Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 6 No. 1, 2021, pp. 40 – 45

Sentiment Analysis on Indonesia Twitter Data Using Naïve Bayes and K-Means Method

Ajib Susanto*[1], Muhammad Atho'il Maula[2]
Dept. Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia 50131
E-mail : ajib.susanto@dsn.dinus.ac.id[1], 111201307805@mhs.dinus.ac.id*[2]
**Corresponding author*

Ibnu Utomo Wahyu Mulyono[3], Md Kamruzzaman Sarker[4]
Dept. Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia 5013, Kansas State University, Manhattan, KS 66506, United States
E-mail : ibnu.utomo.wm@dsn.dinus.ac.id[3], sm.kamruzzaman@gmail.com[4]

Abstract - This study focuses on the analysis of sentiments on Indonesian twitter data. Twitter data on Indonesian simultaneous pilkada used to get its sentiments using Naïve Bayes Classifier method as a method of classification and K-means method to get Label on the data train process. Combining the two methods is expected to get high accuracy results. The results obtained from the research shows a pretty good accuracy of 74.5%.

Keywords - Sentiment Analysis, Naive Bayes, K-Means, Indonesian Election Tweet, Classification

1. INTRODUCTION

Information and technology has grown rapidly. From this development one can easily get or share information. For example, users of social media twitter in Indonesia ranked 3rd in the world. Twitter is a social media used to write and share short messages or commonly called a tweet. These tweets are usually intended to express something that is their concern. Thus the opinions contained in twitter can be used as a source of research material data because the information contained in twitter is very valuable as a determinant of policy. The math is about 85% of the data is unstructured data[1]. So needed a system development that can handle the problem.

Sentiment analysis can be applied in extracting information contained in unstructured data. This method is a possible method to classify the polarity of an unstructured data such as a document or comment, ie whether the document is positive, negative, or neutral[2]. So opinions from twitter will be grouped into positive classes if the opinion is of good value. And conversely if those opinions have bad value then it will be grouped into negative class.

One of the methods used for the analysis of sentiments is the Naïve Bayes Classifier. This simple method is also one of the fastest working methods[3]. Results from related studies also show high accuracy results[1][4]. This method divides the data into two parts: training data and test data.

The training data on naïve bayes will be labeled as the reference of the test data. But there will be problems if the labeling is done manually, because it will be questioned how the validity of the data.

K-means can be applied in labeling existing on the train data. This method in the research that has been done shows the results of high accuracy[5]. In this research will be used

Naïve Bayes Classifier method as a method of classification on sentiment analysis and K-Means method on labeling process on train data.

2. RESEARCH METHOD

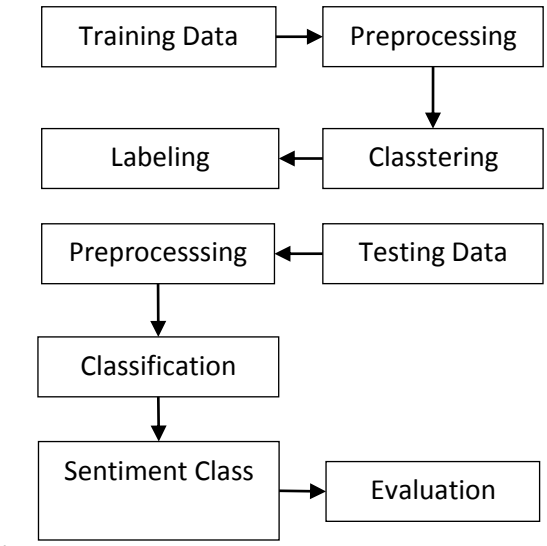

Figure 1. Research Method

2.1. Data

The data used in this research is Indonesian language twitter data which discusses about the election of regional head in Indonesia. The data used are 700 data, with the data sharing train as much as 500 data and data test 200 data.

2.2. Prepocessing

Preprocessing is used to prepare the data to be ready for processing at a later stage. The process includes case folding, normalization, tokenisasi, stopword removal, and stemming in Indonesian[4].

In the process of training data is done weighting the word used to give weight to each document using a sentiment dictionary of research conducted[6]. This weighting is based on positive words and negative words contained in the document.

