MATEMATIKA: MJIAM, 2020, Volume 36, Number 1, 31–42

Surabaya Government Performance Evaluation using Tweet Analysis

Kartika Fithriasari*, Rakhmah Wahyu Mayasari, Nur Iriawan and Wiwiek Setya Winahju

Statistics Department
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember
60111 Surabaya, Indonesia

*Corresponding author: kartika_f@statistika.its.ac.id

Article history
Received: 25 March 2019
Received in revised form: 6 October 2019
Accepted: 12 November 2019
Published online: 1 April 2020

Abstract The purpose of this research is to determine the various positive attributes appreciated by the public, and the negative things that need to be improved by the Surabaya government. The sentiment analysis methods, including the Naïve Bayes Classifier, Support Vector Machine, and Logistic Regression, are employed to classify the pros and cons of the Surabaya government. The comparison of the three methods demonstrated that SVM gives the best classification accuracy compared to others. Police performance is the highlighted word in the positive category, while traffic congestion is in the negative category.

Keywords Logistic Regression; Naïve Bayes Classifier; Support Vector Machine; Sentiment Analysis

Mathematics Subject Classification 62C10, 62J12, 68T20

1 Introduction

Government can be defined as an administrative institution which has authority over the activities of people in a country, city, and others. According to Law of the Republic of Indonesia, Number 32 of 2004 about Local Government is authorized to regulate its affairs according to the principle of the autonomy. The Surabaya City Government implemented this law. The implementation of the autonomous principle carried out by the Surabaya government produced various responses from the community, most of which were easily conveyed through social media. Therefore, twitter is one of the social media platforms widely used by the Surabaya city government, owing to its unique maximum number of users. @SapawargaSby is the Twitter account username of the Surabaya City Government where many of the people express their opinions via a tweet addressed to the account. They also have a radio account known as @e100ss where people convey their opinions about the government. This opinion and feedback

from the community is an important data used as an evaluation material to better performance of the city. This data is obtained from the twitter account's API (Application Programming Interface). Therefore, further analysis can be conducted. Before carrying out the classification analysis, the extracted data needs to be pre-processed with the irrelevant materials eliminated. This procedure which is known as sentiment analysis is used to classify public tweets into positive or negative. Although there are many classification analysis methods which can be employed, this study will be focused on the use of the Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and Logistics Regression. These three methods are chosen based on previous research. Therefore, several studies has demonstrated that the accuracy using NBC was higher than the Logistic Regression method [1]. Another study on the classification of online news reporting comparing between the SVM and KNN methods concluded that the SVM was better than the NBC [2]. Furthermore, research on classification analysis on the online news was also carried out using SVM linear kernel, SVM RBF kernel, and NBC which concluded that the SVM method was better than the others, with linear and RBF kernels producing same results [3]. Another study was also conducted using the Bayesian Bernoulli mixture regression which was better than Bayesian binary logistic regression [4]. In this study, classification analysis will be carried out on the tweet data by employing the NBC, SVM linear kernel, SVM RBF kernel, and Binary Logistic Regression. The best method will be chosen by comparing their classification performance.

2 Literature Review

2.1 Text Mining

Text mining is a branch of data mining science that analyze text. Text mining is conducted automatically by a computer to dig up quality information from a series of texts summarized from a document. The initial idea to its development was to find patterns of information that could be extracted from unstructured text [3]. Text grouping methods are two types such as clustering and classification, with text clustering related to the process of finding an unsupervised group structure from a set of documents. Meanwhile, text classification can be considered as a process to form groups (classes) of documents based on groups that have been previously known. One continuation of the classification process is sentiment analysis, which is a computational technique that examines opinions, feelings, and emotions expressed in a text. A text often represents an opinion or a subjective matter about events and activities. The main focus of sentiment analysis is to group opinions into positive or negative. It can also express emotional feelings such as joy, anger, or sadness [5].

2.2 Text Preprocessing

Text preprocessing is the step taken to prepare the data that need to be processed. This step has to be conducted because the raw data obtained is usually in unstructured form making it impossible to carry out the text mining process. These stages include cleansing, case folding, stemming, stopwords, tokenizing, and weighting of Term Frequency-inverse Document Frequency (TF-IDF) [3]. The formula used in the TF-IDF stage is as follows.

$$w_{ij} = tf_{ij} \times idf. \tag{1}$$

2.3 K-Fold Cross-Validation

K-fold cross-validation is one method used to partition data into training and testing categories, where each data will have the opportunity to get tested. K-fold cross-validation is widely used since it can reduce the biasness that occurs during sampling process [6]. K represents the data partition number used to split the training-testing. An illustration to split the training-testing using K-fold cross-validation is shown in Figure 1.

