Implementation of Sentiment Analysis on Twitter Using Naïve Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election

To cite this article: Yohanssen Pratama *et al* 2019 *J. Phys.: Conf. Ser.* **1175** 012102

Implementation of Sentiment Analysis on Twitter Using Naïve Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election

Yohanssen Pratama, Anthon Roberto Tampubolon, Liana Diantri Sianturi, Rifka Diana Manalu and David Frietz Pangaribuan [1]

Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Sitoluama, Indonesia

Abstract. Developments in information technology are increasing impact on communication and information exchange. One of its development is Twitter. The widespread use of social media has resulted in the availability of a large textual information so that it appears the need of presenting data that will allow users to get accurate information. This research will be conducted sentiment analysis on Twitter against election debates governor Jakarta to know the public response to some public figures. Ahead of the elections, many of the people who responded to a public figure in the social media Twitter. Public figures referred to in the study was the candidate of the Jakarta governor which are Basuki Tjahaja Purnama, Anies Rasyid Baswedan, and Harimurti Agus Yudhoyono. Based on this research, it will use the keyword 'Basuki' or 'Ahok', 'Agus', 'Badwater', 'pilgub', 'election' and 'debat' on Twitter sentiment in Indonesian language. Tweet contains a sentiment that has been collected, will be seen its profile location so it can be considered whether the user was included in the province of Jakarta or outside of Jakarta. Each tweet sentiments of these two categories will then be preprocessing to make it easy to be processed and analyzed. After passing through the phase, it will be calculated for each category with Naive Bayes algorithm in order to get the value of the sentiment of the people residing in the provinces of Jakarta and outside Jakarta to the three candidates for the Jakarta governor.

1. Introduction

The development of information technology is increasingly increasing. The development of information technology is characterized by the emergence of increasingly social media such as Youtube, Facebook, Instagram, Line, Twitter and others. One of the most widely used social media is Twitter. Twitter is one of the social networking services where users can send and read text-based messages with a limit of 140 characters, known as tweets. Based on the infographics of a social network marketing company in Indonesia, Brand24, and based on the CNN Indonesia website (Wednesday, 23/03/2016), the number of Twitter users in Indonesia has reached 50 million with 20 million active users in Indonesia until 2015.

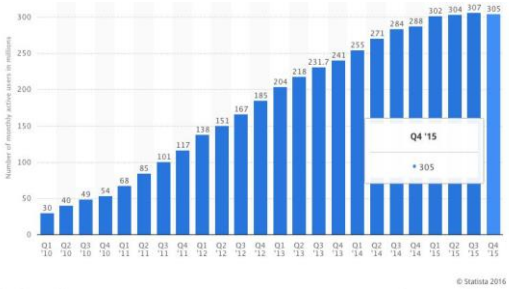

Figure 1. Pertumbuhan pengguna Twitter di Indonesia

Tweet data on Twitter can be a form of perception and opinion. The opinions posted on Twitter can contain many things, one of which is about political activities. One of Indonesia's analytics centers that captures netizen's opinion regarding the popularity and reputation of prospective candidates and strategic issues surrounding DKI Jakarta 2017 elections, Datalyst Indonesia, said that the talk about the elections of DKI in social media in recent years has become a trend of 3.5 million conversations with an amount of about 50,000 each day. Given the amount of textual data it can be done analysis of these data to produce more useful information.

One of the methods that can be used is the analysis of sentiments. One of the most commonly used methods for sentiment analysis is Naive Bayes. Naive Bayes is a probabilistic-based machine learning technique. This method is a simple but high accuracy method for text classification [3].

This is evidenced by the many previous studies that have used this method for classifying sentiments. Xhemali, et al performed a comparison of the three methods of Naive Bayes, Decision Tree, and Neural Networks. The overall research results show that Naive Bayes is the best choice for domain training. Routray et al. and Khairnar & Kinikar discussed many approaches from different researchers, and stated that machine learning methods became an efficient way to analyze sentiments.

Debate is a means of community intelligence and a means of minimizing conflict that can affect the electability of candidates and voter preferences. In the run up to elections 2017 the open debate of Jakarta 2017 election became one of the debates with the highest audience. For the Election Candidate Debate of DKI Jakarta itself, the Election Commission (KPU) of DKI is three times the candidate debate (pilkada.tempo.co).

