Sentiments Analysis for Governor of East Java 2018 in Twitter

by Ghulam Buntoro

Submission date: 07-Oct-2019 04:06PM (UTC+0700)
Submission ID: 1187685340
File name: document_1.pdf (183.48K)
Word count: 4172
Character count: 22286

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Sentiments Analysis for Governor of East Java 2018 in Twitter

Ghulam Asrofi Buntoro
Teknik Informatika, Fakultas Teknik
Universitas Muhammadiyah Ponorogo
ghulamasrofibuntoro@gmail.com

Abstract— East Java Governor Election 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 with Data Prepocessing Selection comments, Cleansing, Parsing, Normalizing, Tokeninzing, Load Dictionary, Weighting and Classification 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%.

Keywords—sentiment analysis, governor, lexicon based features, naïve bayes classifier

I. 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,

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

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]

II. Literature Review

2.1. Literature review

Research by Mesut et al. [6] used machine learning to classify Turkish political news. This research classifies sentiment towards Turkish political news and determines whether the Turkish political news has a positive or negative sentiment. The different features of Turkish political news are extracted with the machine learning algorithm Naïve Bayes Classifier (NBC), Maximum Entropy (ME) and Support Vector Machine (SVM) to produce a classification model. This study obtained 72.05% accuracy for Naïve Bayes Classifier (NBC), Accuracy of 69.44% Maximum Entropy and 66.81% for SVM on the use of bigram.

Pak et al. [7] used emoticons to build an English-language corpus from Twitter with positive, negative and neutral sentiments. For neutral class, Pak and Paurobek took training data from tweets in English media accounts. The method used is Naïve Bayes with n-gram. The best performance is generated when using bigram.

Research by Pang et al [8] uses machine learning to classify moview reviews. This research classifies sentiments towards film reviews and determines whether the film review has a positive or negative sentiment. The different features of the review were extracted and used the Naïve Bayes machine learning algorithm and Support Vector Machine (SVM) to produce a classification model. They obtained an accuracy of 78.7% when using Naïve Bayes on unigram use. The accuracy obtained when using SVM with unigram is 72.8%.

Sentiment analysis is still part of opinion mining [9] referring to a broad field of natural language processing, computational linguistics and text mining. In general, it aims to determine the attitude of the speaker or writer regarding a particular topic. Attitude might be their evaluation or evaluation, their affective statement (the emotional statement of the writer when writing) or the intended emotional communication (the emotional effect the writer wants on the reader).

The basic task in sentiment analysis is to classify the polarity of the text in the document, sentence or feature / level aspect - whether the opinions expressed in the document, sentence or feature of the entity / aspect are positive, negative or neutral. Furthermore sentiment analysis can express emotional sadness, joy, or anger.

Lexicon Based Features is a word feature that has a positive or negative sentiment based on a dictionary or lexicon. Lexicon is a collection of known and collected sentiment words (Desai & Mehta, 2016). For the weighting process of this feature, a dictionary or lexicon that contains words that contain sentiments is called sentiment dictionaries (Buntoro, Adji, & Permanasari, 2014, Cho, et al., 2014). The sentiments used are positive and negative. There are two kinds of features used in this study, namely the feature of the number of positive words and features of the number of negative words in the document. Lexicon Based Features weight is balanced with the weight of tf-idf, so the feature of the number of positive words and negative words needs to be normalized by the Min-max method.

2.2.3. Naïve Bayesian Classifier (NBC)

Classification is a supervised learning process. To do classification, a training set is needed as learning data. Each sample from the training set has attributes and label classes.

The two classification stages are as follows:

a. Learning (training): Learning uses training data (for Naïve Bayesian Classifiers, probability values are calculated in the learning process)

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

b. Testing: Test the model using testing data

As a basis for Bayesian theory X is an unknown data sample class H is a hypothesis that X is data with class (label) C. P (H) is the opportunity of the hypothesis H. P (X) is the opportunity of observed sample data P (X | H) is the chance for sample X data, if it is assumed that the hypothesis is valid (valid). For classification problems, what is calculated is P (H | X), which is the opportunity that the hypothesis is correct (valid) for the X sample data observed:

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

Naïve Bayesian Classifier is a classifier method based on probability and the Bayesian Theorem with the assumption that each variable X is independent (4). In other words, Naïve Bayesian Classifier (NBC) assumes that the existence of an attribute (variable) has nothing to do with the existence of another attribute (variable).

