Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter

Meylan Wongkar[1], Apriandy Angdresey[2]
Department of Informatics Engineering
De La Salle Catholic University
Manado 95000, Indonesia
Email: 14013016@unikadelasalle.ac.id[1], aangdresey@unikadelasalle.ac.id[2]

***Abstract*— Sentiment analysis is an activity carried out to see the level of public sentiment or public opinion relating to goods or services and even a figure, both political and celebrity figures. In this study, a sentiment analysis application for twitter analysis was conducted on 2019 Republic of Indonesia presidential candidates, using the python programming language. There are several steps taken to conduct this sentiment analysis, which is to collect data using libraries in python, text processing, testing training data, and text classification using the Naïve Bayes method. The Naïve Bayes method is used to help classify classes or the level of sentiments of society. The results of this study found that the value of the positive sentiment polarity of the Jokowi-Ma'ruf Amin pair was 45.45% and a negative value of 54.55%, while the Prabowo-Sandiaga pair received a positive sentiment score of 44.32% and negative 55.68%. . Then the combined data was tested from the training data used for each presidential candidate and get an accuracy of 80.90% ≈ 80.1%. In this study a comparison was carried out using the naïve bayes, svm and K-Nearest Neighbor (K-NN) methods which were tested using RapidMiner by producing a naïve bayes accuracy value of 75.58%, svm accuracy value of 63.99% and K-NN accuracy value of 73.34%.**

***Keywords*: Sentiment analysis, Naïve Bayes algorithm, text mining, Twitter.**

I. INTRODUCTION

Sentiment is a term used to describe a topic that is subjective and objective and a factual or non-factual topic that transcends the difference between a positive or negative topic [1]. Sentiment analysis is an analysis carried out based on rumors or gossip circulating [2]. Sentiment analysis is an analytical approach used to analyze a text. The purpose of this sentiment analysis is to determine a subjectivity of opinion, the result of a review or a tweets. Based on sentiment analysis, opinions from someone can be classified into various categories based on data size and document type [3].

Nowadays, the community often provides responses and criticisms of leaders, both political figures and public figures through social media such as twitter. Twitter is one of the social media that has a retweet feature that can be used by every user to re-upload information or tweets which allows the dissemination of information on social twitter media to be faster [2]. Twitter is also a social media that can be used to sentiment analysis using data tweets obtained by doing crawler data. Data crawler is a method used to collect data. In this study, the author aims to analyze the level of sentiment from the community towards the 2019 presidential candidates of the Republic of Indonesia obtained from the public on Twitter social media, by doing crawler data. Furthermore, the author will make a comparison of the accuracy of the Naïve Bayes method, with other classification methods such as SVM and KNN. Naïve Bayes method [3] is a method used to group data according to the categories that already exist.

The paper is organized as follows, part II will be explained about related work and information related to sentimental analysis sentiment on twitter. Part III describes the methods that present the formulas used for classification on sentimental analysis sentiments on twitter. Part IV describes the performance evaluation that contains the results of research that has been done and section V concludes the results of the research that has been done.

II. RELATED WORK

In this study analysis sentiments were carried out for factors related to customer satisfaction with e-commerce. These factors can then help e-commerce companies focus on improving service and company quality that will be associated with increased traffic, sales, and company profits. Then do data collection, data cleaning, and lexicon classification. The three stages will be processed using R Studio, which is software application that uses the basis of the R programming language. Based on the results of the sentiment analysis on the three largest e-commerce companies in the world, namely Amazon, Ebay, and Rakuten, it can be concluded that several factors that influence customer satisfaction that get more customer attention are use fullness, service quality, information quality, and system quality [4]. Then in [10] the sentiment analysis was performed by comparing the SVM and KNN methods. The tested data consists of various amounts to see differences in the level of accuracy in each dataset. Data sets were tested, the first one using 100 data tweets, then the second 500 data tweets, the third one 1000, then 1500 data tweets, after that using 2000 data tweets, 2500 data tweets and the last 3000 data tweets. Based on the research, the results of accuracy for the method are higher than the svm method.

