© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
https://doi.org/10.32628/IJSRSET21841124

Study of Sentiment of Governor's Election Opinion in 2018

Agung Eddy Suryo Saputro*, Khairil Anwar Notodiputro, Indahwati
Department of Statistic, Bogor Agricultural University, Bogor, Indonesia

ABSTRACT

In 2018, Indonesia implemented a Governor's Election which included 17 provinces. For several months before the Election, news and opinions regarding the Governor's Election were often trending topics on Twitter. This study aims to describe the results of sentiment mining and determine the best method for predicting sentiment classes. Sentiment mining is based on Lexicon. While the methods used for sentiment analysis are Naive Bayes and C5.0. The results showed that the percentage of positive sentiment in 17 provinces was greater than the negative and neutral sentiments. In addition, method C5.0 produces a better prediction than Naive Bayes.

Keywords: C5.0, Governor Election opinion, Naive Bayes, sentiment mining, Twitter.

I. INTRODUCTION

The Governor's election in Indonesia, was held on June 27, 2018 which included 17 provinces. News about the Governor's election, often raised as headlines in mass media and social media, especially Twitter. Twitter as one of the social media in the form of microblogging is used to convey and disseminate public opinion regarding the Election of Governors in 2018. This community opinion is often the topic of Twitter trends in Indonesia for several months before Governor Election Day. This is consistent with the results of several studies [1–3] which show that the most popular Twitter topic is political opinion.

Twitter's rapid growth since 2006, is one of the effects of the development of internet technology. However, the development of internet technology has not yet reached all regions of Indonesia. [4]The Indonesian Internet Service Providers Association (Asosiasi Penyelenggara Jasa Internet Indonesia, abbreviated APJII) shows data on the percentage of internet users based on geographical regions in 2017, as follows: 58.08% of users in Java, 19.09% of Sumatra, 7.97% of Kalimantan, 6.73% of Sulawesi, 5.63% of Bali Nusa Tenggara, and 2.49% Maluku Papua. Based on this description, the focus of this research is to analyze the opinion of the Governor's Election by provinces in Indonesia.

This study uses mining sentiment to explore the sentiment of tweets about political opinion in 17 provinces that carried out Governor Election in 2018. The results of sentiment mining are analyzed further using correlation analysis and sentiment analysis. The correlation analysis used in this study is Pearson correlation, while for sentiment analysis using the Naive Bayes and C5.0 methods. The performance of the Naive Bayes and C5.0 methods is evaluated based on the values of accuracy, precision, and recall.

The objectives of this study are (1) to describe the results of sentiment mining in each province, (2) to compare the results of predictions produced by Naive Bayes and C5.0, and (3) to analyze the relationship between sentiments in each province.

II. METHODS AND MATERIAL

2.1 Data Sources and Text Mining

Textual data about opinions regarding the election of Governors comes from the Twitter *crawling* process. The *crawling* process on Twitter requires *keywords* that are used as search words. The *keywords* used are the results of *Topic Modeling* using the *Latent Dirichlet Allocation* (LDA) method [5]. The *crawling* process is carried out for a month before and a month after the Governor's Election, namely the period May 27, 2018 - July 27, 2018. In addition, the *crawling* process was carried out in 17 provinces that carried out the Governor's Election (Table 1). The number of tweets generated in the *crawling* process was 21372 tweets. Data generated from the *crawling* process is then *pre-processing*.

