International Journal of Computer Applications (0975 – 8887)
Volume 136 – No.2, February 2016

Sentiment Analysis Candidates of Indonesian Presiden 2014 with Five Class Attribute

Ghulam Asrofi Buntoro
Muhammadiyah University of Ponorogo
Department of Informatics
Faculty of Engineering

Teguh Bharata Adji
Gadjah Mada University
Department of Electrical Engineering
Faculty of Engineering

Adhistya Erna Purnamasari
Gadjah Mada University
Department of Electrical Engineering
Faculty of Engineering

ABSTRACT
Nowadays, Twitter is not only used for social media to maintain friendship, but also Twitter is used to promote and campaign. Twitter usersare free to express their opinions, including opinions about candidates of Indonesian President 2014. This research accommodate the public opinions by classified it into five class attributes : very positive, positive, neutral, negative and very negative. The classification process using Naïve Bayes Classifier (NBC) with data preprocessing using tokenization, cleansing and filtering. The data used in this research are in Indonesian tweets about candidates of Indonesian President 2014, with 900 tweets of dataset and distributed to five class attributes equally. As result, highest accuracy obtained when the experiment using combination of tokenization n-gram, stopword list WEKA and emoticons, which is the values consisting 71,88% accuration, 71,6% precision, 71,9% recall, 6,1% TP rate and 65% TN rate.

General Terms
Computer Science, Artificial Intelligence, Machine Learning.

Keywords
Sentiment Analysis, Candidate of Indonesian President 2014, Five Class Attribute, Naïve Bayes Classifier (NBC).

1. INTRODUCTION
The development of social media to give space to the user or community of expression, opinion and expression freely. Social media one of the fastest growing consumer is Twitter, Twitter users mentioned amount continues to increase to 300,000 users every day [1]. On Twitter the free speech and opinion about the goodness or badness person, including giving opinions and assess candidates for President of Indonesia in 2014, which Jokowi and Prabowo. Monitor the opinion of the people is not easy, because opinions in social media are numerous once and mostly written by words or phrases that are not standard. Therefore he need a special method or technique that is able to categorize those opinions.
Sentiment analysis is a branch of the field of text mining studies are learning about sentiment, emoticons and text contained in the attitude of opinion. The basic principle of sentiment analysis is to classify the polarity of a given text and decisive opinion in the form of the text. In this study, the domain used is tweet related to the President of Indonesia, 2014. Selection of candidates is because the public domain, it is relatively confused to make their choice. The public confusion triggered by the crux of the information in Twitter of the two candidates for President of Indonesia in 2014, which Jokowi and Prabowo.

In previous studies of sentiment analysis, opinion is divided into two, namely the positive and negative [2]. Subsequent research opinion is divided into three, namely positive, negative and neutral [3]. In the opinion of this study are categorized into five, which is very negative, negative, neutral, positive, and very positive. The aim is to make a more specific opinion and more subtle. More specific division of opinion is intended also to accommodate the opinion or public opinion that has not accommodated if only two or three opinions categories: positive, negative and neutral. Because the opinions expressed on Twitter a little more able to influence the choice of candidates for president terhadapat society [4]. Thus the need for quantification of various opinion or public opinion on Twitter that makes the value of a more specific opinion. For now Twitter became one of the places required for the campaign of the candidate for President of Indonesia in 2014.

2. RELATED WORK
Research by Mesut et al [2] uses machine learning to classify the Turkish political news. This research classify sentiment towards Turkey's political news and political news to determine whether Turkey has a positive or negative sentiment. The distinct feature of the Turkish political news is extracted with machine learning algorithm Naïve Bayes Classifier (NBC), Maximum Entropy (ME) and Support Vector Machine (SVM) to generate a classification model. This study gained 72.05% Accuracy for Naïve Bayes Classifier (NBC), Maximum Entropy Accuracy 69.44% and 66.81% for SVM on the use bigram.

Pak and Paurobek [3] use emoticons to establish English-language corpus of Twitter with sentiment positive, negative and neutral. For the class of neutral Pak and Paurobek take account tweet training data from English-language media. The method used is Naïve Bayes with n-gram. Best performasi generated when using bigram.

