ISSN 2622-626X

Sentiments analysis for prediction the governor of east java 2018 in twitter

Ghulam Asrofi Buntoro

Universitas Muhammadiyah Ponorogo, Jl. Budi Utomo No. 10 Ponorogo 63471 Jawa Timur Indonesia
ghulamasrofibuntoro@gmail.com

ARTICLE INFO

Article history
Received 2018-11-12
Revised 2018-12-05
Accepted 2018-12-23

Keywords
Twitter
East Java
Sentiments Analysis

ABSTRACT

The East Java Governor Election which will be held in 2018 is also felt in the virtual world especially Twitter social media. All people freely argue about their respective governor candidates, the memorandum raises many opinions, not only positive or neutral but also negative opinions. Media growth is so rapid, revealing a lot of online media from the news media to social media. Social media alone is Facebook, Twitter, Path, Instagram, Google+, Tumblr, Linkedin and many more. Today's social media is not only used as a means of friendship or making friends, but also for other activities. Promos of trading or buying and selling, until political party promos or campaigns of candidates for regents, governors, legislative candidates until presidential candidates. The research objective is to conduct a method of analyzing the sentiments of 2018 East Java Governor candidates on Twitter social media with optimal and maximum optimization. While the benefits are to help the community conduct research on opinions on twitter which contains positive, neutral or negative sentiments. Analysis of the sentiments of East Java Governor candidates in 2018 on twitter social media using non-conventional processes that save costs, time and effort. The results of Khofifah's dataset are 77% accuracy, 79.2% precision value, 77% recall value, 98.6% TP rate and 22.2% TN rate. For the results of Gus dataset, the accuracy is 76%, the precision value is 74.4%, the recall value is 76%, the TP rate is 93.8% and the TN rate is 52.9%.

1. Introduction

The election of the Governor of East Java in 2018 was not only felt in the real world, in cyberspace, especially social media, Twitter people started talking about their prospective Governor candidates. The East Java Governor Election Stage in 2018 has been announced by the East Java General Election Commission (KPU) [1]. Since the registration stage until the appointment of 2018 East Java Governor candidates who will advance in East Java Election 2018, the names of candidates have already been widely discussed. The virtual world that is so free and difficult to control, makes everyone free to argue or opinion about their respective prospective Governor candidates, bring up many public opinions, not only positive or neutral opinions but also negative ones.

The development of the information world is so fast, bringing a lot of online media, from news information to social media or friendship, social media starting from Facebook, Twitter, Path, Instagram, Google+ and many more. In 2015 Indonesia became the number two active social media Twitter user from the total number of active Twitter users worldwide until now 330 million, the number of Tweet sent per day for the whole world around 500 million and the number of active daily users around the world around 100 million [2].

The excitement of the 2018 East Java Pilkada has been felt in social media, especially Twitter, social media, especially Twitter, which is now a very important place for candidates and successful teams to conduct campaigns. The success team of a candidate for governor or regional head now, for example, until they justify any means in campaigning for their candidates, as evidenced in every campaign period of many Black Campaigns, especially in social media against a candidate. Today the campaign or imaging is not only done in the real world but also penetrates the virtual world. Social media especially Twitter is now one of the effective and efficient campaigns.

Sentiment analysis is still part of opinion mining research, namely the process of understanding, extracting and processing textual data automatically to get information on sentiments contained in an opinion sentence [3].