Table 1. Provinces that carry out Governor Election in 2018

Code	Province	Code	Province
1	North Sumatra	10	East Nusa Tenggara (NTT)
2	Riau	11	West Kalimantan
3	South Sumatra	12	East Kalimantan
4	Lampung	13	South Sulawesi
5	West Java	14	Southeast Sulawesi
6	Central Java	15	Maluku
7	East Java	16	North Maluku
8	Bali	17	Papua
9	West Nusa Tenggara (NTB)		

Source: Indonesian General Election Commission

Pre-processing is the stage that most influences the quality of the data produced. In the *pre-processing* stage, the processing of unstructured textual data is transformed into structured data. The pre-processing includes several steps, as follows:

1. *Tokenizing, parsing, filtering, cleaning,* and *case folding* that aim to eliminate punctuation, user names, @, url, numbers, noises, convert tweets to lowercase letters, and make tweets into a collection of words.
2. Normalization of words to improve non-standard words to a standard word based on *Big Indonesian Language Dictionary* (KBBI).
3. *Stopword* removal and *stemming*. *Stopword* is a non-essential word with zero values. Whereas stemming is the process of changing word forms into basic form in accordance with KBBI.

After *pre-processing* is complete, the next step is *sentiment mining*. *Sentiment mining* is a process of extracting sentiments from a data so that a sentiment can be generated [5]. *Sentiment mining* in this study was conducted using the *Lexicon* approach [6]. [5] There are several stages in the *Lexicon* approach, as follows:

1. Mark all words containing sentiment, if the positive score is +1 and negative score is -1.
2. Apply shifter sentiment, which is the process of changing words if you meet the negation. For example, the word "smile" which contains a +1 or positive score, if there is a word "no" in front of the word "smile", then the word "no smile" has a score of -1.
3. Handle clauses "but", for example "speech is not long [+1], but fun [+1]".
4. Calculating the sentiment score, if the number of scores > 0 enters a positive sentiment, if the number of scores < 0 enters negative sentiment, and if the number of scores 0 enters neutral sentiment.

All processes carried out from crawling to sentiment mining are part of the text mining.

The software used in this study is R version 3.5.1. The R version 3.5.1 software packages used are twitteR [7], ROAuth [8], wordcould [9], topicmodels [10], tm [11], stringr [12], Hmisc [13], corrplot [14], and e1071 [15].

2.2 Sentiment Analysis of The Naive Bayes and C5.0 Methods

Sentiment analysis can be defined as a textual data extraction process that results from processing textual data so that data can be classified into 3 classes, namely positive, negative, and neutral [16]. The method used for sentiment analysis in this study is the Naive Bayes and C5.0 methods. Several advantages of the Naive Bayes method, namely fast processing time, not failing thoroughly in results, and easy implementation [17]. The *Naive Bayes* algorithm works by calculating opportunities based on *Bayessian Rules*, as follows:

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)} \quad (1)$$

While C5.0 is an interactive algorithm that gives different weights to training and testing data in each iteration [18]. The C5.0 algorithm works by adding weight to each of the incorrect classification examples and decreasing the weight of each sample of the correct classification for each iteration, so that the distribution of training data can work effectively.

Both methods can be evaluated so that the best classification method for sentiment analysis can be determined. The table that can be used to evaluate the performance of the sentiment analysis method is the Confusion Matrix Table (Table 2) [19].

Table 2. Confusion matrix of three sentiment classes

Actual	***Prediction***		
	Positive	***Negative***	***Neutral***
Positive	*True Positive* (TP)	*False Negative1* (FNg1)	*False Neutral1* (FNt1)
Negative	*False Positive1* (FP1)	*True Negative* (TNg)	*False Neutral2* (FNt2)
Neutral	*False Positive2* (FP2)	*False Negative2* (FNg2)	*True Neutral* (TNt)

Source [19]

Using the Confusion Matrix Table (Table 2), several evaluation measures such as *accuracy*, *precision*, and *recall* can be formulated, as follows [19]:

$$accuracy = \frac{TP + TNg + FNt}{TP + FNg1 + \ldots + FNg2 + TNt}.100\% \quad (2)$$

