Sentiment Analysis With Sarcasm Detection On Politician's Instagram

Aisyah Muhaddisi*[1], Bambang Nurcahyo Prastowo [2], Diyah Utami Kusumaning Putri [3]
[1]Bachelor Program of Computer Science; FMIPA UGM, Yogyakarta, Indonesia
[2,3]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]aisyahmuhaddisi@gmail.com**, [2] prastowo@ugm.ac.id , [3]diyah.utami.k@ugm.ac.id

Abstrak

Sarkasme merupakan salah satu tantangan yang mempengaruhi hasil dari analisis sentimen. Menurut Maynard dan Greenwood (2014), performa analisis sentimen dapat ditingkatkan ketika sarkasme dapat diidentifikasi. Beberapa penelitian menggunakan metode Naïve Bayes dan Random Forest pada proses analisis sentimen. Pada penelitian Salles, dkk (2018) dalam beberapa kasus Random Forest dapat mengungguli kinerja dari Support Vector Machine yang dikenal lebih superior. Pada penelitian ini dilakukan analisis sentimen pada kolom komentar akun Instagram politikus di Indonesia. Penelitian ini membandingkan akurasi dari metode analisis sentimen dengan deteksi sarkasme dan tanpa deteksi sarkasme, menggunakan metode Naïve Bayes dan Random Forest untuk analisis sentiment lalu Random Forest untuk deteksi sarkasme. Penelitian ini menghasilkan nilai akurasi pada analisis sentimen tanpa deteksi sarkasme dengan Naïve Bayes sebesar 61%, dengan metode Random Forest sebesar 72%. Hasil akurasi pada analisis sentimen dengan deteksi sarkasme menggunakan metode Naïve Bayes – Random Forest sebesar 60% dan metode Random Forest – Random Forest sebesar 71%.

Kata kunci— *analisis sentimen, deteksi sarkasme, Random Forest, Naïve Bayes*

Abstract

Sarcasm is one of the problem that affect the result of sentiment analysis. According to Maynard and Greenwood (2014), performance of sentiment analysis can be improved when sarcasm also identified. Some research used Naïve Bayes and Random Forest method on sentiment analysis process. On Salles, dkk (2018) research, in some cases Random Forest outperform the performance by Support Vector Machine that known as a superior method. In this research, we did sentiment analysis on comment section on Instagram account of Indonesian politician. This research compare the accuracy of sentiment analysis with sarcasm detection and analysis sentiment without sarcasm detection, sentiment analysis with Naïve Bayes and Random Forest method then Random Forest for sarcasm detection. This research resulted in accuracy value in sentiment analysis without sarcasm detection with Naïve Bayes 61%, with Random Forest method 72%. Accuracy on sentiment analysis with sarcasm detection using Naïve Bayes – Random Forest method is 60% and using Random Forest – Random Forest method is 71%.

Keywords— *sentiment analysis, sarcasm detection, Random Forest, Naïve Bayes.*

Received May 22th,2021; Revised October 31th, 2021; October 31th, 2021

1. INTRODUCTION

Analysis from the Public Connection Survey shows that media consumption, along with demographic, trust, success and social capital measures, influences public connections and political participation (Couldry et al., 2007). Social media is a tool that has a major influence on political activities. According to Lopez-Lopez, et al (2014) stated that residents, with roles as supporters or consumers, mostly visit the social media of an organization (eg government, political parties) to complain (shitstorm phenomenon) or support and good experiences (candystorm phenomenon).

Instagram as one of the largest social media for people to express opinions, share thoughts and reports in real-time. There were as many as 62,230,000 active Instagram users in Indonesia in January 2020, accounting for 22.7% of the entire population. (NapoleonCat.com, 2020). Instagram is also a platform that is often used by Indonesian social media users aged 16 to 64 years, 23% higher than Twitter (Hootsuite, 2020). The amount of Instagram data increases with the height of its popularity. The maximum number of characters in Instagram upload comments and captions is 2200 characters, much different from Twitter which can only contain 280 characters. With very large data and a higher maximum number of characters, it certainly makes the text in Instagram comments more complex to analyze.