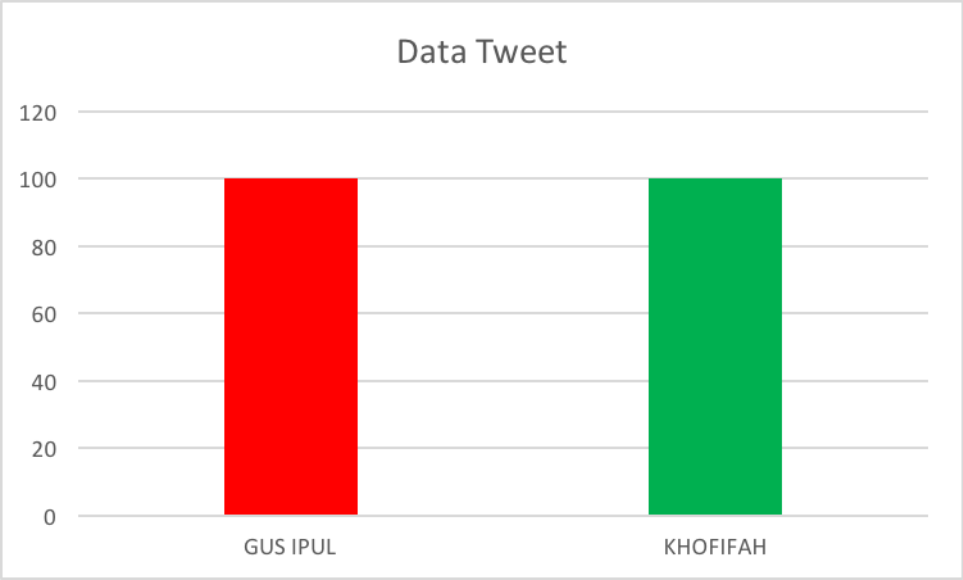
In this study, sentiment analysis was carried out to see and retrieve information from a person's opinion in Indonesian on Twitter aimed at 2018 East Java Governor candidates, whether that opinion was categorized as positive, neutral or negative opinion. The method of weighting uses Lexicon Based Features and to test the accuracy of sentiment analysis in this study using two methods, namely the Naïve Bayes Classifier (NBC) method, because the method is widely used for sentiment analysis with fairly good accuracy results. [4]

2. The Proposed Method/Algorithm

The research steps in accordance with the research flow are as follows:

2.1. Collect tweet data

Tweet data taken by the Crawling method from Twitter social media. The data taken only tweets in Indonesian starting June 17 2016-23 June 2018. The amount of data is balanced, ie 100 tweets with keywords Gus Ipul and 100 tweets with the keyword Khofifah. Data is taken randomly from ordinary users or online media on Twitter.



Gambar 1. Data *tweet*

2.2. Preprocessing Data

At the preprocessing stage, 4 steps are carried out as follows.

1. Selection of comments

At this stage, the selection of comments containing the hashtag Gus Ipul and Khofifah (#GusIpul and #Khofifah) was carried out, because Twitter has a retweet function, which provides comments on someone's tweet comments, tweet comments will interfere with the tweet sentiment analysis process. So in preprocessing this tweet comment is deleted.

2. Cleansing

Sentences that are obtained usually still have noise, namely random errors or variants in measured variables [11], for that, we must eliminate the noise. The words omitted are HTML characters, keywords, emotion icons, hashtag (#), username (@username), url (http://website.com), and email (nama@website.com) [10].

3. Parsing

namely the process of dividing documents into a word by analyzing a collection of words by separating the word and determining the syntactic structure of each word. [11]

4. Sentence Normalization

Aiming to normalize sentences so that slang sentences become normal [10], so that slang language can be identified as a language that matches the KBBI.

What must be done to normalize sentences is:

• Stretch punctuation and symbols other than the alphabet

Stretching punctuation marks is a distance from punctuation from the words after or before, the goal is that punctuation and symbols other than the alphabet do not become one with words during the tokenization process.

• Change to all lowercase letters

• Word normalization

The rules in the normalization process can be seen in Table 1.

Table 1. Rules for word normalization [12]

Not normal / slang	Not normal / slang
Normal	Normal
Suffix -ny	Suffix -ny
The suffix	The suffix
Suffix -nk	Suffix -nk
Endings -ng	Endings -ng
Suffix -x	Suffix -x
The suffix	The suffix
Suffix -z	Suffix -z

• Eliminates repetitive letters

When you're happy or upset, someone is free to write opinions based on their emotions, usually someone writes by repeating the same letter. For example: "kereeen" to express pleasure. Repeated words like "kereeen" will be normalized to "cool".