In this research the researcher will conduct research on tweet data of the society after the three times of the election debate. In this research will be shown diagram depicting positive and negative response of society from DKI Jakarta and Luar DKI Jakarta to three candidate of governor of DKI Jakarta.

2. Data Collection, Labeling on Training Data

Data source as data development test used in this research is taken from the status or tweet of Indonesian society on social media Twitter. The selection of tweets is based on the emergence of the keywords 'Basuki Tjahaja' or 'Ahok', 'Agus Harimurti', 'Anies Baswedan' or 'Anies Rashid', 'pilgub', 'pilkada', and 'debate'.

In Indonesia itself has been debated three times. Based on that, the data will also be taken based on the timing of the three debates. So that there will be debate data 1 which took place on Monday 13 January 2017, debate 2 on Friday 27 January 2017, and debate 3 on Friday 10 February 2017.

Tweet data retrieval used in this analysis is an Indonesian language tweet from Indonesian society which will then be divided into two categories ie tweet from Twitter users with profile location comes from DKI Jakarta province and outside DKI Jakarta.

3. Data Preprocessing

Before entering the main process or applying the algorithm that will be used in research on the data, need to do the initial processing of preprocessing stage. Preprocessing is done to avoid less than perfect data, data interruptions, and less consistent data. The preprocessing process is done by using a program built by researchers using the Phyton programming language. This program will run the normalization stage of text, case folding, non-alphanumeric removal, stop words removal, and stemming.

1. Normalization of Text

In tweets there are many non-standard words. Of course, it is necessary to convert non-standard words into standard words to simplify the process of sentiment analysis. Non-standard words can be character looping, abbreviated words, and emoticons. For the case of the character iteration itself is necessary to do the removal of repetitive characters.

In the process of normalization of the text there are three stages of replacement with the default word, repetitive character disappearance and replacement with the default word back. Replacement with default words is done first to avoid some cases like 'ttg', 'ttd', and others.

If repeating characters are done repeatedly then replaced with the default word then a word can have a different meaning. For example, 'sign' which means the signature will be 'td' with that meaning.

For the next stage is the replacement stage with the default word back. This is done because there are cases where words that have been in accordance with the standard word but have repetitive characters become changed again. For example the word 'planning' is changed to 'planning' so it needs to be replaced to the default word again.

2. Case Folding

The preprocessing process starts from the case folding stage. The process of case folding is the process of converting letters into text into lower case [5]. It is intended that the same word but have different case letters can be grouped into one attribute at the time of N-gram representation.

3. Non-alphanumeric Removal

The non-alphanumeric removal process is performed to remove non-alphanumeric characters or symbols on tweets. This is because the symbol has no information or semantics in the text. In a non-alphanumeric removal process, the program will separate words by connecting characters, such as punctuation and discarding all non-letter characters or numbers. Non-alphanumeric can also consist of a regular expression pattern (Regex Pattern), but should not contain spaces. The Regex Pattern example is as follows:

Table 1. Regex Pattern

Type	Example	Regex Pattern
URL	http://www.google.com, https:/students.del.ac.id	http://.* dan https://.*
Hashta	#wow, #curcol, #love, #senangsekali, #semangatya	#.*
Mention	@will.gozali, @lianacleolin, @syahputrisyahputra	@.*
Angka	021, 545465436436, 2130003232	[0-9] +

The separation of words is indicated by the spaces between the two words performed in order to generate the attributes to be used in the process of N-gram representation. Here is the result of non-alphanumeric removal process using case folding data.

4. Stop Words Removal

The next preprocessing step is stop word removal which is the phase of eliminating meaningless words like link remove, repeated tweets or retweets, mentions, hashtags, and words that have no effect in the extraction of a tweet's sentiments. Stop words are vocabulary that is not a feature of a sentiment. For example "in", "by", "from", "an", and others as [4].

In this study, researchers used Lucene's Indonesian Stopwords (attached to appendix). The input for stop word removal stage, the data used is the data generated at the preprocessing stage of non-alphanumeric removal. On each token or word will be compared with the data stop words are made. If a word in a sentence appears as a stop words, the stop words contained in the sentence will be deleted.