Complaints or support on Instagram accounts of politicians who represent the government in Indonesia, need to be analyzed to assist in the selection of the next policy. Various data mining techniques are applied to understand public opinion. One popular technique for analyzing data is sentiment analysis. Opinions on sentiment analysis are classified as positive, negative or neutral (Pang and Lee, 2008). In some circumstances sentiment analysis has significant drawbacks. One of them when the text contains sarcasm. It is possible that a sarcastic text that actually mocks a politician is detected as a positive opinion. The results of the research of Antonakaki, et al (2017), 11% of Twitter users who are active in the topic of the United States presidential election express opinions in the form of sarcasm. Sarcasm, as a special type of communication, where the explicit meaning is different from the implicit meaning, cannot be identified effectively with conventional data mining techniques such as sentiment analysis (Yee Liau and Pei Tan, 2014).

Maynard and Greenwood (2014) say that the performance of sentiment analysis can be improved when sarcasm can be identified. In a previous study, Yunitasari, et al (2019) detected sarcasm using the Random Forest method with 4 features, namely unigram, sentiment-related features, punctuation-related features and lexycal and syntactic features. Using these 4 features, the accuracy of sentiment analysis increased from 75% to 80%. Alita and Rahman's (2020) research succeeded in increasing the accuracy of sentiment analysis by 16.61% by detecting sarcasm in tweets about public services. There were 69 sarcasm tweets from 122 tweets with positive sentiment predictions about "Jokowi", and 82 sarcasm tweets from 100 tweets with positive sentiments about "Ahok".

In a study by Bouazizi and Otsuki (2016), a comparison was made using 4 classification methods on sarcasm detection, the highest accuracy result of 83% was obtained by the Random Forest algorithm. In addition, in the research of Salles, et al (2018) in some cases Random Forest can outperform the performance of the Support Vector Machine which is known to be superior. Based on the research mentioned, there is an algorithm that obtains high accuracy in classification with certain datasets. To find out the best method to improve the accuracy of sentiment analysis, in this study a comparison of the Machine Learning algorithm on sentiment analysis with sarcasm detection was carried out with a modification of the research flow from Yunitasari, et al (2019) Therefore, this study uses the Naïve Bayes and Random Forest algorithms with a dataset of Instagram comments from Indonesian politician's accounts.

2. METHODS

This study focuses on detecting sarcasm in Instagram comments on politician's posts in Indonesia using the Naïve Bayes and Random Forest method. The flow of this research can be seen on figure 1. The steps taken are collecting data, labeling data, preprocessing, feature extraction, classification of sentiment analysis, evaluation of sentiment classification results, classification of sarcasm comments, evaluation of sarcasm classification results, sentiment reversal if comments are detected as sarcasm, evaluation after sentiment class reversal.

Comments were obtained from data scraping using Selenium by taking comments from uploaded accounts of politicians such as Puan Maharani, Joko Widodo and the DPR RI. After the dataset is collected, the next step is the data labeling process.

The next stage is data preprocessing. At the data preprocessing stage using several methods such as data cleaning, case folding, tokenization, stopword removal, conversion of emoticons into strings, slang words into standard words, stemming. The next step, feature extraction. The features used in the model are unigram, bi-gram, sentiment-related features, punctuation-related features and lexical features.

The next step is to classify the sentiments of the data using the Naïve Bayes classifier and Random Forest classifier methods. The next step is to classify sarcasm from the data using the Random Forest method. Furthermore, the data with the label of sarcasm will be changed to a negative class. After that process, then each model is evaluated. The evaluation of the model is calculated based on the performance value, which contains accuracy, precision, recall, and f1-score.