• Eliminate emoticons

When writing someone's status (tweet) sometimes wrong or incorrect in the use of emoticons, whether intentional or not many do it. For example: They can only slander because they cannot meet bad facts :), said slanderous opinions but the emoticon smiles :), so the emoticon will interfere in the tweet Sentiment Analysis process, so in this process emoticons are deleted or ignored. Some emoticons, feeling and sentiment can be seen in Figure 2.

Emoticon	Feeling	Sentiment
:) :-)	Happy	Positive
:(:-(	Sad	Negative
:D :-D	Very Happy!	Positive
D: D=	Very Sad	Negative
_ *.* *-*	Fascinated	Positive
D:< D: D8	Horror, disgust, sadness	Negative
xD XD	Laughing, big grin	Positive
:| =| :-|	Straight face no expression	Neutral

Figure 2. Emoticon, Feeling and Sentiment

3. Tokenisasi
After normalization of sentences, then the sentence is broken into tokens using delimiter or delimiter space. The token used in this study is [11]:
• unigram: tokens which consist of only one word, for example: Election.
• bigram: tokens consisting of two words, for example: General Election.
• trigram: tokens consisting of three words, for example: Regional General Elections.
4. Term Frequency-Inverse Document Frequency (tf-idf)
This weighting is the most popular weighting method and is often used by researchers, this weighting is the result of multiplying weighting term frequency and inverse document frequency of a term.
The equation:
$w = w \times idf\ (3)\ t,d\ tf_{t,d}\ t$ (1)

Information :
$w_{tf_{t,d}}$: *Term Frequency.*
idf : *Inverse Document Frequency.*

5. Determination of Class Attribute
Twitter data that has been done Preprocessing will then be determined class attribute, class attributes that appear in this study there are 3, including positive, neutral, and negative. With these 3 class attributes, it is expected to be able to accurately assess a particular object.

6. Load Dictionary
After tokenisation and class attribute are specified, the next step is dictonary load. Many types of dictionaries can be used, for example: positive sentiment keyword dictionary (positive keywords), negative sentiment keyword dictionary (negative keywords), negation keywords dictionary, and slang or alay normalization dictionary.
Here is an example of a dictionary and its contents [10]:
• Positive keywords: good, great, honest, smart, cool.
• Negative keywords: lying, corruption, evil, bad.
• Negation keywords: no, no, no, far away.
• Dictionary of slang conversion to KBBA: sp = who, like = like, brp = what, hrg = price, ciyus = serious.

7. Word Weighting
After knowing the word that contains positive, negative and neutral in a sentence, then the weight of the value contained in the sentence is calculated by adding the value of the word opinion. If the number of opinion values in the sentence is 1, then the sentiment value of the sentence is positive, if the opinion value in that sentence = 0, then the sentiment value of the sentence is neutral, if the opinion value in that sentence = -1, then the sentiment value from the sentence is negative.

Table 2. Sentiment Value

Sentiment	Value
Positive	1
Neutral	0
Negative	-1

8. Classification
Enter the classification process. Classification process using WEKA 3.7.11. The classification method used in this study is the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM). The data classification process was tested using the 10-fold cross validation method [15]. So the dataset will be divided into two, namely 10 parts with 9/10 parts used for the training process and 1/10 parts used for the testing process. Iteration lasts 10 times with variations in training and testing data using a combination of 10 parts of data.