III. RESEARCH METHODS

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

1. Collect tweet data

The tweet data is taken by the Crawling method from Twitter social media. The data taken is only tweets in Indonesian starting on June 17 2016 - June 23, 2018. The amount of data is balanced, which is 100 tweets with Gus Ipul keywords and 100 tweets with the keyword Khofifah. For more variety the data is taken randomly from either ordinary users or online media on Twitter [5].

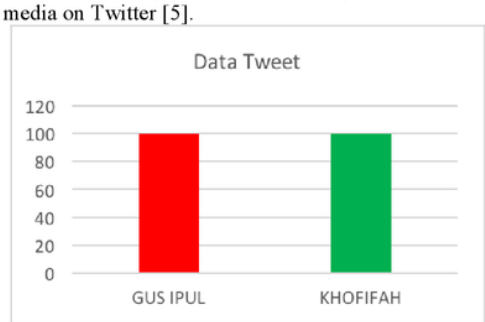

Figure 1. Tweet data

2. Preprocessing Data

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

2.2.1. Selection comments

At this stage, a comment selection containing the hashtag Gus Ipul and Khofifah (#GusIpul and #Khofifah) was conducted, on Twitter there was a retweet function, useful for commenting on tweets of someone's opinions. Comment tweets will interfere in the Sentiment Analysis process, so in this preprocessing comment the tweet is deleted.

2.2.2. Cleansing

The sentence obtained is usually still there is noise [6], for that, we have to eliminate the noise. Words that are omitted are HTML characters, keywords, emotion icons, hashtag (#), username (@username), url (http://www.alamatwebsite.com), and email (name@alamatwebsite.com).

2.2.3. Parsing

Parsing process is the process of breaking a document into a word by analyzing a collection of words by separating the word and determining the syntactic structure of each word [7].

2.2.4. Sentence Normalization

The aim is to normalize opinion sentences so that sentences with slang are normal or standard [6]. Besides that, the slang can be recognized as a language that is in accordance with the KBBI.

What must be done for the normalization of sentences is:

- Stretching punctuation and symbols other than the alphabet is to give a distance to the punctuation of the words after or before, the purpose is that punctuation and symbols other than the alphabet do not become one with the words during the tokenisation process.
- Change to all lowercase letters
- Normalization of words

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

Table 1. Rules for normalizing words [8].

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Abormal	Normal
Suffix -ny	Suffix –nya
Suffix –nk	Suffix –ng
Suffix –x	Suffix –nya
Suffix –z	Suffix –s
Suffix –dh	Suffix –t
Repeated Words: sama2	Repeated Words : sama-sama
Spelling: oe	Alphabet: u
Spelling: dj	Alphabet: j

• Eliminating repeated letters, for example: "hebaaat" to express admiration. Repeated words like "hebaaat" will be normalized to be "great".

3. Tokenisasi

After normalizing the sentence, then the sentence is broken down into tokens using a delimiter or space delimiter. The token used in this study is [9]: trigram: meaning the sentence is broken down into tokens with tokens consisting of three words, for example: the General Election Commission.

4. Load Dictionary

After tokenisation, the next step is to load the dictionary for. Many types of dictionaries can be used, for example: dictionaries of positive sentiment keywords (positive keywords), dictionary of negative sentiment keywords (negative keywords), dictionary of negative words (negation keywords), and normalization dictionaries for slang or alay.

The following is a sample dictionary and its contents [8]:

• Positive keywords: honest, great, smart, good, cool, smart.

• Negative keywords: cheats, lying, corruption, evil, bad.

• Negation keywords: same, no, no, no, far away.