In [5],[9] conducted a twitter sentiment analysis to see the level of sentiment in twitter users using the K-Nearest Neighbor (KNN) method and analyzed the level of community sentiment towards the performance of Malang City (SAMSAT) using the Naïve Bayes classifier method. In the study [5] the results were collected by crawling data on twitter and then doing text processing so that the data was ready to be analyzed using the K-NN method. Based on the tests that have been carried out, the highest accuracy value is produced, namely 67.2% and the precision value of 56.94% in the test using the value = 5, the highest recall value is 78.24% in the test with a value of k = 15. While in paper [9], the collection process data used web scrapping techniques to retrieve data from twitter and then save it to the database. Then in the data processing, the author performs duplicate tweets filtering which functions to delete repeated tweets, folding cases to convert all letters to lowercase letters, cleaning to clean data tweets from characters or words that are not needed, tokenizing to separate word words, filtering to retrieve important words from the results of tokens. Based on the research that has been done in the first stage of the class category positive, negative and neutral opinions were obtained 81%, 89%, 80% and in the second test the results for the positive, negative and neutral categories were 82%, 92% and 80% [5] [9].

In the study [7], [8] analyze the tweets using Indonesian language was conducted to determine the level of public sentiment towards a film and to public figures on Twitter ahead of the 2014 general election in Indonesia using the Naïve Bayes classifier method. In the sentiment analysis about film [7] words were corrected on twitter with text processing and then testing non-standard sentences. There are 140 opinion data as training data consisting of 70 positive opinion data and 70 negative opinion data and 60 test data consisting of 30 positive opinion data and 30 negative opinion data. The test was carried out 3 times and produced the highest accuracy on the third stage testing with an accuracy value of 86.67%. Whereas in the research [8] the list of tweets obtained was obtained by using the cron job facility on the Windows operating system. At the data processing stage the author performs cleaning data by deleting special characters, URLs and eliminating word affixes. Based on 1329 data tested tweets obtained classification testing results with the term frequency feature obtained at 79.91% while the TF-IDF feature obtained an accuracy of 79.68%. The classification using the RapidMiner tools with Naive Bayes and the term frequency feature obtained was 73.81%, while the TF-IDF feature was obtained at 71.11% [7] [8].

III. Methods

There are several processes that are carried out in this text processing: firstly we collect data, in this study we using data tweets are collected from Twitter social media by using a crawler. Furthermore, we parse the tweets are get by describing it verbatim. Hereinafter, we do the tokenization process that is cleaning the tweet and selecting the meaningful words. Then, we do text mining using naïve bayes method, the process of text processing can be seen in Figure 1.

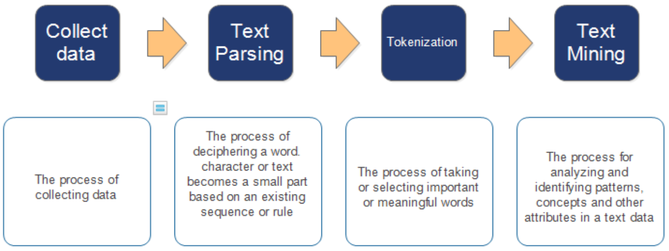

Collect data
Text Parsing
Tokenization
Text Mining
The process of collecting data
The process of deciphering a word character or text becomes a small part based on an existing sequence or rule
The process of taking or selecting important or meaningful words
The process for analyzing and identifying patterns, concepts and other attributes in a text data

Figure 1. Text Processing

a. Collect data

In this data collection process, we use data tweets are obtained using crawler data from Twitter taken from January to May 2019. In Table 1 presents sample data taken from Twitter.

Table 1. Sample Data

No.	Sentiment	Text
1	Positive	the president has worked well
2	Negative	the president cannot keep his promises
3	Positive	president helps flood victims
4	Negative	the president raised the price of fuel oil
5	Negative	the president raised the electricity tariffs
6	Negative	the president was unsuccessful
7	Positive	president built infrastructure
8	Positive	president spent his holiday with his family
…	…	….
443	Negative	president of imaging

b. Text parsing and Tokenization

The process of processing sentences into several words that have been separated from characters and taken words that have value. Table 2 is the results from the process of text parsing and tokenization of the sample data Table 1.