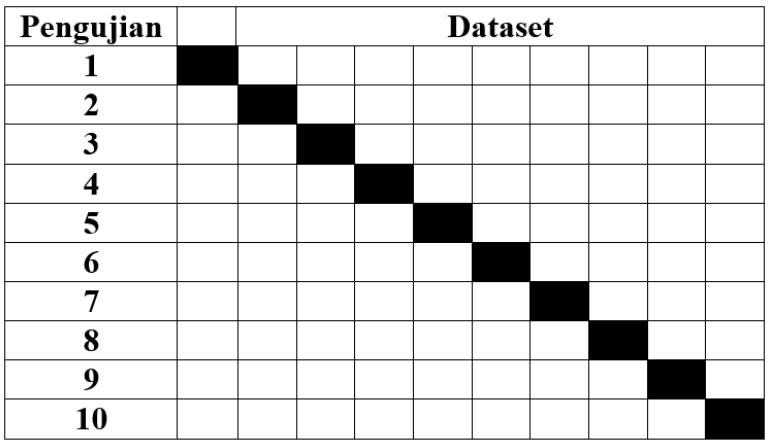


Figure 3. 10-fold cross validation illustration

9. Evaluation of Results

Evaluate the performance of Accuracy, Precision and Recall from experiments that have been conducted. 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 classes that are classified as positive class, while TN rate is the percentage of negative classes that are classified as negative classes. FP rate is a negative class which is classified as a positive class. FN rate is a positive class which is classified as a negative class [16].

Table 3. *Confusion Matrix*

		Predicted	
		Negative	*Positive*
Actual	*Negative*	*a*	*b*
	Positif	*c*	*d*

3. Results and Discussion

The results in the form of research data that has been processed and poured in the form of tables, graphs, photos or images. The discussion contains the results of the analysis and research results that are associated with established knowledge structures (literature reviews referred to by the author), and bring new theories or modifications to existing theories.

The dataset in this study uses the ARFF format collected from Twitter with the Crawling method from Twitter social media. The data taken are only tweets in Indonesian, namely tweets with the keyword Guspul for Candidates for East Java Governor 2018 Syaifullah Yusuf and Khofifah for Candidates for East Java Governor 2018 Khofifah Indar Parawansa. Data is taken randomly from ordinary users or online media on Twitter.

The dataset used is 200 Tweets, the data is balanced equally for each class, because with unbalanced data, the classifications built have a tendency to ignore minority classes (Kohavi, 1998). Data is divided into GusIpul 100 Tweets, and Khofifah 100 Tweets. Labeling is carried out using the Lexicon Based Features method and the assistance of Indonesian language experts.

The results of the 2018 East Java Governor Candidate Sentiment Analysis using the Lexicon Based Features method with three class attributes namely positive, neutral and negative.

Table 4. Analysis Results of the Lexicon Based Features method

Sentiment	**GusIpul**	**Khofifah**
Positive	65	72
Neutral	18	19

Negative	17	9

To find out the accuracy, Analysis of Sentiment of candidates for Governor of DKI Jakarta 2017 with the Lexicon Based Features method is classified using the Naïve Bayes Classifier (NBC) method with WEKA software version 3.8.1. WEKA uses the Document Attribute File Format (ARFF) type as input to classify data.
The results of the classification process are then tested using the 10-fold cross validation method, the data is divided into 10 parts with 9/10 parts used for the training process and 1/10 parts are used for the testing process. Iteration lasts 10 times with variations in training and testing data using a combination of 10 parts of data.
Comparison of results from the Naïve Bayes Classifier (NBC) classification method with datasets of Candidates for East Java Governor 2018 GusIpul and Khofifah.

Table 5. Comparison of Classification Results

Governor candidate	Accuracy (%)	Precision (%)	Recall (%)	TP Rate (%)	TN (%)	Rate
GusIpul	76	74,4	76	93,8	52,9	
Khofifah	77	79,2	77	98,6	22,2	

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

Table 5. contains information on the value of accuracy, precision, recall, TP rate and TN rate of each trial that has been carried out. The column section contains information about the 2018 East Java Governor Candidate. While the row section contains the value of accuracy, precision, recall, TP rate and TN rate of each trial that has been done. From the preprocessing data process produces a number of tokens which are then used as input to a classification process. The classification process is carried out using the Naïve Bayes Classifier (NBC) method. From the classification process obtained the value of accuracy, precision, recall, TP rate and TN rate of each trial.