The normalization[7] process is used to clean up features that are usually included in tweets, such as hashtags, mentions and links. In the stopword removal process is done to eliminate words that have no meaning and if it is removed will not remove the important information contained in the document.

The stemming[8] process is done to convert non-standard words into standard words. It also removes the affixed words contained in one Suffixes document ("it", "-mu", "-ku", "-kah", "lah"), prefix ("ke-," "di-" , "A"). The basic dictionary used in the stemming process is based on http://kateglo.com.

2.3. K-Means

The process of the K-Means method in five steps[9]:

1. Determining the number of clusters
2. Allocate data into clusters at random

3. Calculate the centroid data in each cluster
4. Allocate each of the closest centroids
5. Repeat step -3, if the data is still there that change repeat until there is no change in the data.

2.2. Naïve Bayes Classifier

This method is simple, but has high accuracy accuracy. Because this method is included in supervised learning, so the Naive Bayes method requires early knowledge to determine predictions. Th[10]is method is formulated as follows:

$$P(A|B) = \frac{P(B|A)\,.\,P(A)}{P(B)} \qquad (1)$$

The tweets contained in this method will be represented by the attribute pair x_1, x_2, x_3, ... x_n. Where x_1 is the first word of the document, x_2 is the second word and so on. The Naive Bayes method will search for the highest probability of performing the classification process in the tweets to be tested (VMAP).

$$\boldsymbol{V_{MAP} = \underset{V_j e\, V}{\operatorname{argmax}} \prod_{n=1}^{n} P(x_i|V_j)\, P(V_j)} \qquad (2)$$

$$\boldsymbol{P(x_i|V_j) = \frac{n_k + 1}{n + |kosakata|}} \qquad (3)$$

3. RESULTS AND DISCUSSION

Data obtained from the crawling results of the Indonesian language twitter which discussed about the election of the district head of Indonesia, using the keyword "pilkada2017". This research will be implemented using PHP programming language with the help of several libraries.

In the Naïve Bayes Classifier method the data will be divided into two processes, namely data processing and test process data. Here is an example of preprocessing data:

Table 1. Result of Preprocessing

No	Preprocessing Result
1	ayo hak suara depan golput ya
2	pilih kerjaa jujur bersih
3	apakah bukti curang pilih kepala daerah serentak
4	golput kertas sisa pakai curang
5	tps hujan ahmadi yakin target suara capai
6	moga pilih umum kali safety sukses Indonesia

After passing preprocessing, the training data is then continued with the word weighting phase with the word dictionary[6].

Table 2. World Weigting

No	Preprocessing Results	Weight	
		positif	negatif
1	ayo hak suara depan golput ya	0	1 golput
2	pilih kerjaa jujur bersih	2 Jujur, bersih	0
3	apakah bukti curang pilih kepala daerah serentak	0	1 curang
4	golput kertas sisa pakai curang	0	2 Golput curang
5	tps hujan ahmadi yakin target suara capai	1 yakin	0
6	moga pilihumum kali safety sukses indonesia	1 sukses	0

The clustering process is performed using randomly selected centroids in the training data. Then calculated using Ecludian Distance.

$$\boldsymbol{d(i,j) = \sqrt{\sum_{i=1}^{N}(i,j)^2}} \qquad (4)$$

After iterating with the same result, the calculation of the iteration is stopped with the result of the member on C1 tweet 4 and on the member on C2 tweet 1,2,3,5,6.

Table 3. K-MEANS Label Result

No	Tweet	Label
1	ayo hak suara depan golput ya	Negatif
2	pilih kerjaa jujur bersih	Negatif
3	apakah bukti curang pilih kepala daerah serentak	Negatif
4	golput kertas sisa pakai curang	Positif
5	tps hujan ahmadi yakin target suara capai	Negatif
6	moga pilih umum kali safety sukses indonesia	Negatif

In this document, the test will be D7: "teror jelang pilwali kendari". The Naïve Bayes Classifier method will calculate the data in each category.

$$\boldsymbol{P(x_i|V_j) = \frac{n_k + 1}{n + |kosakata|}} \qquad (5)$$

1. Positive class probability

P(teror | Positif) : $\frac{0+1}{4+32} = 0.028$ (6)