Read [5] using emoticons like ": -)" and ": - (" to establish a training set classification sentiment. For this purpose a read collect text that contains emoticons from Usenet Newsgroups. The dataset is divided into positive samples (text with a happy emoticon) and negative samples (text with sad or angry emotions). Read [5] attempted classification is based on the topic of dependency. For the category of mixed topics on the training data and mixed topics on testing the data, Read produces an accuracy of 84.6% for the Naïve Bayes method and 81.1% for the SVM method. For other classification topics read gain accuracy of the range of 62% to 70% in the test set for both methods. Read also obtained a 81.5% accuracy for SVM method and accuracy of 78.9% for methods Naïve Bayes, when using a domain dependency on sentiment classification.

The study by Pang et al [6] uses machine learning to classify moview reviews. This study did sentiment classification for movie reviews and movie reviews to determine whether it has a positive or negative sentiment. The distinct feature of this review is extracted and used Naïve Bayes machine learning algorithm and Support Vector Machine (SVM) to generate a classification model. They earn between 78.7% accuracy when using the Naïve Bayes on unigram use. The accuracy obtained when using SVM with unigram was 72.8%.

Frangky and Manurung [7] tried to repeat the experiment sentiment classification movie review by Pang et al [6] to Indonesian. Ketidaktersediaannya related to training corpora for Indonesian, then applied to machine translation tools to mentranslatasikan English corpus made Pang et al [6] which is native to the Indonesian and translasinya results are used to train the classification. A wide selection of used machine translation start of commercial tools to simple translation and verbatim text classification methods attempted. Average accuracy results were obtained for the Naïve Bayes method was 74.6% and 75.62% for SVM method. Best results are obtained together with that obtained when using experiments in English.

3. EXPERIMENTAL METHODOLOGY

Measures in accordance with the flow of the research study are as follows:

3.1 Collecting Data Tweet

Collects data from Twitter users tweet opinion. Tweet retrieved through the search facility with the keyword Jokowi and Prabowo. Taken starting on May 1, 2014 - July 1 2014.

3.2 Changing Tweet the Data into a format Arff

Tweet the data collected in the form of text, then made the ARFF file [8]. To make the ARFF file manually.

3.3 Preprocessing Data

Perform preprocessing of data tweet. Preprocessing includes tokenization cleansing and filtering. Tokenization done to break down the tweet into some word or set of words that stand alone. This study uses 5 tokenization method, namely unigram, bigram, trigram, 3-5 grams with a minimum value of n = 3 and n = 5 and a maximum of n-grams with a minimum value of n = 1 and n = 3. Process tokenization maximum use menus in WEKA. At tokenizer select tokenization choose and select the method that will be used.
Then entered the cleansing process, the process of removing symbols is less important in a tweet that could interfere with the data classification process will be. This process is done by using the menu on WEKA delimiters.

The final stage after tokenization in this study is filtering. Filtering is done to remove the words are less important or less affect the classification process will be. This process is done by using stopword list. Stopword list used in this study was Weka stopword list and stopword list of Indonesian made by Tala [9].

3.4 Data Tweet Converted into Vector

Tweet data is then converted into vector form [10]. By selecting StringToWordVector on WEKA tool. The result can be seen in Table 1.

Table 1. Table Example of a vector of text data

38:Capres Numeric	39: Cina Numeric	47:Dosa Numeric	48:Fakta Numeric
0.0	0.0	0.0	0.0
3.052966	0.0	0.0	4.234608

In Table 1. the part that is in the red box are the words that exist in the data tweet. For each row of data representing each tweet. On the 3rd line in the blue box can be seen that the word "candidate" has a value of 3.052966 and the word "fact" has a value of 4.234608. While others worth 0.0 it means that there is the word "candidate" and "facts" in the data tweet 3rd.

3.4.1 Giving Weight

Giving weight to each word (term). Weighting is done to get the value of the word that successfully extracted. The method used for assigning weights in this study is the TF-IDF (Term Frequency - Inverse Document Frequency).

3.4.2 Classification

Entered in the classification process. The classification process using Weka 7.3.11. Classification method used in this study is Naïve Bayes Classifier (NBC). In the process of data classification was tested using 10-fold cross validation [11]. So dataset will be divided into two, namely 10 parts by 9/10 part is used for the training process and 1/10 parts used for process testing. Iteration lasts 10 times with a variety of training data and testing using a combination of 10 pieces of data.