• Dictionary of slang conversion to KBBA: sp = who, like = like, brp = how many

5. Word Weighting

After knowing the words that contain positive, negative and neutral in a sentence, then the weight of the values contained in the sentence is calculated by adding up 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 of the sentence is negative [9].

Table 2. Sentiment Value

Sentiment	Value
Positive	1
Neutral	0
Negative	-1

6. Classification

Enter the classification process. The classification process uses WEKA 3.7.11. The classification method used in this study is Naïve Bayes Classifier (NBC). Naïve Bayesian Classifier is a method classifier based on probability and the Bayesian Theorem with the assumption that each variable X is independent. In other words, Naïve Bayesian Classifier (NBC) assumes that the existence of an attribute (variable) has nothing to do with the existence of another attribute (variable). Here are the formulas.

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

In the classification process the data is tested using the 10-fold cross validation method [10]. So the dataset will be divided into two, namely 10 parts with 9/10 parts used for the training process and 1/10 parts are used for the testing process. Iteration takes place 10 times with variations in training and testing data using a combination of 10 parts of data.

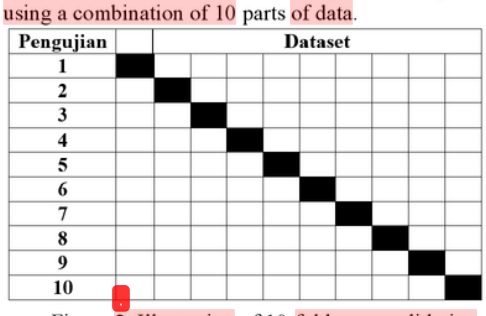

Figure 2. Illustration of 10-fold cross validation

7. Evaluate Results

Evaluating Accuracy, Precision and Recall performance from experiments that have been carried out. 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 the positive class that has been successfully classified as a positive class, while the TN rate is the percentage of the

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

negative class that has been successfully classified as a negative class. FP rate is a negative class that is classified as a positive class. FN rate is a positive class that is classified as a negative class [11].

Table 3. Confusion Matrix

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

IV. Results and Discussion

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

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

The dataset is used as many as 200 Tweets, the data is divided in a balanced manner each class, because with data that is not balanced (imbalanced), the classification that is built has a tendency to ignore minority class [11]. The data is divided into GusIpul 100 Tweets, and Khofifah 100 Tweets. Labeling is done using the Lexicon Based Features method and the assistance of Indonesian language experts.

The results of the Sentiment Analysis of the 2018 East Java Governor Candidates used the Lexicon Based Features method with three attribute classes namely positive, neutral and negative.

Table 4. Results of Lexicon Based Features Sentiment Analysis

Sentiment	**GusIpul**	**Khofifah**
Positive	65	72
Neutral	18	19
Negative	17	9

To find out the accuracy, the Sentiment Analysis of 2017 DKI Jakarta Governor candidates with the Lexicon Based Features method is classified using the Naïve Bayes Classifier (NBC) method with WEKA version 3.8.1 software. WEKA uses the Attribute-Relation File Format (ARFF) document 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 takes place 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 a dataset of 2018 East Java Governor Candidates Guspul and Khofifah.

Table 5. Comparison of Classification Results

Candidate for Governor	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

*) Precision and Recall values are the average values of positive class values and negative classes.

