASSES

11 Classes	**Average**
SVM-Linear	93.48%
SVM-Poly	94.12%
SVM-rbf	86.73%
SVM-Sigmoid	93.48%
Naïve Bayes	89.09%

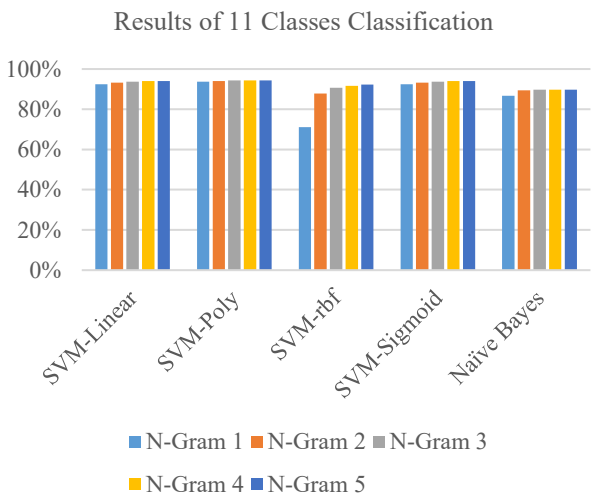

Fig. 4. Results of 11 classes classification

C. *Results of 11 Classes Classification*

Based on the table IV and Fig 4, it can be seen that the SVM method with the kernel poly has the highest accuracy with an average accuracy of 94.12%. SVM method with the kernel RBF has the lowest accuracy with an average accuracy of 86.73%.

TABLE V. AVERAGE RESULTS BASED ON MODEL

Model	Average
SVM-Linear	95.91%
SVM-Poly	97.11%
SVM-RBF	77.47%
SVM-Sigmoid	95.91%
Naïve Bayes	92.81%

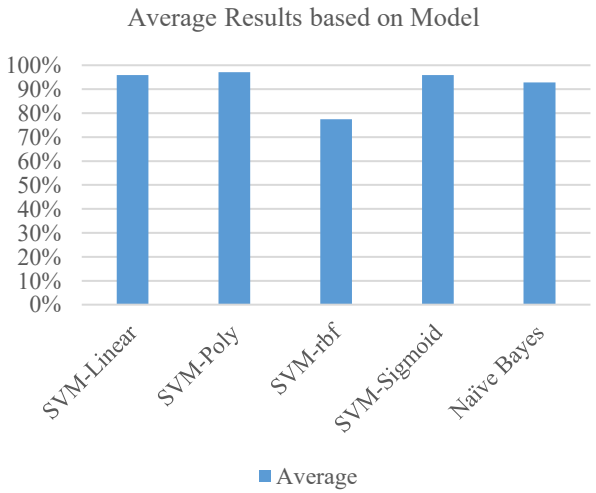

Fig. 5. Average Results based on Model

D. *Average Results based on Model*

Based on the table V and Fig 5, it can be seen that the SVM method with the kernel poly has the highest accuracy with an average accuracy of 97.11%. SVM method with the kernel RBF has the lowest accuracy with an average accuracy of 77.47%. Average accuracy of Naive Bayes is 92.81%.

E. *Average Results based on N-Gram*

Based on the table VI and Fig 6, it can be seen that the N-gram 5 has the highest accuracy with an average accuracy of 92.75% and the lower accuracy is N-gram 1 with an average 89.05%. Can be explained the higher n-gram will produce a higher level of accuracy.

TABLE VI. AVERAGE RESULTS BASED ON N-GRAM

N-gram	Average
N-Gram 1	89.05%
N-Gram 2	92.01%
N-Gram 3	92.72%
N-Gram 4	92.66%
N-Gram 5	92.75%