3.4.3 Evaluation Results

To evaluate the performance of Accuracy, Precision and Recall of the experiments that have been dilakuakan. Evaluation is done by using the true positive rate (TP rate), true negative rate (TN rate), false positive rate (FP rate) and false negative rate (FN rate) as an indicator. TP rate is the percentage of positive class successfully classified as positive grade, while the TN rate is the percentage of negative class successfully classified as negative class. FP rate is negative classes that are classified as positive class. FN rate is positive grade classified as negative class [12].

4. RESULTS AND DISCUSSION

The dataset in this study using ARFF format collected from Twitter. The data taken is the opinion or public opinion in the Indonesian tweet about Indonesian presidential candidate in 2014, namely Jokowi and Prabowo.

The dataset used as many as 900 tweets, the data is split equally (balanced) of each class, because the data is out of balance (imbalanced), a classification that is built has a tendency to ignore the minority class [25]. Very negative data is divided into 180 tweets, 180 negative tweet, tweet somewhat negative 180, 180 tweets neutral, slightly positive 180 tweet, tweet positive and very positive 180 180 tweets. Labelling is done manually with the help of Indonesian experts.

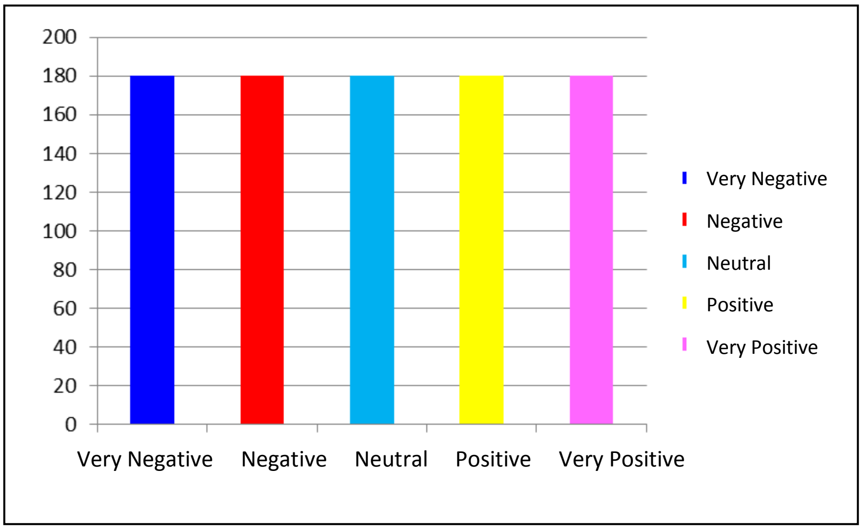

Figure 1. Data Text

In Figure 1. The blue section shows the class is very negative, the red part shows a negative grade, marine blue section shows the class neutral, yellow colored sections showed a positive grade, and part of the pink indicates a very positive grade. From the picture above shows that the dataset used in this research balanced (balanced dataset). The accuracy of the results obtained from studies using a dataset that is balanced better than using a dataset that is not balanced [13]. In this study, using data that is balanced and implement methods Naïve Bayes Classifier (NBC). The goal is to prove whether the result is good enough what is not.In this study, the method used to perform data classification is Naïve Bayes Classifier (NBC) using Weka software version 3.7.11 to classify. WEKA using document types Attribute-Relation File Format (ARFF) as input to perform data classification.

Results from the classification process and then tested using 10-fold cross validation, the data form the 10 sections divided by 9/10 part is used for the training process and 1/10 parts used for process testing. Iteration lasts 10 times with a variety of training data and testing using a combination of 10 pieces of data.

Table 2. Illustration 10 fold Cross Validation

Test	**Dataset**									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

The part that is black in Table 2. shows the test data. In the above image dataset is divided into 10 subsets. In the first test, the tested subset is a subset of the first, while the second to the tenth subset is used as training data. In the second test, which tested subset is a subset of the latter, while the first subset, the third to the tenth is used as training data and so on.