5. Stemming

Stemming is the process of mapping and decomposing the form of a katamenjadi its basic form. Stemming is used to change the shape of a word into the word of the word in accordance with good and correct Indonesian morphology structure. The purpose of the stemming process is to remove the affixes that exist in each word. Stemming is done to the following elements [1]:

- prefix : meng-, di-, per-, ber-, ter-, per-, example: saving, beating, flaming
- suffix : -an, -kan, -i, -nya, -ik, -is, -if, -al, - (is) ation, -al, -iah, -wi, -wiah, -isme , -the receptionist example: food, clothing, perfectionist, normalization
- Infix: -el-, -em-, -er-
  example: serrations, trembling, index finger
- affixes consisting of prefixes and suffixes (confix): ke an, ke-i
  example: ownership, belongs
- particle: -what is it?
  example: who, when, when, when, is
- pronouns: I-, you-, -you, him
  examples: his, my book, your knife, his wallet

Although the particle and pronoun do not include affixes but are treated equally so that the particles are regarded as suffixes and pronouns are regarded as prefixes or suffixes according to their position. Examples of stemming processes in the word 'support', 'supported', and 'support' are the basic verb 'support'. After the stemming stage is complete it will generate basic verbs, single nouns and adjectives.

For the stemming process in this study, researchers used Indonesia Language's Stemming Algorithm (Adriani M, et al., 2007).

4. N-Grams Representation

After passing the preprocessing stage, each word in the tweet will be represented as an attribute in the text representation [4]. The method used is N-gram. In this study, researchers used unigram (n = 1) in N-gram, because unigram is the most common basic technique since it will have the most complete attribute representation regardless of considering the meaning of a phrase or combination of two or more words.

For unigram, grouping is done on every single word. In each data, the number of occurrences of words will be attributes to the data.

5. Naïve Bayes

The naive bayes classifier algorithm is an algorithm used to find the highest probability value for classifying test data in the most appropriate category [2].

In this research that becomes test data is document tweets. There are two stages in the classification of documents. The first stage is the training of documents that are known to the category. While the second stage is the process of classification of documents that have not been known category. Bayes's Theorem has the following general formula:

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

Where:

1. $P(H|X)$ is the probable final probability (posteriory probability) of hypothesis H occurs when given evidence E occurs.
2. $P(X|H)$ is the probability that a proof E occurs will affect the hypothesis H.
3. $P(H)$ is the prior probability hypothesis H occurs regardless of any evidence.
4. $P(X)$ is the prior probability evidence of E regardless of the hypothesis or other evidence.

To use Bayes's theory, the two variables used are aspects / features as hypotheses (H) and sentiments as evidence (E). The other three variables will be used as metadata of the sentiment.

Because in a sentence consisting of many words, which is very difficult in practice to determine which one might be called an aspect / feature, it is assumed that each word is an aspect / feature. Then the application of Bayes theory:

$$P(K|F) = \frac{\mathrm{P(F|K)}X\,P(K)}{\mathrm{P(F)}}$$

Where :

1. F is a feature or a word.
2. K is a category or sentiment value.

Because the features or words that support one category can be many, eg there are features *F1, F2, F3,* the Bayes theory can be developed into

$$P(K|F_1, F_2, F_3) = \frac{\mathrm{P}(F_1, F_2, F_3|\mathrm{K})X\,P(K)}{\mathrm{P}(F_1, F_2, F_3)}$$

Since Bayes theory requires that the evidence (in this case is a feature or word) that exists is independent of each other, then the form of the above formula can be changed to:

$$P(K|F_1, F_2, F_3) = \frac{\mathrm{P}(F_1|\mathrm{K})X\mathrm{P}(F_2|\mathrm{K})X\mathrm{P}(F_3|\mathrm{K})X\,P(K)}{\mathrm{P}(F_1)X\mathrm{P}(F_2)\mathrm{X}\,\mathrm{P}(F_3)}$$

If described in general can be formulated as follows:

$$P(K|F) = \frac{\prod_{i=0}^{q} \mathrm{P}\,(F_1|\mathrm{K})}{\mathrm{P(F)}}$$

Because the value of *P*(*F*) is always fixed for a category, the calculation of the *P*(*F*) value can be done once, so that only the $\prod_{i=0}^{q} \mathrm{P}\,(F_1|\mathrm{K})$ is calculated only.