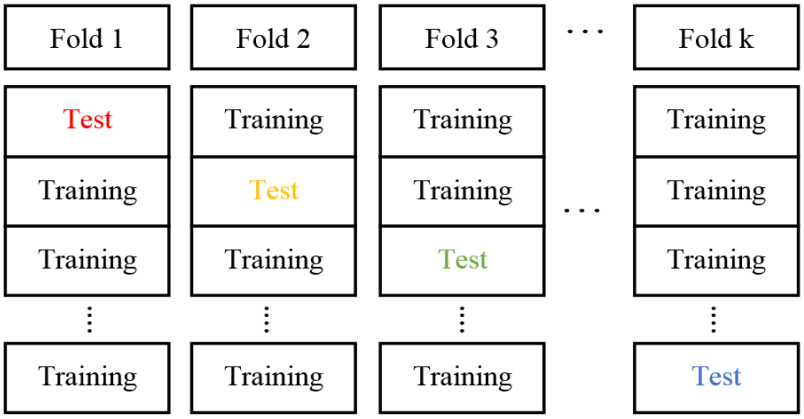

Figure 1: Data Splitting

2.4 Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) is a method used to measure the highest probability of classified data used [7]. In this algorithm, each document is represented by attributes $a_1, a_2, \ldots, a_n$ where a_1 is used to denote the first word, down to the nth word with V used to represent a set of tweets. At the time of the algorithm classification, the highest probability of all categories of documents tested (V_{MAP}) is found in equation (2). With the $P(v_j)$, value calculated during training. The $P(a_i|v_j)$, value for the probability of the word a_i for each category can be obtained from equations (3) and (4).

$$V_{MAP} = \arg\max P(v_j) \prod_{i}^{n} P(a_i|v_j), \tag{2}$$

$$P(v_j) = \frac{|doc_j|}{|training|}, \tag{3}$$

$$P(a_i|v_j) = \frac{n_i + 1}{|n + kosa\,kata|}, \tag{4}$$

2.5 Support Vector Machine

The basic concept of Support Vector Machine (SVM) is actually a combination of several concepts that existed before. This machine was developed with the linear classifier principle.

However, in most cases, data is not linear. SVM is development for nonlinear cases by incorporating the kernel concept [8]. Data can be said to be linear when the boundary between classifications is linear, and it is non-linear when it cannot be split linearly. In real life, a lot of data is found to be non-linear, so the kernel tricks is used to carry out the analysis. The linear kernel and Radial Basis Function can be seen in Table 1 below.

Table 1: Kernel Functions

No	Kernel Name	Kernel Function
1.	Linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_j + C$
2.	Radial Basis Function (RBF)	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i'\mathbf{x}_j\right\|\right)^2, \gamma > 0$

2.6 Logistic Regression

Logistic regression is a method that can be used to determine the relationships between dichotomous (two categories) or polychotomous (more than two categories) response variables with one or more category or continuous predictor variables [9]. The model obtained, can be used as a model in classifying the predictor variables into its response form. Assuming that a set of independent variable p is shown as $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_p)$, then the multivariable logistic regression form is as follow in equation (5).

$$g(\mathbf{x}) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \tag{5}$$

Furthermore, the logistic regression model (p), is the number of independent variables as in equation (6) below.

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}. \tag{6}$$

2.7 Classification Performance

Classification performance measurement is conducted to visualize the results obtained from the process. There are several ways to measure performance, some of which are to calculate accuracy, precision, and recall. Accuracy is a percentage of the total documents correctly identified in the classification procedure [10]. These performance measurements can be obtained from the confusion matrix. It can also be found using the Area Under Curve (AUC) calculation as in equation (7).

$$\text{AUC} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right), \tag{7}$$

Where TP denotes (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). The performance classification for the interpretation of AUC values can be seen in Table 2.

Table 2: Interpretation of AUC Values

AUC Value	Model Performance
0.5 - 0.6	Poor
0.6 - 0.7	Fair
0.7 - 0.8	Good
0.8 - 0.9	Very Good
0.9 - 1.0	Excellent

2.8 Word Cloud

Word Cloud is a method often used for describing text data by plotting words that often appear. The more a word appears, the greater the letters formed from it. And the lesser the appearance of a word, the smaller its size when compared to others [11].

3 Methodology

3.1 Data Source

The data used in this study is a collection of tweets obtained from Twitter users in Indonesia from September 28, 2017, to May 7, 2018. These data were obtained using Twitter API (Application Programming Interface) on the official account of the Surabaya City Government (@SapawargaSby) and Radio Suara Surabaya (@e100ss).

3.2 Research Variable

In this research, data sentiment was used to categorize the obtained data into positive and negative groups. The following research variables are used as in Table 3.

Table 3: Research Variable

Variable	Explanation	Data Scale
Y	Sentiment (Positive/Negative) 0 = Negative 1 = Positive	Nominal
X	The frequency of the i-word that appears on the object (twitter)	Ratio

The data structure used in this study after pre-processing the tweet text consists of the response variable (Y) and predictor variable (X).

4 Result and Discussion

The data obtained is a compilation from the two accounts government accounts namely @SapawargaSby and @e100ss. Figure 2 is a comparison of the number of tweets from those accounts.