Table 2. Tokenization

No.	Sentiment	Text
1	Positive	[the, president, has, worked, well]
2	Negative	[the, president, can, not, keep, his, promises]
3	Positive	[president, helps, flood, victims]
4	Negative	[the, president, raised, the, price, of, fuel, oil]
5	Negative	[the, president, raised, the, electricity, tariffs]
6	Negative	[the, president, was, unsuccessful]
7	Positive	[president, built, infrastructure]
8	Positive	[president, spent, his, holiday, with, his, family]
…	…	…
443	Negative	[president, of, imaging]

c. Text mining

The following is an explanation of the text processing process using the Naïve Bayes method. Then calculate the value of the class probability by dividing the number of class data with the total or number of existing documents.

$$P(\textit{positive}) = \frac{7}{13} = 0{,}53.$$

$$P(\textit{negative}) = \frac{6}{13} = 0{,}46.$$

The following are test data that will be tested using training data that has been previously cleaned, as shown in Table 3. Then, the value of $P(a_j|v_j)$ will be determined from the existing test data. The following is the testing phase of test data No.1 to determine the value of $P(a_j|v_j)$ in the sentence in the positive class, as presented in Table 4. Further, the stage of testing in data test no.1 to determine the value of $P(a_j|v_j)$ in the sentence in the negative class, as presented in Table 5.

Table 3. Test Data

No.	Text	Sentiment
1	a good president	?
2	not a good president	?

Table 4. Probability Table for Test Data No.1 (Positive)

Term	*n*	*nc*	*p*	*m*
A	1	0	0,15	2
good	1	1	0,15	2
president	13	7	1,07	14

$$P_1(positive|a) = \frac{0+2.0,15}{1+2} = \frac{0,3}{3} = 0,1$$

$$P_1(positive|good) = \frac{1+2.0,15}{1+2} = \frac{1,3}{3} = 0,43$$

$$P_1(positive|president) = \frac{7+14.1,07}{13+14} = \frac{21,98}{27} = 0,81$$

Table 5. Probability Calculation Table on No.1 (Negative) Test Data

Term	*n*	*nc*	*p*	*m*
A	1	1	0,15	2
Good	1	0	0,15	2
president	13	6	1,07	14

After that, all the results of the predetermined class probabilities are multiplied to conclude the class classification results from the test data that has been tested.

V_1(*positive*) = 0,1 x 0,43 x 0,81 = 0,03483.

V_1 (*negative*) = 0,43 x 0,1 x 0,77 = 0,03311.

Based on the results of calculations that have been made, it can be concluded that the No.1 test data entered into the positive class. In Table 6 present the test data no. 2 to determine the value of $P(a_j|v_j)$ in the sentence in the positive class. Whereas, the following will be tested on test data no.2 to determine the value of $P(a_j|v_j)$ in the sentence in the negative class, presented in Table 7.

Table 6. Probability Calculation Table on Test Data No.2 (Positive)

Term	*n*	*nc*	*p*	*m*
not	1	0	0,15	2
a	1	0	0,15	2
good	1	1	0,15	2
president	13	7	1,07	14

$$P_2(positive|not) = \frac{0+2.0,15}{1+2} = \frac{0,3}{3} = 0,1$$

$$P_2(positive|a) = \frac{0+2.0,15}{1+2} = \frac{0,3}{3} = 0,1$$

$$P_2(positive|good) = \frac{1+2.0,15}{1+2} = \frac{1,3}{3} = 0,43$$

$$P_2(positive|president) = \frac{7+14.1,07}{13+14} = \frac{21,98}{27} = 0,81$$

Table 7. Probability Calculation Tables on No.2 (Negative) Test Data

Term	*n*	*nc*	*p*	*m*
not	1	1	0,15	2
a	1	1	0,15	2
good	1	0	0,15	2
president	13	6	1,07	14