P(jelang | Positif) : $\frac{0+1}{4+32} = 0.028$ (7)

P(pilwali | Positif): $\frac{0+1}{4+32} = 0.028$ (8)

P(kendari | Positif):$\frac{0+1}{4+32} = 0.028$ (9)

2. Negative class probability

P(teror | Negatif): $\frac{0+1}{28+32} = 0.017$ (10)

P(jelang | Negatif): $\frac{0+1}{28+32} = 0.017$ (11)

P(pilwali | Negatif): $\frac{0+1}{28+32} = 0.017$ (12)

P(kendari | Negatif): $\frac{0+1}{28+32} = 0.017$ (13)

Then calculate the probability of P ($x_i|V_j$) P (V_j) to determine the category of the test data.

1. Positive class

P(Positif | uji) =
($c_{positif}$)*P(teror | Positif) *P(jelang | Positif) * P(pilwali | Positif) *P(kendari | Positif)
= 0.16 * 0.0277 * 0.0277 * 0.0277 * 0.0277
= 9.524920026519511e-8

2. Negative class

P(Negatif | uji) =
P($c_{negatif}$)* P(teror | negatif) *P(jelang | negatif) * P(pilwali | negatif) *P(kendari | negatif)
= 0.83 * 0.0166 * 0.0166 * 0.0166 *0.016666
= 6.403296357776138e-8

Based on the calculation that has been done, it can be concluded that the test document included in the negative class.

After clustering process using k-means to get the label used in training data and classification using naïve bayes classifier method, then tested using confussion matrix. This test is used to find the value of error rate and accuracy value. Here is the result of accuracy testing on the train data using confussion matrix with total test data of 200 data:

$$Akurasi = \frac{74.5}{100}\ x\ 100\% \quad (14)$$

= 74.5 %

$$Error\ rate = \frac{25.5}{100}\ x\ 100\% \qquad (15)$$

= 25.5 %

4. CONCLUSION

After the research done, the conclusion is that this research succeeded to implement Naïve Bayes Classifier and K-Means Method on Sentiment Analysis on twitter data of "Pilkada Serentak Indonesia" into positive or negative class, with result of accuracy 74.5% and 25.5% error rate shown on testing using confusion matrix.

REFERENCES

[1] C. Fiarni, H. Maharani, and R. Pratama, "Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique," *2016 4th Int. Conf. Inf. Commun. Technol. ICoICT 2016*, no. May 2016, 2016.

[2] B. A. Sevsa and M. D. R Wahyudi, "Analisis Sentimen pada Indeks Kinerja Dosen Fakultas SAINTEK UIN Sunan Kalijaga Menggunakan Naive Bayes Classifier," *J. Buana Inform.*, vol. 10, no. 2, p. 112, 2019.

[3] L. Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method," *Proceeding 2015 1st Int. Conf. Wirel. Telemat. ICWT 2015*, no. November 2015, 2016.

[4] G. Septian, A. Susanto, and G. F. Shidik, "Indonesian news classification based on NaBaNA," in *Proceedings - 2017 International Seminar on Application for Technology of Information and Communication: Empowering Technology for a Better Human Life, iSemantic 2017*, 2018, vol. 2018-Janua.

[5] A. N. Ulfah, "Analisis Kinerja Algoritma Fuzzy C-Means Dak-Mean Pada Data Kemiskinann," *Skripsi*, vol. 1, no. 2, 2014.

[6] D. H. Wahid and A. SN, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 10, no. 2, p. 207, 2016.

[7] R. Maskat and N. Abdul Rahman, "Categorization of malay social media text and normalization of spelling variations and vowel-less words," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 4, pp. 1380–1386, 2020.

[8] A. F. Hidayatullah, "The influence of stemming on Indonesian tweet sentiment analysis," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2015.

[9] M. Mardiani, "Perbandingan Algoritma K-Means dan EM untuk Clusterisasi Nilai Mahasiswa Berdasarkan Asal Sekolah," *Creat. Inf. Technol. J.*, vol. 1, no. 4, p. 316, 2015.

[10] S. Fajar Rodiyansyah and E. Winarko, "Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, 2013.