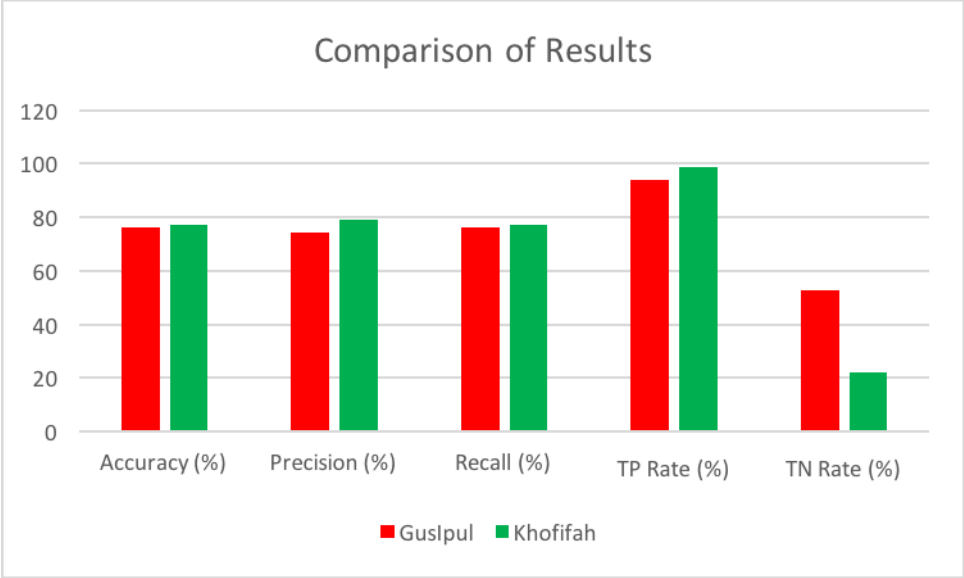


Figure 4. Graph of accuracy

From Figure 4. can be seen the accuracy of the value of Sentiment Analysis with the Lexicon Based Features method classified by the Naïve Bayes Classifier (NBC) method. The accuracy value of Khofifah dataset reaches 77%, the precision value is 79.2%, the recall value is 77%, the TP value is 98.6% and the TN rate is 22.2%. For the GusIpul dataset, the accuracy reaches 76%, 74.4% precision value, 76% recall value, TP rate 93.8% and TN rate value 52.9%. The Khofifah Dataset gets the highest accuracy because of 72 positive data, 71 data were successfully classified as the Naïve Bayes Classifier (NBC) method correctly according to the sentiment of positive

sentiment. While for the GusPul dataset, out of 65 positive data, 61 data successfully classified the Naïve Bayes Classifier (NBC) method correctly according to the sentiment of positive sentiment. This method tends to be Naïve Bayes Classifier (NBC) more stable because it is based on the probability of the occurrence of words in a sentence. Accuracy value is one of the assessment parameters of the method that has been used, the accuracy value is obtained from the amount of data that is successfully classified correctly according to the sentiment class of the total amount of data classified. High accuracy values are obtained when a lot of data is successfully classified correctly according to the sentiment class.

From Figure 4. you can also know the value of Precision and Recall. Nilai Precision follows the accuracy value, the higher the accuracy value, the high Precision value will be followed too, and vice versa. Nilai Precision is the correct amount of positive data classified as positive data divided by the total data classified as positive data. While the recall value is the number of positive data that is correctly classified as positive data divided by the number of actual positive data.

From Figure 4. we can also find out the value of TP Rate and TN Rate. TP Rate is a positive data value that is correctly classified according to the sentiment class, which is positive. The TN Rate value is the value of sentiment data that is correctly classified according to the sentiment class, which is negative.