$$precision.positive = \frac{TP}{TP + FP1 + FP2}.100\% \quad (3)$$

$$precision.negative = \frac{TNg}{FNg1 + TNg + FNg2}.100\% \quad (4)$$

$$precision.neutral = \frac{TNt}{FNt1 + FNt2 + TNt}.100\% \quad (5)$$

$$recall.positive = \frac{TP}{TP + FNg1 + FNt1}.100\% \quad (6)$$

$$recall.negative = \frac{TNg}{FP1 + TNg + FNt2}.100\% \quad (7)$$

$$recall.neutral = \frac{TNt}{FP2 + FNg2 + TNt}.100\% \quad (8)$$

III. RESULTS AND DISCUSSION

Fig. 1 shows the words used as *keywords*, where the *keywords* are the result of the *Topics Modeling* analysis using the *Latent Dirichlet Allocation* (LDA) method. Some examples of keywords obtained from *Topic Modeling*, among others: *rabucoblos, pilkadadamai, pilkadajurjur*, etc (Fig. 1). In addition, the keywords indicate that opinions on the Governor's Election relate to the Presidential Election held in 2019. This is indicated by keywords about Candidates for Governors who are linked to the Presidential Candidates they support.

After tweets have been collected, the next step is to pre-processing. Table 3 shows the results of Indonesian language pre-processing, where pre-processing converts unstructured textual data into structured data. For example, before the pre-processing tweet contains the phrase "*Senangnya, Ridwan Kamil dapat berkunjung ke desa kami*", the results of the pre-processing produce "*senang ridwan kamil kunjung desa*".

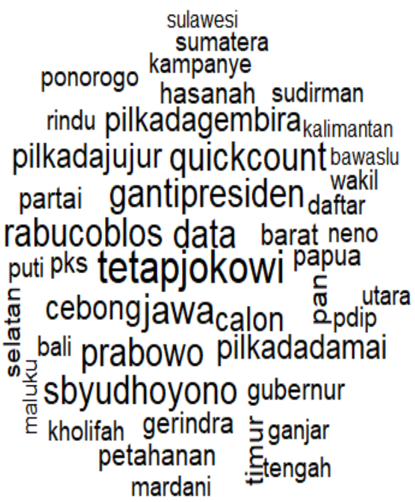


Figure 1 : *Wordcloud* about political topics
Source: processed results R 3.5.1

Table 3. Examples of tweets before and after pre-processing

Before *Pre-processing*	After *Pre-processing*
Senangnya, Ridwan Kamil dapat berkunjung ke desa kami	*senang ridwan kamil kunjung desa*
Tadi malam ada debat seru antara Edy dan Jarot	*debat seru edy jarot*
Tinggi mana jabatan Menteri atau Gubernur? Kok Khofifah berhenti jadi Menteri	*jabat menteri gubernur khofifah menteri*

Source: processed results R 3.5.1

3.1 Exploration of *Sentiment Mining*

After the pre-processing stage is complete, the next step is the lexicon based sentiment mining process. The results of the sentiment mining are presented in the form of pie charts according to the provinces implementing the Governor's Election in 2018. Fig. 2 shows that the percentage of positive sentiments in Riau, South Sumatra and Lampung provinces is the largest compared to negative and neutral sentiments. This shows that people in these 3 provinces mostly responded to the Governor's Election with positive sentiment. While North Sumatra Province, the percentage of neutral sentiment is the largest compared to positive and negative sentiments, which is 37%. The biggest percentage of positive sentiment is in South Sumatra Province, which is 44%. While the largest percentage of negative sentiment is in Lampung Province, which is 33%.