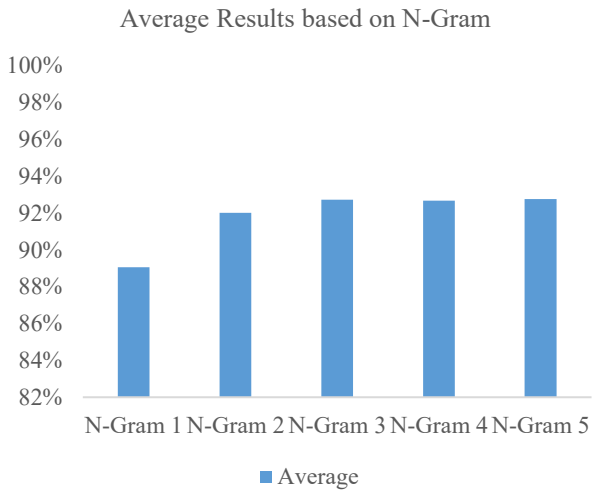

Fig. 6. Average Results based on N-Gram

TABLE VII. AVERAGE RESULTS BASED ON NUMBER OF CLASSES

Number of Class	Average
2 Classes	91.65%
4 Classes	92.50%
11 Classes	91.38%

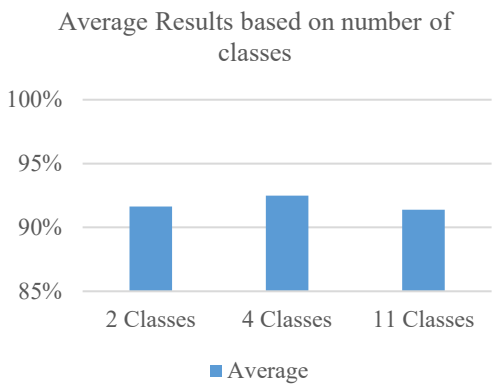

Fig. 7. Average Results based on number of classes

F. *Average Results based on number of classes*

Based on the table VII and Fig 7, it can be seen that the 4 classes has the highest accuracy with an average accuracy of 92.50%. The lowest accuracy is 11 classes with an average accuracy 91.38%.

G. *Comparison with Previous Research*

We tried to compare the Kelly Reynolds Research (2012) work. Due to the data of Kelly Reynolds research, in this paper similar Kelly Reynolds method are applied. Decision tree (J48) and k-nearest neighbor (k = 1 and k = 3) are used to classify 2 classes. Table VIII shows a comparison of accuracy with Kelly Reynolds's study's:

TABLE VIII. COMPARISON WITH PREVIOUS RESEARCH

Research	2 Classes	Average
Kelly Reynolds	J48	78.28%
	k-NN (k=1)	89.01%
	k-NN (k=3)	74.53%
This Research	SVM-Linear	99.04%
	SVM-Poly	99.41%
	SVM-rbf	63.77%
	SVM-Sigmoid	99.04%
	Naïve Bayes	96.98%

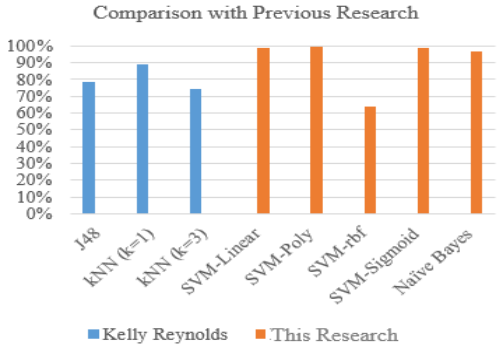

Fig. 8. Average Results based on number of classes

Based the table VII and Fig 8, it can be seen that the SVM method with poly kernel has the highest and the most stable accuracy although applied to n-grams from 1 to 5 with an average accuracy of 99.41%, while Kelly Reynolds's research by decision tree method (J48) only reached 78.28%, k-NN (k = 1) reached 89.01%, k-NN (k = 3) reached 74.53%.

V. CONCLUSIONS

The most optimal SVM kernel in classifying cyberbullying is the Poly kernel with an average accuracy of 97.11%, because of the data used in this study are non-linear separable. Therefore, the optimal function for separating the sample into different classes is SVM with poly kernel. The application of n-gram may increase the accuracy level in cyberbullying classification, due to the highest accuracy level at n-gram 5 (92.75%), the lowest accuracy set at n-gram 1 (89.05%).

In the future, classification of cyberbullying on text conversations in Bahasa Indonesia is interesting for applying this method. Moreover, classifying text conversations is more challenging due to text conversations usually has shorted words. Therefore, the pre-processing of text conversations is required to implement spelling correction algorithm.

REFERENCES

[1] R. M. Kowalski, S. P. Limber and P. W. Agatston, Cyberbullying: Bullying in the Digital Age, West Sussex: Wiley-Blackwell, 2012.

[2] S. Poland, "Cyberbullying Continues to Challenge Educators," 1 5 2010. [Online]. Available: https://www.districtadministration.com/article/cyberbullying-continues-challenge-educators.

[3] J. Shapka and H. Amos, "Cyberbullying and bullying are not the same: UBC research," 13 April 2012. [Online]. Available: http://news.ubc.ca/2012/04/13/cyberbullying-and-bullying-are-not-the-same-ubc-research/.

[4] Yin, D., Xue, Z., & Hong, L. (2009). Detection of Harassment on Web 2.0. *Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, 1-7.

[5] K. Dinakar, R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," International AAAI Conference on Web and Social Media, pp. 11-17, 2011.

[6] K. Reynolds, "Using Machine Learning to Detect Cyberbullying," the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science, pp. 1-4, 2012.

[7] C. V. Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. D. Pauw, W. Daelemans and V. Hoste, "Automatic Detection and Prevention of Cyberbullying," HUSO 2015: The First International Conference on Human and Social Analytics, pp. 13-18, 2015.

[8] C. Vercellis, Business intelligence: Data Mining and Optimization for Decision Making, Politecnico di Milano, Italy.: Wiley, 2009.