From the results of the classification by using tokenization, cleansing and filtering different, it can be obtained by the comparison table like Table 3.

Table 3. Comparison of Results Classification

Methods Tokenization	Emoticons	Stopword List	Accuracy (%)	Precision (%)	Recall (%)	TP Rate (%)	TN Rate (%)
Unigram	Yes	WEKA	70,55	70,4	70,6	64,4	63,9
		Indo	66,66	66,2	66,7	53,9	61,7
	No	WEKA	70,66	70,5	70,7	65,0	64,4
		Indo	66,66	66,2	66,7	53,9	61,7
Bigram	Yes	WEKA	54,44	58,4	54,4	45,0	35,6
		Indo	54,44	58,4	54,4	45,0	35,6
	No	WEKA	54,44	58,4	54,4	45,0	35,6
		Indo	54,44	58,4	54,4	45,0	35,6
Trigram	Yes	WEKA	36,44	58,4	36,4	29,4	13,9
		Indo	36,44	58,4	36,4	29,4	13,9
	No	WEKA	36,44	58,0	36,4	29,4	13,9
		Indo	36,44	58,0	36,4	29,4	13,9
3-5gram	Yes	WEKA	35,55	56,9	35,6	26,7	13,9
		Indo	35,55	56,9	35,6	26,7	13,9
	No	WEKA	35,55	56,9	35,6	26,7	13,9
		Indo	35,55	56,9	35,6	26,7	13,9
N-gram	Yes	WEKA	71,88	71,6	71,9	66,1	65,0
		Indo	69,11	68,7	69,1	60,0	63,3
	No	WEKA	71,77	71,6	71,8	66,1	65,0
		Indo	69,22	68,8	69,2	59,4	63,9

*) The value of Precision and Recall the average value of the value of the positive class and negative class

Table 3. contains information regarding the value of accuracy, precision, recall, TN TP rate and rate of each of the trials that have been conducted. Part column contains information about the type of method tokenization, use emoticons, stopword list used in each test, the value of accuracy, precision, recall, TP and TN rate rate. Being part of the line contains the value of accuracy, precision, recall, TN TP rate and rate of each of the trials that have been conducted. This study uses 5 tokenization method, namely unigram, bigram, trigram, 3-5gram and from Figure 2. it can be seen that the accuracy is checked with 2 types stopword list (stopword list and stopword list WEKA Indonesian) using emoticons and delete emoticons.