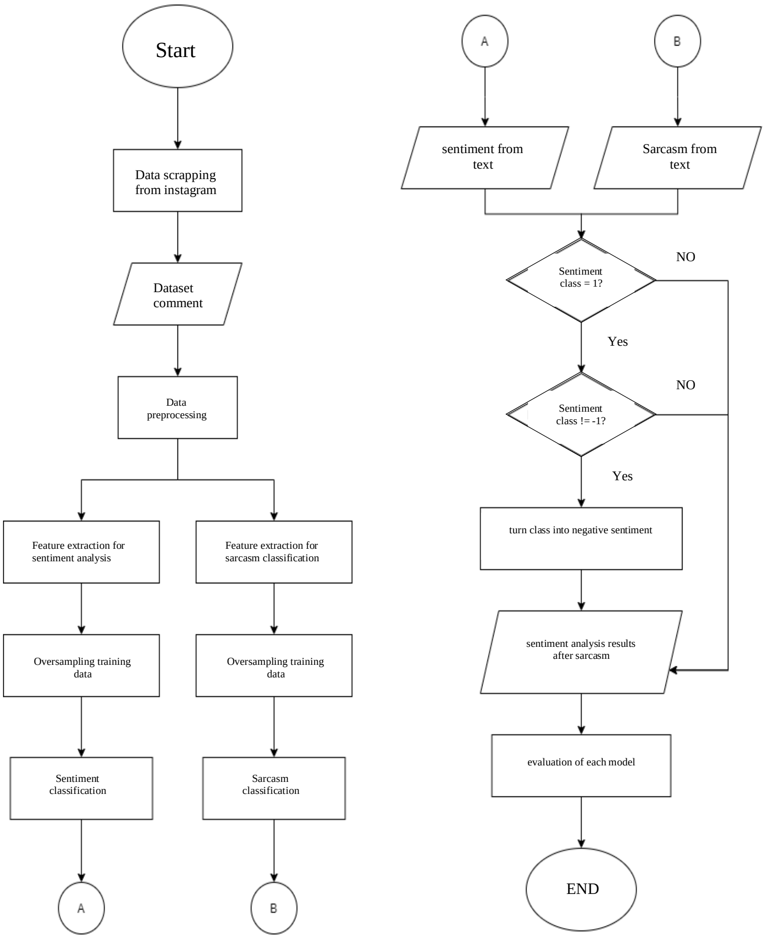
Start
Data scrapping from instagram
Dataset comment
Data preprocessing
Feature extraction for sentiment analysis
Feature extraction for sarcasm classification
Oversampling training data
Oversampling training data
Sentiment classification
Sarcasm classification
A
B
A
B
sentiment from text
Sarcasm from text
Sentiment class = 1?
NO
Yes
Sentiment class != -1?
NO
Yes
turn class into negative sentiment
sentiment analysis results after sarcasm
evaluation of each model
END

Figure 1 Sentiment analysis with sarcasm detection flowchart

2.1 Data Collecting and Labeling

Collecting data is done using Selenium to retrieve the data from Instagram. Number of comments taken as many as 3140 data. The number of classes that have been labeled with sentiment, namely 815 negative classes, 521 positive classes, 1804 neutral. The number of classes that have been labeled sarcasm class, namely 2528 non-sarcasm class and 612 sarcasm class. The results of data collection and labeling can be seen in Table 1.

Table 1 Data colection result

Coment
Nggak usah pda ngoceh Mulu, lu nyobain sono jadi presiden, mampu nggak..?.
Siapa org yg membayar hutang negara??? Dan sampai kpn hutang negara itu terlunaskan??
❤️

The class distribution of the data can be seen on Figure 2. The green bar describe number of non sarcasm data and the blue bar describe number of sarcasm data.