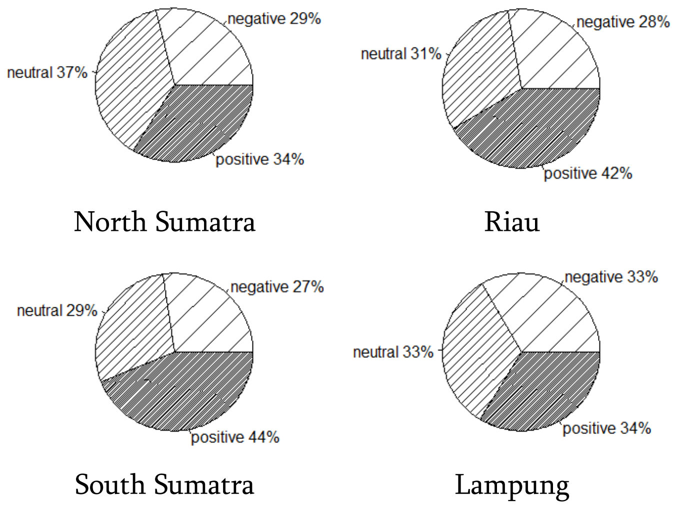
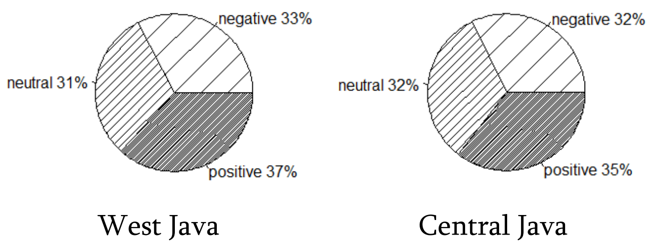


Figure 2: Sentiment on the island of Sumatra
Source: processed results R 3.5.1

Fig. 3 shows the percentage of positive sentiment in West Java and Central Java Provinces is greater than negative and neutral sentiments. This shows that people in these 2 provinces mostly respond to the Governor's Election with positive opinions. While East Java Province has the same percentage of positive and neutral sentiments, which is 35%. Central Java Province also has the same negative and neutral sentiment percentage, which is 32.5%. The highest percentage of positive and negative sentiment in Java is West Java, which is 37% positive sentiment and 33% negative sentiment.



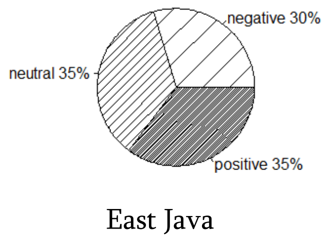


Figure 3 : Sentiment on the island of Java
Source: processed results R 3.5.1

Fig. 4 shows Bali province has a higher percentage of negative sentiment than positive and neutral sentiments, which is 35%. While NTB province has a percentage of neutral sentiment which is higher than positive and negative sentiments, which is equal to 36%. NTT Province has the largest percentage of positive sentiment compared to negative and neutral sentiments, which is 37%.

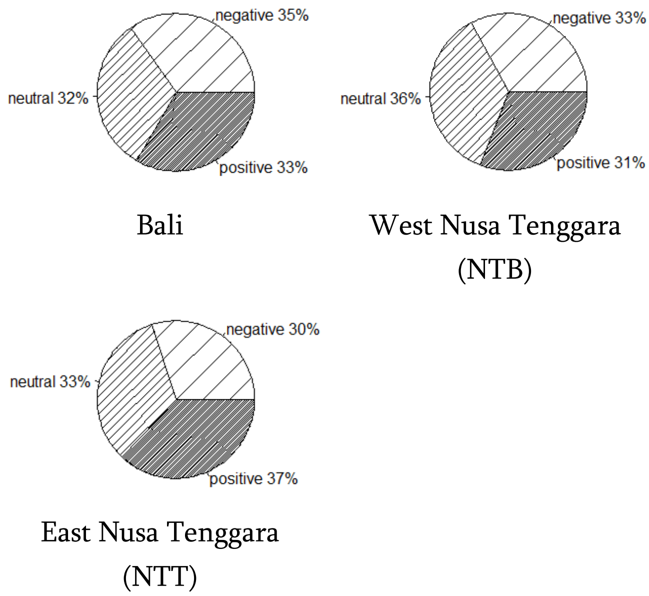


Figure 4 : Sentiment in Bali and Nusa Tenggara
Source: processed results R 3.5.1

Fig. 5 shows that the percentage of neutral sentiments in the provinces of East Kalimantan, South Sulawesi and Southeast Sulawesi is greater than the negative and positive sentiments. This shows that most people are neutral in responding to the Governors Election in these three provinces. While the positive sentiment in West Kalimantan Province is higher than the negative and neutral sentiment, which is 35%.