$$P_2(negative|not) = \frac{1+2.0,15}{1+2} = \frac{1,3}{3} = 0,43$$

$$P_2(negative|a) = \frac{1+2.0,15}{1+2} = \frac{1,3}{3} = 0,43$$

$$P_2(negative|good) = \frac{0+2.0,15}{1+2} = \frac{0,4}{3} = 0,1$$

$$P_2(negative|president) = \frac{6+14.1,07}{13+14} = \frac{20,98}{27} = 0,77$$

After that, all the results of the predetermined class probabilities are multiplied to conclude the class classification results from the test data that has been tested.

V_2(*positive*) = 0,1 x 0,1 x 0,43 x 0,81 = 0,0034.

V_2 (*negative*) = 0,43 x 0,43 x 0,1 x 0,77 = 0,0142.

Based on the results of calculations that have been done, it can be concluded that the test data No.2 entered into the negative class.

Table 8. Test Data Test Results Table

No.	Sentiment	Text	Sentiment Result
1	Positive	A good president	Positive
2	Negative	Not a good president	Negative

Furthermore, the calculation is done to see the value of accuracy, precision, recall, and f_1-score from the results of the analysis. For this reason, the value of TP, TN, FP and FN is determined. For TP and TN values taken from the initial assumptions of the test data where in this study the value of TP is 1 and TN is 1. While, FP and FN are taken from the results of the classification of test data with FP values are 1 and FN is 1. After the text mining process, the data is classified using the naïve Bayes method. Naïve Bayes Method is a method that can be trained or used on small-scale data and can provide predictive results in real-time. The naïve Bayes method can also help in classifying a class whose results can be used in parallel in increasing the scale of the dataset, especially in large-scale data case studies [10].

In this research, the classification of every sentiment from the community is present on social media. The following is the equation formula of the Naïve Bayes methods [14]:

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)} \qquad \text{. . . (2.1)}$$

x = Data with an unknown class.

c = The data hypothesis is a specific class.

P(c|x) = Probability of hypothesis based on condition (posteriori probability).

P(c) = Probability of hypothesis (prior probability).
P(x|c) = Probability based on conditions in the hypothesis.
P(x) = Probability of value *c*.

IV. PERFOMANCE EVALUATION

A. Experimental Setup

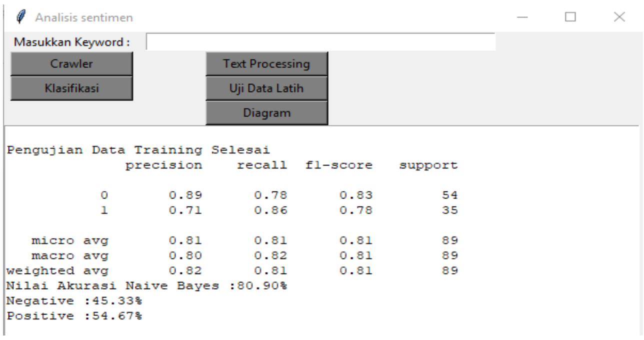
In this study, we using the data are collected through social media twitter, by collecting all the tweets related to the Republic Indonesia of presidential candidate pair period 2019 - 2024, where the data began to be collected during the presidential election campaign in 2019 until after the election period. Then the number of data are obtained as 443 with sentiment attributes that contain information, while for labels consists of positive and negative labels.