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

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

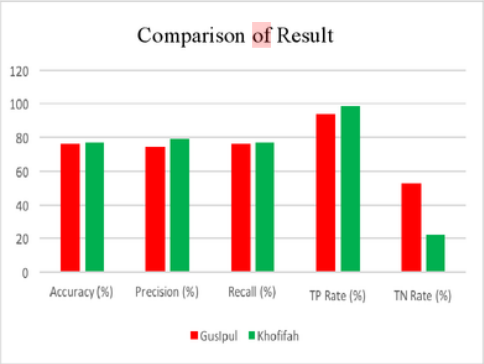

Figure 4. Graph of accuracy level

From Figure 4. can be seen the accuracy of the value of Sentiment Analysis with Lexicon Based Features method which is classified by the Naïve Bayes Classifier (NBC) method. The accuracy of the Khofifah dataset reaches 77%, the value of precision is 79.2%, the recall value is 77%, the TP rate is 98.6% and the TN value is 22.2%. For the GusIpul Accuracy dataset, it reaches 76%, the precision value is 74.4%, the recall value is 76%, the TP rate is 93.8% and the TN value is 52.9%. The Khofifah dataset obtained the highest accuracy because of 72 positive data, 71 data were successfully classified by the Naïve Bayes Classifier (NBC) method correctly according to the sentiment of positive sentiment. As for the GusIpul dataset, of the 65 positive data, 61 data were successfully classified by the Naïve Bayes Classifier (NBC) method correctly according to the sentiment of positive sentiment. This method tends to be more stable because Naïve Bayes Classifier (NBC) is based on the probability of the appearance of words in a sentence. Accuracy value is one of the parameters of the evaluation method that has been used, the accuracy value is obtained from the amount of data that is successfully classified correctly according to the class of sentiments of the entire amount of data classified. High accuracy values are obtained when many data that are successfully classified correctly according to the sentiment class.

From Figure 4. you can also see the value of Precision and Recall. Precision values follow the value of accuracy, the value of accuracy gets higher then it will follow a high Precision value, and vice versa. Precision value is the amount of positive data that is correctly classified as positive data divided by the total data classified as positive data. Whereas 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 the value of positive data that is correctly classified according to the sentiment class, which is positive. TN Rate 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 sentiments (positive, neutral and negative) tweet Indonesian Language towards the Candidates for East Java Governor 2018. Furthermore, the Khofifah dataset gets higher accuracy than accuracy GusIpul dataset, with an accuracy of 77% compared to 76%. In the Khofifah dataset, the positive sentiments were mostly 72 and the negative sentiments were only 9. While the GusIpul dataset was positive and negative sentiments were 17. So it can be concluded, on Khofifah's Twitter social media more loved than GusIpul. Even though it produced a fairly high accuracy, the model built still made a bit of a misclassification for the dataset whose distribution of sentiments was not balanced. Because using an unbalanced dataset will cause minority class data that is incorrectly classified as majority class data (Kohavi, 1998). In the end it makes the value difference big.

IV. CONCLUSIONS AND SUGGESTIONS

From the research that has been done, it is proven that the Naïve Bayes Classifier (NBC) classification method can be used to classify sentiments (positive, neutral and negative) tweet Indonesian Language towards the Candidates for East Java Governor 2018. Furthermore, the Khofifah dataset gets higher accuracy than accuracy the GusIpul dataset, the results of the Khofifah dataset are accuracy of 77%%, the value of precision is 79.2%, the recall value is 77%, the TP rate is 98.6% and the TN value is 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 value is 52.9%. In the Khofifah dataset, the positive sentiments were mostly 72 and the negative sentiments were only 9. While the GusIpul dataset was positive and negative sentiments were 17. So it can be concluded, on Khofifah's Twitter social media more loved than GusIpul. Proven Sentiment Analysis can be used to find out the public sentiment, especially Twitter netizens, against the East Java Governor Candidate 2018, thus helping ordinary

Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10025
e-ISSN : 2541-2019
p-ISSN : 2541-044X

people to know the sentiments of other people towards 2018 East Java Governor Candidates. Further research should be developed using more data and Real Time. . It is also necessary to develop an Indonesian language stopword list that is able to improve accuracy in Indonesian Sentiment analysis.

REFERENCE

[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 using 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] G. A. Buntoro, (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, *2*(1).

[10] 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.

[11] G. A. Buntoro, (2016). " Sentiment Analysis Candidates of Indonesian Presiden 2014 with Five Class Attribute" in International Journal of Computer Applications (0975 – 8887).

[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] Class StringToWordVector. 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) *Confusion Matrix http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html*

Sentiments Analysis for Governor of East Java 2018 in Twitter

ORIGINALITY REPORT

24%	22%	3%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

16%

★ www.ijcaonline.org

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		