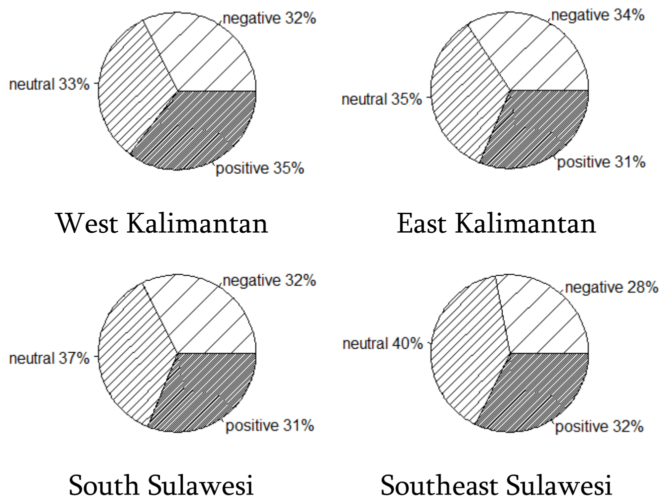


Figure 5 : Sentiment in Kalimantan and Sulawesi
Source: processed results R 3.5.1

Fig. 6 shows the percentage of positive sentiment in Maluku and North Maluku Provinces is higher than negative and neutral sentiments. This shows that people in these 2 provinces mostly respond to the Governor's Election with positive opinions. While the Papua Province has the same percentage of positive and neutral sentiments, which is equal to 34%.

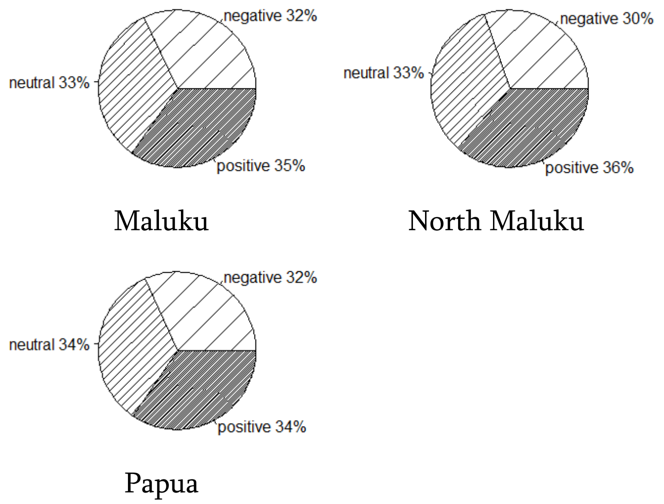


Figure 6 : Sentiment in Maluku and Papua
Source: processed results R 3.5.1

Fig. 7 shows the percentage of positive sentiment in 17 provinces is greater than the negative and neutral sentiment, which is equal to 35%. The percentage of the Governor's Election sentiment in 17 provinces was strongly influenced by the sentiment of the Governor's election from Java. This is because the number of tweets from 3 provinces in Java is greater than the number of tweets from 14 provinces outside Java. The negative and neutral sentiment percentage in 17 provinces was 31% negative sentiment and 34% neutral sentiment.

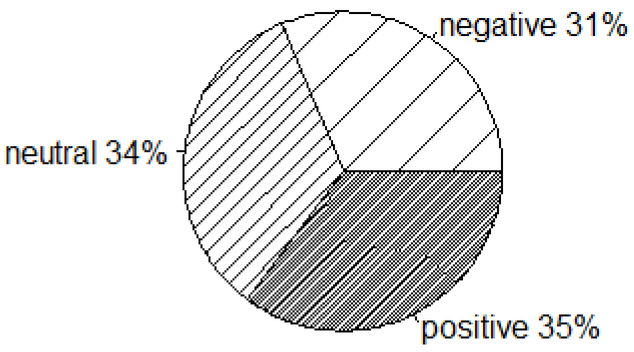


Figure 7 : Sentiment in seventeen provinces
Source: processed results R 3.5.1

3.2 Sentiment Analysis of the Naive Bayes Method

Table 4 shows the Naive Bayes method is able to predict the positive class sentiment as much as 3475 tweets, the exact prediction of the negative sentiment class is 1579 tweets, and the exact prediction of the neutral class is 1895 tweets. There are as many as 14443 tweets that are wrongly predicted using the Naive Bayes method.