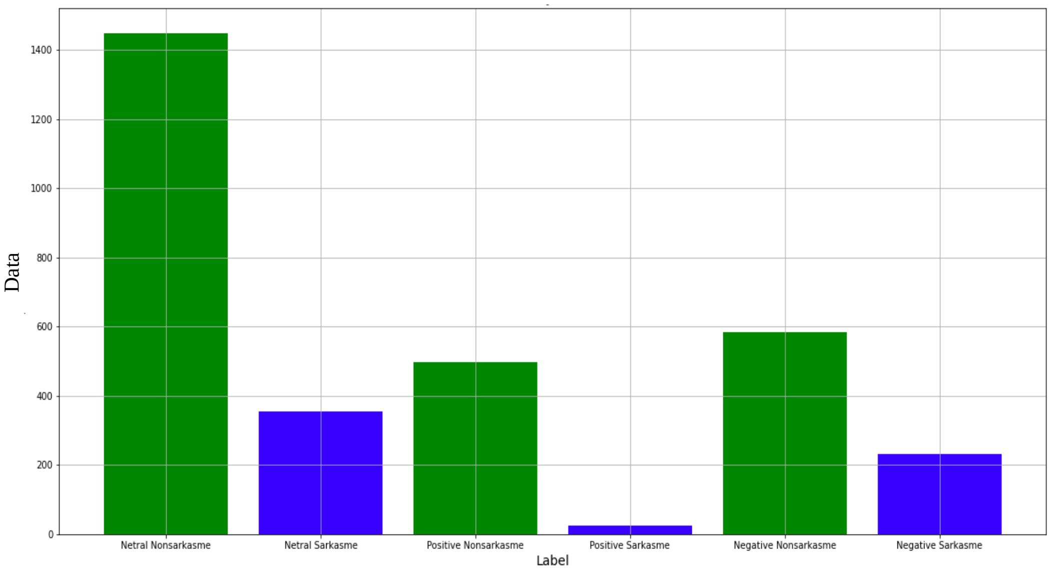

Figure 2 Distribution of each class pair

2.2 Data Preprocessing

After the data is labeled, the data preprocessing stage is carried out. Data preprocessing is done by converting emoji into strings, deleting unimportant parts of the text, separating sentences into word parts called tokens (tokenization), changing language slang words into standard words, deleting words that often appear and don't have important meaning (stopword removal) and changing words into basic words (stemming). The result of data preprocessing can be seen on Table 2.

The process of converting Indonesian slang words into standard words is assisted by a dictionary from github Louis Owen (https://github.com/louisowen6/NLP_Bahasa_resources/ blob/master/combined_slang_words.txt). The process of getting sentiment from the text to calculate the sentiment of each word using the InSet Lexicon dictionary which contains 3609 positive words and 6609 negative words with a range of -5 to +5, from github Fajri Koto (https://github.com/fajri91/InSet).

Table 2 Data preprocessing result

Coment
ngoceh melulu nyobain sono presiden
orang bayar hutang negara hutang negara lunas
red heart

2.3 Features Extraction

The next step after preprocessing the data is feature extraction. The features used are TF-IDF unigram, sentiment-related features, punctuation-related features and lexical features.

2.3.1 TF-IDF

TF-IDF weighting combines term frequency (tf) and inverse document frequency (idf) models. The first element, TF counts the occurrence of terms (words) in the document. TF with term t, is calculated as follows:

$$\mathrm{TF}(t) = \frac{n_t}{n_d} \tag{1}$$

where nt is the number of occurrences of t in the document and nd is the number of terms in the document. The second element, IDF calculates the importance of a term. The IDF is calculated as follows:

$$\mathrm{IDF}(t) = ln\ ln\ \frac{N}{df_i+1} \tag{2}$$

where N is the number of documents and dfi is the number of documents containing term t. The final result of tf-idf is the multiplication of tf and idf.

2.3.2 Sentiment-related Features

In the sentiment-related features feature set, there are several features taken, namely the sentiment value of the emoji, the sentiment contrast value of the emoji, the word sentiment value and the sentiment contrast value of the word.

2.3.3 Punctuation-related Features

In the set of punctuation-related features, there are several features taken, namely the calculation of the number of occurrences of exclamation marks, question marks, periods, quotation marks, capital letters in words, the number of repetitions of letters in one word.

2.3.4 Lexical Features

In this feature, the number of repetitions of laughter in the text is counted.

2.4 Splitting Data and Oversampling

The datasets that have gone through the data preprocessing and feature extraction processes are then divided into training data and testing data. With a comparison between training data and testing data, which is 8:2.

The data obtained from scraping comments on Instagram is very diverse, so the class of the dataset obtained is not balanced. In the process of sentiment analysis, there are more data with negative classes than the other two classes. In the sarcasm prediction process, the data with the non-sarcasm class is much more than the sarcasm class. In order to overcome the problem of unbalanced data, the SMOTE oversampling library is used which duplicates samples from the minority class.