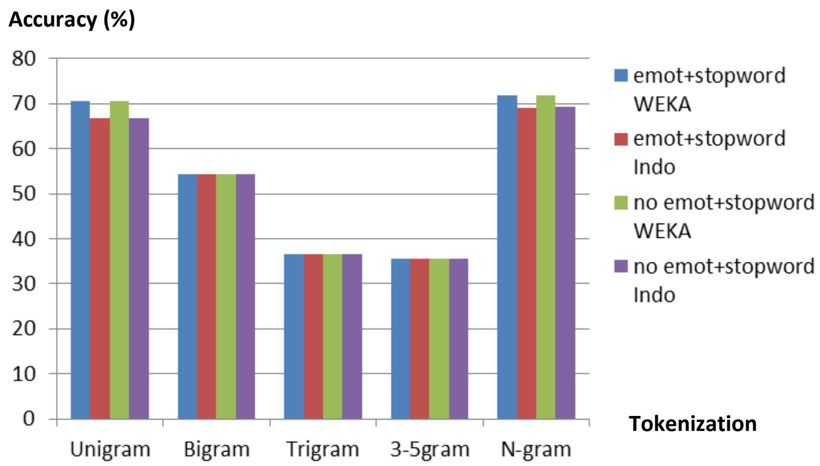

Figure 2. Graph Accuracy Rate

In this study, the highest accuracy is obtained by using tokenization n-gram, stopword list WEKA with emoticons, why when using n-gram gain accuracy is the highest, because tokenization this break down or describe a tweet into multiple tokens, each token size n word. In this study, n minimum 1 and maximum 3. So the classification method will be more accurate because there are 3 variations in the size of the token, which is 1 word per token, n-gram. In the preprocessing of data each tokenization method using two types stopword list (stopword list and stopword list WEKA Indonesian) using emoticons and delete emoticons. Of the process of generating the data preprocessing token number which is then used as input a classification process. The classification process is done using Naïve Bayes Classifier (NBC). Of the values obtained classification accuracy, precision, recall, TN TP rate and rate of each test. Token 2 and 3 words per word per token. The second highest accuracy is obtained when using tokenization unigram, stopword list WEKA without emoticons, because tokenization unigram tweet breaking data into tokens with each token is one word, so any data will tweet classified word by word. While the value of the lowest accuracy obtained when using tokenization 3-5gram, why tokenization 3-5gram get the lowest accuracy, because the data is too big for each tokennya, because the minimum value of n 3 and n 5 makes the maximum size of the token there are 3 words, 4 words and 5 words per token.

With 5 words each token will make the combination of a pattern that appears to be too much so it is difficult to find the optimal classification function. Rated the accuracy of the value derived from the number of times accuracy is one of the parameters of the assessment methods that have been used, data is successfully classified correctly according to the class sentiments of the entire amount of data to be classified. High accuracy values obtained when much data successfully classified correctly according to grade sentiment.

Figure 2. above shows that the use of Indonesian stopword list is not able to improve accuracy, even on the contrary, when using stopword list Indonesian accuracy to be down. Because stopword list Indonesian delete the words of negation, for example, "no", "no", and so on, so the positive word can be negative and vice versa, examples of positive phrase, "do not be evil" but because the word "not" be removed, it became negative. While all methods tokenization combined with stopword list WEKA produce high accuracy, because stopword list stopword list WEKA is English, so it can not erase the words with Indonesian.

Figure 2. also shows that the use of emoticons in tokenization n-gram can improve accuracy value, whereas when used in tokenization unigram accuracy by using emoticons less high than that without emoticons, because when tokenization n-gram, emoticons do not stand alone, be classified together words in 1 token, so accuracy is higher than the n-gram without emoticons. The highest accuracy is obtained when using tokenization n-gram, stopword list WEKA and emoticons.

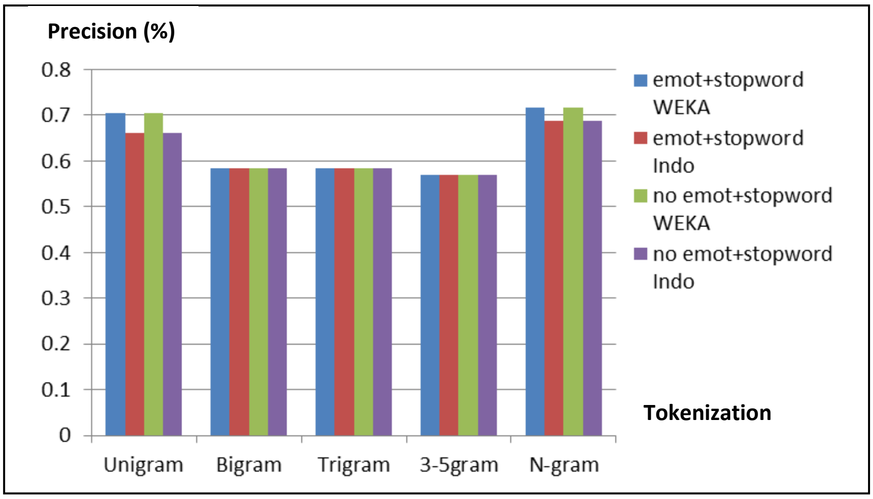

Figure 3. Chart Precision

From Figure 3. it can be seen the value of the highest precision obtained using n-gram tokenization, stopword list WEKA and emoticons, while the value of low precision is obtained when using tokenization 3-5gram, stopword list WEKA and no emoticons. The higher accuracy value will be followed by a high-precision value as well, and vice versa, because the value of precision is the number of true positives class are classified as positive class divided by the total data that are classified as positive class. So why n-gram precision high, because tokenization n-gram has resulted in many classes of data that are classified correctly.