TABLE 4. Confusion matrix with the Naive Bayes method

Actual	*Prediction* (tweets)		
	Positive	*Negative*	*Neutral*
Positive	3475	1883	2162
Negative	3207	1579	1882
Neutral	3503	1786	1895

Source: processed results R 3.5.1

Table 5 shows the application of the Naive Bayes method in 17 provinces produced an accuracy value of 32.51%, negative precision of 30.09%, and positive recall of 46.21%. South Sumatra Province has the highest accuracy value compared to the other 16 provinces, which is 37.20%. While West Java Province has the lowest accuracy value compared to 16 other provinces, which is equal to 28.05%. The highest positive precision value is produced by North Maluku Province, which is 87.56%, but the value of negative precision and neutral precision of North Maluku Province is quite low, which is 5.94% (negative precision) and 6.48% (positive precision). While the highest negative recall value was produced by the Province of Bali, which amounted to 37.13%.

Table 5. The evaluation results using the Naive Bayes method

Prov Code**	Accuracy	*Precision*			*Recall*		
		(-)*	(0)*	(+)*	(-)*	(0)*	(+)*
1	32.12	33.78	36.57	25.78	28.94	37.73	28.89
2	32.53	26.63	31.01	37.60	26.79	29.13	39.31
3	37.20	13.16	36.01	53.02	23.68	27.08	50.00
4	33.78	36.66	29.51	35.13	32.09	32.36	37.13
5	28.05	20.24	42.96	22.46	27.84	28.39	27.68
6	31.37	12.65	7.14	70.50	25.87	26.75	33.04
7	32.31	24.91	0.45	70.46	27.61	33.33	34.03
8	34.74	47.43	4.15	50.82	37.13	23.94	33.80
9	31.59	23.19	13.06	62.33	30.19	42.36	30.24
10	30.51	3.75	53.40	31.68	18.03	29.47	34.64
11	32.49	33.94	7.76	53.99	30.11	28.26	34.75
12	33.24	24.86	67.30	4.50	32.37	33.60	32.61
13	32.12	70.05	17.23	10.26	31.25	39.66	27.27
14	33.87	3.07	62.53	25.00	29.03	39.05	24.28
15	30.07	0.95	32.63	54.36	27.27	26.60	32.58
16	35.82	5.94	6.48	87.56	30.16	35.94	36.19
17	33.15	8.36	25.47	63.56	31.52	32.30	33.71
17 Prov	**32.51**	**30.09**	**31.91**	**34.12**	**23.68**	**26.38**	**46.21**

Source: processed results R 3.5.1
*Column Code: (+) Positive; (0) Neutral; (-) Negative
** Provincial Code follows Table 1

3.3 Sentiment Analysis of the C5.0 Method

Table 6 shows that the C5.0 method is able to predict the positive class correctly as many as 5084 tweets, the exact prediction of the negative class is 4188 tweets, and the right prediction is the neutral class of 4626 tweets. While there were 7474 incorrect tweets predicted using the C5.0 method.

Table 6. Confusion matrix with the C5.0 method

Actual	*Prediction* (tweets)		
	Positive	*Negative*	*Neutral*
Positive	5084	997	1439
Negative	972	4188	1508
Neutral	1176	1382	4626

Source: processed results R 3.5.1

Table 5 shows that the application of the C5.0 method in 17 provinces produced an accuracy value of 65.03%, positive precision of 70.30%, and negative recall of 62.81%. The Province of South Sumatra has the highest score for several evaluation values compared to the other 16 provinces, namely 73.84% (accuracy), 81.90% (precision positive), 70.43 (negative recall), and 80.72% (positive recall). Whereas North Maluku Province has the lowest score for several evaluation values compared to the other 16 provinces, namely 54.10% (accuracy), 38.75 (negative precision), 47.15% (negative recall), and 51.87% (neutral recall).