2.5 Sentiment Analysis

Sentiment analysis is a text analysis technique to detect the polarity of a text in a document, paragraph, sentence, or clause. Sentiment analysis is often used to detect sentiment in social data, measure brand reputation, and understand customers.

In this research, conducted sentiment analysis with Naïve Bayes and Random Forest classifier.

2.5.1 Naïve Bayes Model

Naïve Bayes is a fast algorithm, high-scale model formation and assessment, can be used for binary and multiclass classification, and lightweight for training because it does not need complicated optimizations (Oracle, 2021). In Bayes' theorem, the conditional probability or probability is expressed as:

$$P(X) = \frac{P(X|H)P(H)}{P(X)} \tag{3}$$

where X is the proof, H is the hypothesis, P(H|X) is the probability that the hypothesis H is true for the proof X, P(X|H) is the probability that the proof X is true for the hypothesis H, P(H) is the prior probability of the hypothesis H , and P(X) is the prior probability of the proof X.

2.5.2 Random Forest Model

Random Forest is one of the ensemble methods in figure 3, that combines a number of k learning models with the aim of creating an improved classification model. The ensemble method tends to be more accurate than the base classifier (Han et al., 2011). Random Forest returns the class prediction results based on the majority vote results from the base classifiers (Decision Tree). The Random Forest method uses the Bagging algorithm, as follows (Han et al., 2011):

- for i = 1 to k do //create k models
- create a bootstrap sample, D_i D
- use D_i and the learning scheme to create a new model, M_i
- endfor

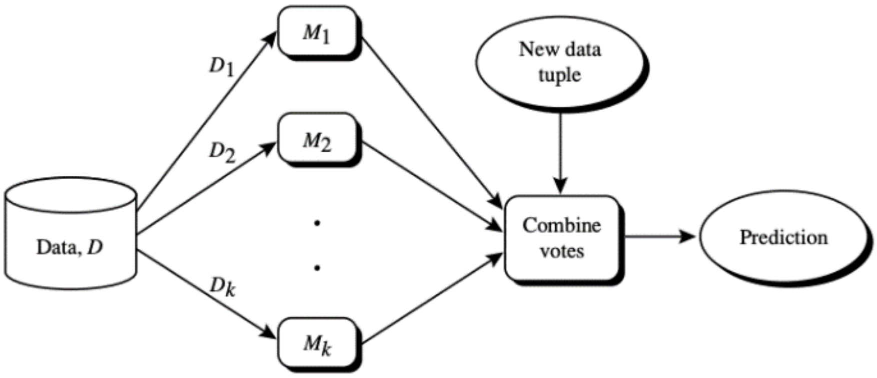

Figure 3 Ensemble Method [10].

2.6 Sarcasm Detection and Changing Sentiment Label

In this research, sarcasm detection is done with Random Forest algorithm. After label from sarcasm detection retrieved, the sentiment label and sarcasm label is being checked. If the data is sarcasm and the sentiment is positive or netral, the sentiment is changed to negative sentiment.

2.7 Model Evaluation

The results of the evaluation of the data and their classification can be represented in a 2x2 matrix called the Confusion Matrix (Table 3).

Table 3 Confusion Matrix

Prediction	Actual	
	Positive	Negative
Positif	*True Positive*	*False Positive*
Negatif	*False Negative*	*True Negative*

The accuracy value can be calculated by dividing the number of correct classification results by the sum of all data with the equation:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{4}$$

Precision is the level of accuracy between the information requested and the answer given by the system. Equation of precision:

$$precision = \frac{TP}{TP+FP} \tag{5}$$

Recall is the success value of the system to retrieve information with the equation:

$$recall = \frac{TP}{TP+FN} \tag{6}$$

Then the f1-score shows the performance of precision and recall:

$$f1 - score = 2\frac{precision\ x\ recall}{precision+recall} \tag{7}$$

3. RESULTS AND DISCUSSION

3.1 Sentiment Analysis With Random Forest

In this Random Forest model, hyperparameter tuning is performed to determine the parameters in order to obtain the best model. The results of the Random Forest parameters obtained are 'n_estimators' as much as 1400, 'min_samples_split' as much as 2, 'min_samples_leaf' as much as 1, 'max_features' with a value of 'sqrt', 'max_depth' as much as 80, 'bootstrap' with a boolean value of False. The time required for training data, predictions on testing and evaluation data is 21772.37 seconds. The accuracy of the Random Forest model using the above parameters is 72%.