Table 9. Sample Data

No.	Text	Sentiment
1	jokowi ma'ruf winner successful perfect full	Positive
2	second term free as a bird or a lame-duck president nice analysis	Positive
3	it seems that 17 million votes have been confirmed in central and east java all in favor of jokowi	Positive
4	prabowo subianto argues jokowi campaign stole votes during election	Negative
5	let’s be honest both two parties, jokowi and prabowo, do inappropriate ways of campaign	Negative
6	i hate jokowi	Negative
7	do you see what's the difference between prabowo and jokowi's speech? jokowi didn't forget to say thanks to all participant	Positive
8	love my president	Positive
…	....	…
443	media today not doing journalistic but be a tool of a ruler they degrade themselves	Negative

B. Experimental Result and Discussion

The following is the result of a comparison of the value of accuracy made on the Naïve Bayes method, SVM and KNN which can be seen in Figure 4, where the accuracy of the Naïve Bayes method is better with the results of 80.90% ≈ 80.1%, compared to the KNN which has an accuracy rate of 75.58% and with the lowest accuracy value of 63.99% using SVM. Moreover, In Figure 2 the results of the training data are explained using the Naïve Bayes method which produces an accuracy value of 80.90% ≈ 80.1% and a precision value for the positive class represented by the number 0, which is 0.89 and the others is number 1 for the negative class with a score of 0.71%, the value of the positive class is better than the negative class, significant things can be seen from the positive class support values above 50. Meanwhile, in Figure 3 we can see the comparison of the accuracy of the Naïve Bayes, KNN and SVM methods. The accuracy of the predictions made indicates that the accuracy of the Naïve Bayes is better than both methods that is 80.9%. While, the level of accuracy by KNN is only 75.58% and the lowest accuracy level is SVM with 63.99%.



```
Pengujian Data Training Selesai
                precision    recall  f1-score   support

           0       0.89      0.78      0.83        54
           1       0.71      0.86      0.78        35

   micro avg       0.81      0.81      0.81        89
   macro avg       0.80      0.82      0.81        89
weighted avg       0.82      0.81      0.81        89
Nilai Akurasi Naive Bayes :80.90%
Negative :45.33%
Positive :54.67%
```

Figure 2. Data Training Result of Naïve Bayes

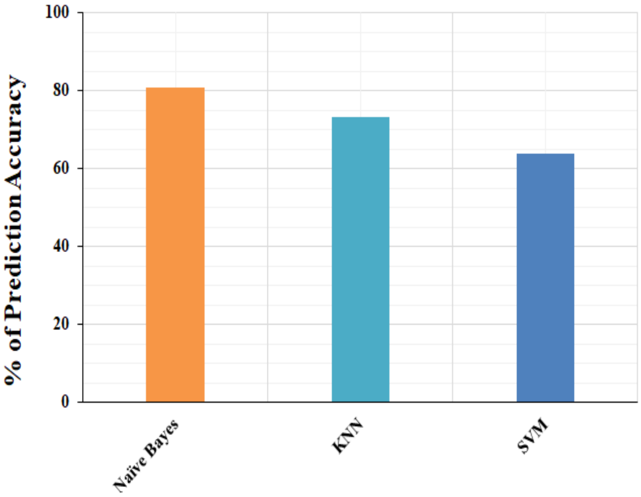


Figure 3. The Comparison of Level Accuracy

Figure 4 shows the results of calculating the comparison of the precision prediction and recall values of each presidential candidate using Naive Bayes, namely for the precision value of the positive class of the Jokowi-Ma'ruf is lower than that Prabowo and Sandiaga, but on the contrary for the recall value of positive class. Whereas, the precision value of negative class from Jokowi - Ma'ruf is slightly higher than that of Prabowo and Sandiaga, whereas for the recall value.

In Figure 5 can be seen the results of the comparison of the precision and recall values in each positive and negative class of each method used, for the precision value in the positive class Naive Bayes is still better, however for the recall of positive classes SVM has a better value but not very significant. Meanwhile, the precision value of negative Naive Bayes class and SVM has almost the same value, on the contrary for the recall value of negative Naive Bayes class is better.