Table 7. The evaluation results using the Naive Bayes method

Prov Code**	Accuracy	*Precision*			*Recall*		
		(-)*	(0)*	(+)*	(-)*	(0)*	(+)*
1	64.62	62.21	66.24	64.87	60.19	62.11	72.24
2	64.03	54.80	56.15	76.03	61.46	58.43	69.04
3	73.84	67.54	67.59	81.90	70.43	66.49	80.72
4	63.78	69.65	58.61	63.07	63.93	57.66	70.38
5	64.82	64.42	61.50	67.96	66.55	57.89	69.63
6	66.37	71.97	60.88	66.30	67.47	59.87	71.76
7	70.01	64.06	70.96	74.06	67.04	65.02	78.29
8	60.51	61.74	63.17	56.67	62.30	54.64	65.94
9	57.86	62.09	59.46	51.46	60.29	55.46	58.26
10	62.65	57.00	62.96	66.94	60.73	60.00	66.58
11	64.20	61.52	58.51	71.90	65.48	57.99	68.68
12	65.26	58.01	64.86	73.57	63.25	61.86	71.01
13	66.28	64.01	65.53	69.52	68.33	63.08	68.16
14	72.81	65.53	77.09	73.81	66.90	74.60	75.61
15	70.43	69.94	75.53	65.99	70.16	67.20	74.67
16	54.10	38.75	62.54	59.07	47.15	51.87	61.62
17	60.53	61.67	62.60	57.45	54.45	66.19	61.71
17 Prov	65.03	63.77	61.09	70.30	62.81	64.39	67.61

Source: processed results R 3.5.1

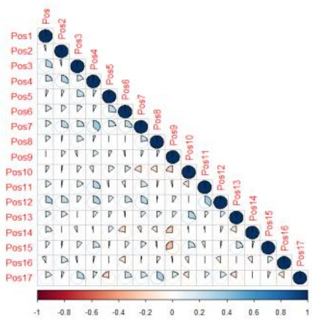
*Column Code: (+) Positive; (0) Neutral; (-) Negative

** Provincial Code follows Table 1

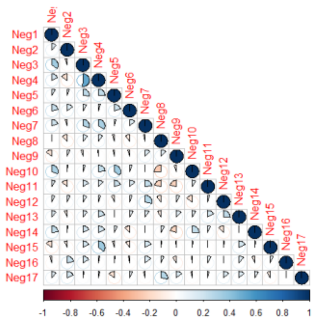
So it can be concluded that the C5.0 method is a better method than the Naive Bayes method in predicting the opinion sentiment class of Governor Election in 17 provinces based on the values of accuracy, precision, and recall.

3.4 Sentiment Correlation for Each Province

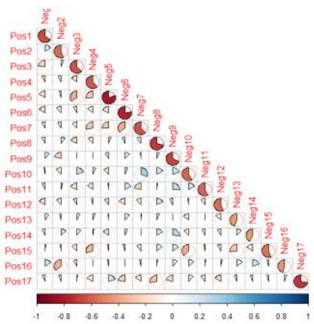
Fig. 8A shows there is a positive relationship between positive sentiment and between provinces, this is marked by a bluish color of several variables. Most of the relationship between negative sentiment and between provinces also shows a positive relationship (Fig. 8B). While the relationship between positive sentiment and negative sentiments and between provinces showed a negative relationship, this was shown by most correlation of reddish yellow images (Fig. 8C). The correlation test results for 408 correlations that occurred between sentiment and between provinces showed that there were 254 correlations that were statistically significant at the 5% level.