3.2 Sentiment Analysis With Naïve Bayes

In this Multinomial Naive Bayes model, hyperparameter tuning is performed to determine the parameters in order to obtain the best model. The result of the Multinomial Naive Bayes parameter obtained is 'alpha' with a value of 0.00001. The time required for training data, predictions on testing and evaluation data is 14.48 seconds. The accuracy of the Random Forest model using the above parameters is 61%.

3.3 Sarcasm Detection With Random Forest

In this Random Forest model, hyperparameter tuning is performed to determine the parameters in order to obtain the best model. The results of the Random Forest parameters obtained are 'n_estimators' as much as 1400, 'min_samples_split' as much as 2, 'min_samples_leaf' as much as 1, 'max_features' with a value of 'sqrt', 'max_depth' as much as 80, 'bootstrap' with a boolean value of False. The time required for training data, predictions on testing and evaluation data is 19183.17 seconds. The accuracy of the Random Forest model using the above parameters is 83%.

3.4 Sentiment Label Changed Results

The label results from sentiment prediction using Random Forest and Naive Bayes and then checking for sarcasm from the text. If the text is predicted to be sarcasm and the sentiment is neutral (0) or positive(1), then the sentiment value will be changed to negative (-1). A comparison of the evaluation of the Random Forest and Naive Bayes sentiment analysis model after the label was changed can be seen in table 4.

Table 4 Comparison of sentiment analysis model

	Random Forest	Naive Bayes
No Sarcasm detection	0.72	0.61
With Sarcasm detection	0.71	0.60

4. CONCLUSION

After all the research steps have been carried out, the following conclusions can be drawn is the best accuracy in the sentiment analysis model is obtained using Random Forest with an accuracy of 71%. The accuracy of the sarcasm model with Random Forest is 83%.Sentiment analysis without sarcasm detection obtained better results in both models, Random Forest and Naive Bayes. The result of sentiment analysis accuracy without sarcasm detection is one percent higher than sarcasm detection.

REFERENCES

[1] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends Inf. Retr.*, vol. 2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[2] B. Y. Liau and P. Tan, “Gaining customer knowledge in low cost airlines through text mining,” *Ind. Manag. Data Syst.*, vol. 114, pp. 1344–1359, 2014.

[3] D. Alita and A. Isnain, “Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier,” *J. komputasi*, vol. 8, 2020, doi: 10.23960/komputasi.v8i2.2615.

[4] D. Antonakaki, D. Spiliotopoulos, C. V. Samaras, P. Pratikakis, S. Ioannidis, and P. Fragopoulou, “Social media analysis during political turbulence,” *PLoS One*, vol. 12, no. 10, pp. 1–23, 2017, doi: 10.1371/journal.pone.0186836.

[5] D. Maynard and M. A. Greenwood, “Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis,” 2014.

[6] Y. Yunitasari, A. Musdholifah, and A. Sari, “Sarcasm Detection For Sentiment Analysis in Indonesian Tweets,” *Indones. J. Comput. Cybern. Syst.*, vol. 13, pp. 53–62, 2019.

[7] M. Bouazizi and T. Ohtsuki, “A Pattern-Based Approach for Sarcasm Detection on Twitter,” *IEEE Access*, vol. 4, p. 1, 2016, doi: 10.1109/ACCESS.2016.2594194.

[8] N. Couldry, S. Livingstone, and T. Markham, “Media Consumption and Public Engagement: Beyond the Presumption of Attention,” 2007, doi: 10.1057/9780230800823.

[9] I. López, S. de Maya, and L. Warlop, “When Sharing Consumption Emotions With Strangers Is More Satisfying Than Sharing Them With Friends,” *J. Serv. Res.*, vol. 17, 2014, doi: 10.1177/1094670514538835.

[10] J. Han, M. Kamber, and J. Pei, “9 - Classification: Advanced Methods,” in *Data Mining (Third Edition)*, Third Edition., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 393–442.