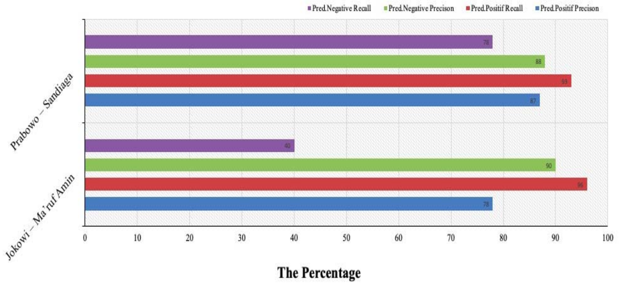


Figure 4. The Result of President Candidate using Naïve Bayes

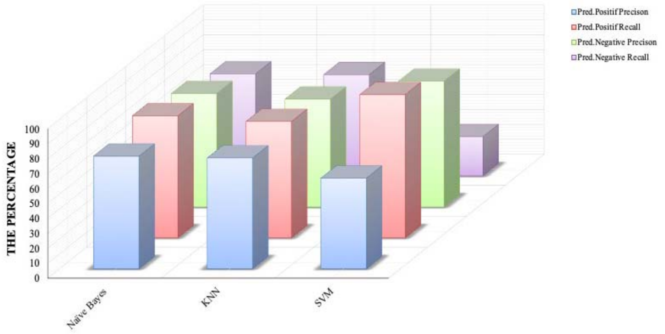


Figure 5. The Testing Performance of Data Training

V. Conclusions

In this paper, we discuss the sentiment analysis of public towards the Republic of Indonesia's presidential candidates for the 2019-2024 period, using tweet data obtained from social media: Twitter, by crawlers. In addition, we do the text processing from data obtained and use Naive Bayes method to predict the class. Afterward, compare with other methods such as SVM and KNN. We classify by two classes namely positive and negative. From the results of our experiments, it can be seen that the Naïve Bayes method has a better accuracy level (i.e. 80.90%) compared to using other methods, such as KNN which only has an accuracy rate of 75.58% and an accuracy rate using SVM which is 63 99%. For future work, we plan to analyze sentiment of public satisfaction toward the performance of the elected president of the Republic of Indonesia, using the data from another social media, such as Facebook and Instagram.

VI. References

[1] F. A. Pozzi, E. Fersini, E. Messina and B. Liu, in *Sentiment Analysis In Social Network*, United States, Todd Green, 2017, p. 228.

[2] S. Widoatmodjo, in *Cara Cepat Memulai Investasi Saham Panduan Bagi Pemula*, Jakarta, Kompas Media, 2012, p. 139.

[3] Rajput, D. Singh, Thakur, R. Singh, Basha and S. Muzamil, *Sentiment Analysis and Knowledge Discovery in Contemporary Business*, United States of America: IGI Global, 2018.

[4] C. A. Haryani, H. Tihari, Marhamah and Y. A. Nurrahman, Sentimen Analisis Kepuasan Pelanggan E-commerce Menggunakan Lexicon Classification dengan R, in Konferensi Nasional Sistem Informasi, Pangkalpinang, 2018.

[5] A. Deviyanto and M. D. Wahyudi, Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor, JISKa (Jurnal Informatika Sunan Kalijaga), Vols. Vol. 3, No. 1, no. ISSN : 2527-5836, p. 1–13, 2018.

[6] I. F. Rozi, E. N. Hamdana and M. B. I. Alfahmi, Pengembangan Aplikasi Analisis Sentimen Twitter Menggunakan Metode Naive Bayes Classifier (Studi Kasus SAMSAT Kota Malang), Jurnal Informatika Polinema, Volume 04, Edisi 02, no. ISSN: 2407-070X, 2018.

[7] P. Antinasari, R. S. Perdana and M. A. Fauzi, Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 1, No. 12, no. e-ISSN: 2548-964X, pp. 1-9, Desember 2017.

[8] A. F. Hidayatullah and A. SN, Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter, Seminar Nasional Informatika 2014 (SemnasIF 2014) , no. SSN: 1979-2328, 2014.

[9] M. R. Huq, A. Ali and A. Rahman, *Sentiment Analysis on Twitter Data using KNN and SVM*, (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, 2017.

[10] G. Chakraborty, M. Pagolu, S. Garla, *Text Mining And Analysis Practical Methods, Examples, And Case Studies Using SAS*, North Carolina, USA: SAS Institute Inc., 2013.