6. Accuracy Calculation

In this study, there are three categories as possible from classification result, that is positive category, negative, and neutral. Three lines and three columns of Confusion Matrix are then created:

Table 2. Confution Matrix

		Actual Value		
		Positive Category	Negative Category	Neutral Category
Predicted Value	Positive Category	False Negative	False Positif	False Positive
	Negative Category	False Negative	True Negative	False Negative
	Neutral Category	False Neutral	False Neutral	True Neutral

True Positive is the number of positive records that are classified as positive. False Positive is the number of positive records classified as negative and neutral. False Negative is the number of negative records classified as positive and neutral. True Negative is the number of negative records that are classified as negative. True Neutral is a neutral record number that is classified as neutral. False Neutral is the number of neutral records classified in negative and positive.

Assumed that researchers have 1000 tweets as test data, of which 800 tweets are predicted to be positive (positive predictions) and 100 tweets are predicted to fall into negative categories (negative predictive data). 100 tweets predicted to fall into neutral category.

Then from positive predictive data found that the correct tweet into the negative category is as much as 90 where 5 tweets enter the positive classification and 5 tweets into the neutral category. The right tweets included in the positive category are as many as 880 tweets where 5 tweets fall into the negative classification and 15 tweets fall into the neutral category. While the correct tweet into the neutral category is 90 tweets where 7 tweets entered in the positive classification and 3 tweets entered in the neutral category. Then the calculation of the accuracy is as follows:

Table 3. Predicted Value

		Actual Value			
		Positive Category	Negative Category	Neutral Category	Total
Predicted Value	Positive Category	780	5	15	800
	Negative Category	5	90	5	100
	Neutral Category	7	3	90	100
	Total	792	98	110	1000

$$Precision = \frac{\text{True Positive}}{\text{True positive + False Positive}}\ x\ 100\%$$

$$Precision = \frac{780}{800}\ x\ 100\% = 97\%$$

So the level of accuracy between the information requested by the user and the answer given by the system (precision) is 97%.

$$Recall = \frac{\text{number of positive predictive data that correct}}{\text{number of positive actual data}}$$

$$= \frac{\text{True positive}}{\text{True Positive + False Negative + False Neutral}}\ x\ 100\%$$

$$Recall = \frac{780}{792}\ x\ 100\% = 98\%$$

So the system success rate in rediscovering an information (recall) is 98%.

$$Accuracy = \frac{\text{number of correct prediction data}}{\text{total number of data}}$$

$$Accuracy = \frac{\text{True Positive + True Negative + True Neutral}}{\text{Total}}\ x\ 100\%$$

$$Accuracy = \frac{960}{1000}\ x\ 100\% = 96\%$$

So the level of closeness between the prediction value and the actual value of the system (accuracy) is 96%.

$$Recall = \frac{\text{jumlah data prediksi yang benar}}{\text{jumlah data positive yang sebenarnya}}$$

$$= \frac{\text{True positive}}{\text{True Positive} + \text{False Negative} + \text{False Neutral}}\ x\ 100\%$$

$$Recall = \frac{780}{792}\ x\ 100\% = 98\%$$

So the system success rate in rediscovering an information (recall) is 98%.

7. Conclusion

The use of Twitter social media that rife in Indonesia produces large textual data. Large data can be used as data that produces more useful information. With the use of sentiment analysis, useful information will be easier to obtain. So based on the studies and implementation that have been done, it can be concluded that this study produces a simulator that can be used to perform sentiment analysis by doing positive, negative, and neutral classification on input data.

References

[1] 2017. *GitHub.* [Online] Available at: https://github.com /sastrawi/sastrawi/wiki/Stemming-Bahasa-Indonesia [Accessed 27 March 2017].

[2] Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* United States of America: s.n.

[3] F. & Nur, D. D., 2015. Pengembangan Aplikasi Sentiment Analyst Mengunakan Metode Naive Bayes. *Seminar Nasional Sistem Informasi Indonesia,* pp. 336-340.

[4] Pantau, 2017. *Pantau.* [Online] Available at: pantaw.com/syntatic-proses-n-grams/ [Accessed Sunday May 2017].

[5] Sunni, I. & Widyantoro, D. H., 2012. Analisis Sentimen dan Ekstraksi Topik Penentu. *Jurnal Sarjana Institut Teknologi Bandung Bidang Teknik Elektro dan Informatika,* 1(2).