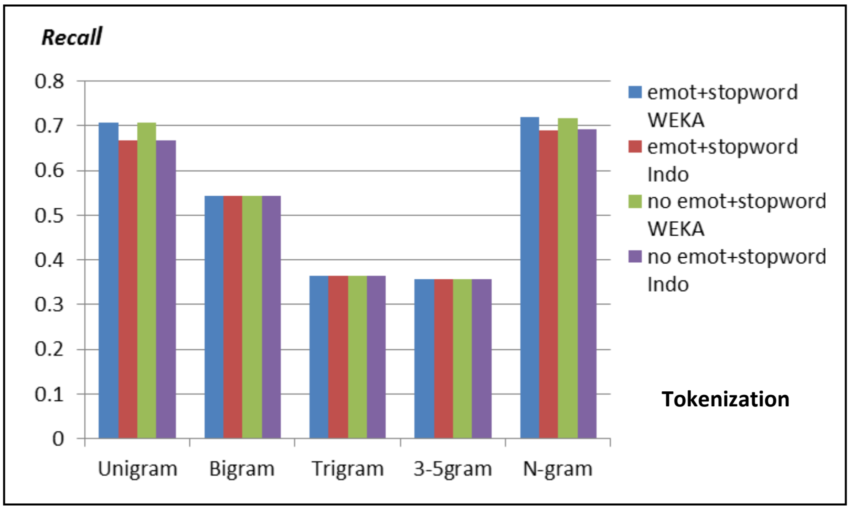

Figure 4. Graph Recall

From Figure 4. shows that the highest recall value obtained when using n-gram tokenization configuration, stopword list WEKA and emoticons. Recall value is the number of true positives class are classified as positive class divided by the number of positive real class.

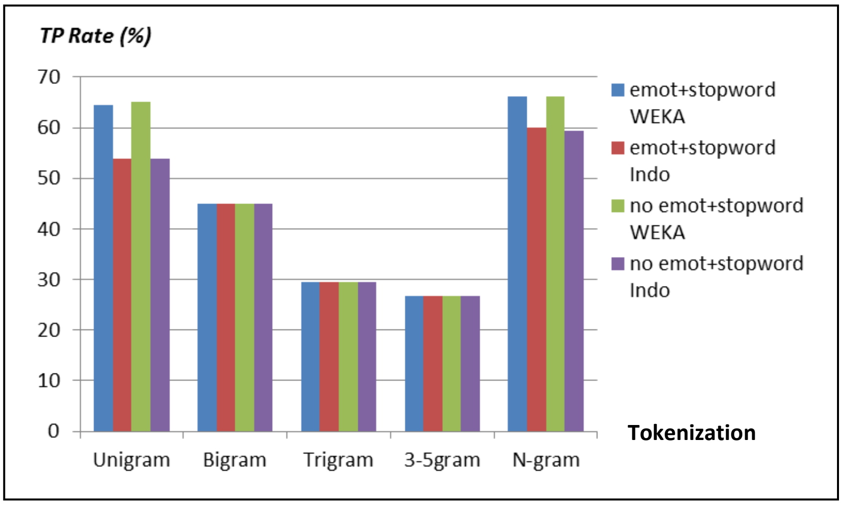

Figure 5. Graph TP Rate

In addition, Figure 5. also shows that the recall value obtained using n-gram tokenization higher than recall value obtained by methods other tokenization, low recall value obtained when using the method tokenization 3-5gram. Why 3-5gram get the most low recall value, because with this tokenization classification method used is less capable of classifying the data class correctly, the recall value is obtained from summing data correctly classified positives divided by all the data with a positive grade.

In Figure 5. we can know TP highest rate obtained when using n-gram tokenization configuration, stopword list WEKA and emoticons. TP value is the value of the data rate of positive tweets that are classified correctly in accordance class sentiments, which is positive.