A. Correlation between Positive and Positive Sentiments



B. Correlation between Negative and Negative Sentiments



Correlation between Positive and Negative Sentiments

Figure 8 : Source: processed results R 3.5.1

IV. CONCLUSION

Sentiment mining applied to the opinion of the Governor's Election in 17 provinces resulted in a greater percentage of positive sentiment compared to negative and neutral sentiments. Only Bali Province has negative sentiments greater than positive and neutral sentiments. The results of classification prediction in the sentiment class show that C5.0 method is a better method than the Naive Bayes method. This is indicated by the value of accuracy, precision, and recall method C5.0 which is higher than the Naive Bayes method. Most of the positive sentiment relations of each province and negative sentiment relations in each province showed a positive relationship. While the relationship between positive sentiment and negative sentiment in each province shows a negative relationship.

Future studies can add the number of provinces and periods of Twitter crawling. It is recommended to conduct sentiment analysis on other social media, such as Facebook. Methods other than Naive Bayes and C5.0 can be used in subsequent studies so that the results can be compared with the results of this study.

V. REFERENCES

[1] Kwak, H., Lee, C., Park, H., and Moon, S.. 2010. “What is Twitter , a Social Network or a News Media ?”. *International World Wide Web Conference Com-mittee (IW3C2)*. pp. 591–600.

[2] Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. 2010. “Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment”. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media Predicting*. pp. 178–185.

[3] Hanna, A., Wells, C., Maurer, P., Friedland, L., Shah, D., and Matthes, J. 2013. “Partisan Alignments and Political Polarization Online : A Computational Approach to Understanding the French and US Presidential Elections”. *PLEAD’13*.

[4] Indonesian Internet Service Providers Association. 2017. "*Penetration & Behavior of Indonesian Internet Users*".

[5] Ding, X., Liu, B., and Yu, P.S. 2008. “A Holistic Lexicon-Based Approach to Opinion Mining,” *Proceeding of the 2008 International Conference on Web Search and Data Mining*. pp. 231–240.

[6] Maks, I., and Vossen, P. 2012. “A lexicon model for deep sentiment analysis and opinion mining applications,” *Decis. Support Syst.* 53:680–688.

[7] Gentry, J. 2016. “Package ‘ twitteR’”.

[8] Gentry, J, and Lang, D.T. 2015. “Package ‘ ROAuth ’” .

[9] Fellows, I. 2018. “Package ‘ wordcloud ’”.

[10] Grun, B., and Hornik, K. 2017. “Package ‘ topicmodels ’”.

[11] Feinerer, I., and Hornik, K. 2018. “Package ‘ tm ’”.

[12] Wickham, H. 2018. “Package ‘ stringr ’” .

[13] Harrell, F.E.J. 2018. “Package ‘ Hmisc ’”.

[14] Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., and Zemla, J. 2017. “Package ‘ corrplot ’”.

[15] Meyer, D., Dimitriadou, E., Hornik, K., Wingessel, A., Leisch, F., Chang, C.C., and Lin, C.C. 2018. “Package ‘e1071’”.

[16] Pang, B., and Lee, L. 2008. “Opinion Mining and Sentiment Analysis,” *Inf. Retr. Boston.*, 2(1-2): 1–135.

[17] Manning, C.D., Raghavan, P., and Schutze, H. 2008. "Introduction to Information Retrieval". New York: Cambridge University Press.

[18] Alam, F., Mehmood, R., Katib, I., and Albeshri, A. 2016. “Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT),” *Procedia Comput. Sci.* 98:437–442.

[19] Han, J., Kamber, M., and Pei, J. 2012. "Data Mining – Concepts and Techniques Third Edition". Morgan Kaufmann Publishers.

Cite this article as :

Agung Eddy Suryo Saputro, Khairil Anwar Notodiputro, Indahwati, "Study of Sentiment of Governor's Election Opinion in 2018", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 4 Issue 11, pp. 231-238, November-December 2018. Available at doi : https://doi.org/10.32628/IJSRSET21841124, Journal URL : http://ijsrset.com/IJSRSET21841124