From the research that has been done, it is proven that the Naïve Bayes Classifier (NBC) classification method can be used to classify Indonesian positive tweets (positive, neutral and negative) towards 2018 East Java Governor Candidates. Furthermore, the Khofifah dataset gets higher accuracy compared to accuracy GusIpul dataset, with 77% accuracy compared to 76%. In the Khofifah dataset, the most positive sentiment was 72 and the negative sentiment was only 9. While the positive GusI was 65 and negative sentiment 17. So it can be concluded, on Twitter social media Khofifah is more loved than GusIpul. Although it produces high accuracy, the model that is built still does a little misclassification for datasets that share the unbalanced sentiment. Because using unbalanced datasets will cause minority class data to be misclassified as majority class data (Kohavi, 1998). In the end, the difference in value becomes large.

4. Conclusion

From the research that has been done, it is proven that the Naïve Bayes Classifier (NBC) classification method can be used to classify Indonesian positive tweets (positive, neutral and negative) towards 2018 East Java Governor Candidates. Furthermore, the Khofifah dataset gets higher accuracy compared to accuracy GusIpul dataset, Khofifah's dataset results in 77% accuracy, 79.2% precision value, 77% recall value, TP rate 98.6% and TN rate 22.2%. For the results of the GusIpul dataset, the accuracy is 76%, the precision value is 74.4%, the recall value is 76%, the TP rate is 93.8% and the TN rate is 52.9%. In the Khofifah dataset, the most positive sentiment was 72 and the negative sentiment was only 9. While the positive GusI was 65 and negative sentiment 17. So it can be concluded, on Twitter social media Khofifah is more loved than GusIpul. Proven Sentiment Analysis can be used to find out Twitter's public netizens, especially towards 2018 East Java Governor Candidates, thus helping ordinary people to find out other people's sentiments towards 2018 East Java Governor Candidates. For further research, try to develop more data using Real Time. . It is also necessary to develop an Indonesian stopword list and stemmer that can improve accuracy in the analysis of Indonesian Sentiment.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] “Kegiatan tahapan pilgub jatim 2018 makin padat, ketua kpu jatim ajak jaga soliditas,” *kpu provinsi jawa timur*, 08-feb-2018.

[2] S. Aslam, “ Twitter by the Numbers (2018): Stats, Demographics & Fun Facts,” 01-Jan-2018.

[3] B. Liu, “Sentiment Analysis and Subjectivity.,” *Handb. Nat. Lang. Process.*, vol. 2, pp. 627–666, 2010.

[4] B. Wagh, S. J. V., and W. N. R., "Sentimental Analysis on Twitter Data u sing Naive Bayes," *IJARCCE*, vol. 5, no. 12, pp. 316–319, Dec. 2016.

[5] RENSTRA PENELITIAN_Universitas_Muhammadiyah_Ponorogo.pdf.".

[6] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment Analysis of Turkish Political News," 2012, pp. 174–180.

[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREc*, 2010, vol. 10.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79–86.

[9] A. F. Hadi and M. Hasan, "text mining pada media sosial twitter studi kasus: masa tenang pilkada dki 2017 putaran 2." Seminar Nasional Matematika dan Aplikasinya, Universitas Airlangga 2017.

[10] J. Ariawan, "Data Preprocessing." [Online]. Available: https://www.google.com/search?q=apa+itu+noise+dalam+data+mining&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a&channel=sb. [Accessed: 10-Mar-2014].

[11] M. Yusuf nur sumarno putro, "analisis sentimen pada dokumen berbahasa indonesia dengan pendekatan support vector machine," masters, BINUS, 2011.

[12] N. Adiyasa, "Analisis Sentimen Pada Opini Berbahasa Indonesia Menggunakan Pendekatan Lexicon-Based," *Catatan Kecil*, 2011. [Online]. Available: http://adiyasan.wordpress.com/2013/02/08/sentiment-analysis-menggunakan-pendekatan-lexicon-based/. [Accessed: 10-Mar-2014].

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

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

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

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