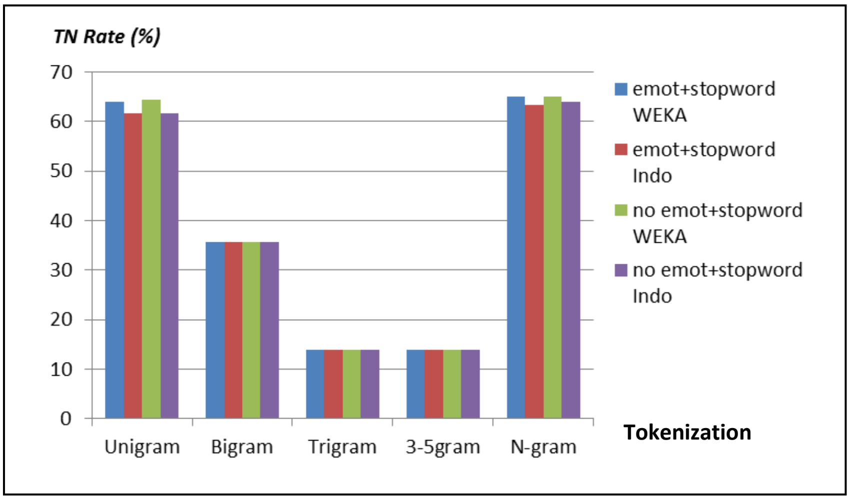

Figure 6. Graph TN Rate

It can be seen from Figure 6. Rate TN highest value obtained when using n-gram tokenization configuration, stopword list WEKA and emoticons. Rate TN value is negative tweet data values that are classified correctly in accordance class sentiments, that is negative.

From the research that has been done, it is known that the highest accuracy results obtained when using tokenization n-gram, stopword list WEKA and emoticons. Despite producing fairly good accuracy, but the model is built are still doing a little mistake at the time of the negative class classification process, as shown by TP rate greater than the value of TN Rate, value TN Rate by 65% and the value of TP Rate by 66.1% , This value is a little difference because the data used balanced, so it does not cause a lot of errors in classification, when using data that is not balanced will cause data minority class misclassified as the majority of data class [9] in the end makes the difference in value to be great.

5. CONCLUSION

From the research that has been done, it can be concluded that the sentiment analysis with five class attribute yield lower accuracy compared with two or three class attribute, with an average accuracy of 71.9%. However, using five class attribute value opinions submitted sentiment became more specific and detailed. So the information presented more clearly. The highest accuracy value obtained by tokenization n-gram, stopword list WEKA and emoticons, with the average value reached 71.9% accuracy, precision value of 71.6%, the value of 71.9% recall rate of 66.1% TP value and the value TN rate of 65%. Furthermore, needs to be developed stopword list and stemmer Indonesian were able to improve the accuracy in sentiment analysis.

6. REFERENCES

[1] Marian Radke Yarrow, John A. Clausen and Paul R. Robbins (2010). *The Social Meaning of Mental Illness.* Journal of Social Issues. Volume 11, Issue 4, pages 33–48, Fall 1955..

[2] Mesut Kaya, Guven Fidan, Ismail H. Toroslu (2012). *Sentiment Analysis of Turkish Political News.* IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[3] Pak, A., dan Paurobek, P., (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining,* Universite de Paris-Sud, Laboratoire LIMSI-CNRS.

[4] Jennifer Yang Hui (2014). *Indonesian Presidential Election: Will Social Media Forecasts Prove Right?*

[5] Read, J. (2005). *Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification*. Meeting of the Associationfor Computational Linguistics - *ACL*, 43.1-6.

[6] Pang, B., Lee, L., & Vithyanathan, S. (2002). *Thumbs Up? Sentiment Classification Using Machine Learning Techniques.* Proceedings of The ACL-02 conference on mpirical methods in natural language processing (pp. 79-86).

[7] Franky dan Manurung, R., (2008). *Machine Learning-based Sentiment Analysis of Automatic Indonesia n Translations of English Movie Reviews.* In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications.

[8] Olson, David L.; & Delen, Dursun (2008); *Advanced Data Mining Techniques*, Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1.

[9] Tala, F. Z. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, language and Computation. Universiteti van Amsterdam The Netherlands..*

[10] ARFF *files from Text Collections*. http://weka.wikispaces.com/ARFF+files+from+Text+Collections.

[11] ClassStringToWordVector.http://weka.sourceforge.net/doc.de.v/weka/filters/unsupervised/attribute/StringToWordVector.html.

[12] Ian H. Witten. (2013) *Data Mining with WEKA*. Department of Computer Science University of Waikato New Zealand.

[13] Kohavi, & Provost. (1998) *Confusion Matrix http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html.*