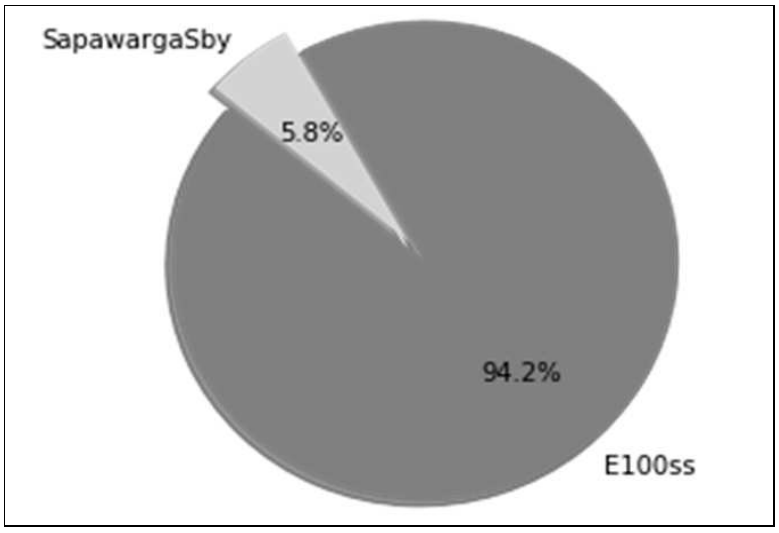

Figure 2: Comparison of Data Sources

In Figure 2, it can be seen that 94.2% of the data was obtained from the official Radio Suara Surabaya account while the remaining 5.8% was from the government account. All the data must be carried out in several stages, therefore, the analysis can be continued with filtering carried out by removing data that does not contain sentiments about the City Government of Surabaya. Furthermore, it is classified as positive or negative sentiments. From all the data containing sentiments, only 21.9% were positive. This can be seen in Figure 3.

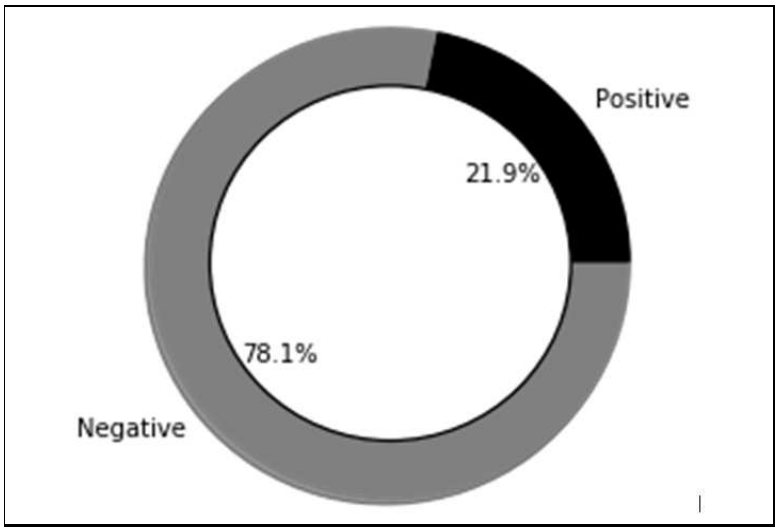

Figure 3: Characteristics of Tweet Data

From Figure 3, it can be seen that as many as 78.1% of the people expressed their opinion, with the remaining 21.9% appreciating the Surabaya city government.

Data that contains sentiments are then preprocessed by cleansing, case folding, stemming, stopwords, and tokenizing. The cleansing stage is conducted to eliminate words that are not needed such as HTML characters, URL links, usernames (@username), emoticons, and hashtag (#). The case folding stage changes the text character into lowercase letters. The stemming stage eliminates prefixes, suffixes, inserts, and confixes (a combination of prefixes and suffixes). Hence, the words formed are basic words. The next step which is the stopwords, removes vocabularies that does not include a unique word or does not give any significant message to the text such as "dan", "dari", "di" and "yang". The words that were omitted were the words "Surabaya", "SapawargaSby", and "e100ss" because they would often appear and were considered unimportant. The tokenizing stage decides the meaning of each word in the sentence. This stage aims to divide the original form into words.

Pre-processing data has been carried out on all public tweet data on the Surabaya City Government which gives results as in Table 4.

Table 4: Word Calculation Result

Tweet	**Predictor Variable**						
	x_1 "Acara"	...	x_{1158} "Jatim"	...	x_{1580} "Lancar"	...	x_{3689} "Zona"
1	0	...	0	...	1	...	0
2	0	...	1	...	0	...	0
3	0	...	0	...	0	...	0
⋮	⋮	...	⋮	...	⋮	...	⋮
1687	0	...	0	...	0	...	0

Based on the results of preprocessing data calculations, it can be seen that the number of predictor variables or the word used is 3689. In the first variable is made up of the word "Acara", followed by 1158 variables of "Jatim" and the last variable is "Zona". After preprocessing the data, the next step that be done is the sentiment analysis.

The purpose of this study, is to obtain the best results using the sentiment verification. Hence, it is necessary to have the best classification method. Therefore, the Naïve Bayes Classifier, Support Vector Machine and Logistic Regression techniques were utilized. The data used was pre-processed and divided into training and testing categories using K-folds cross-validation. The number of folds used is 10, where the results of the AUC classification performance for each fold with the NBC method are as in Figure 4.

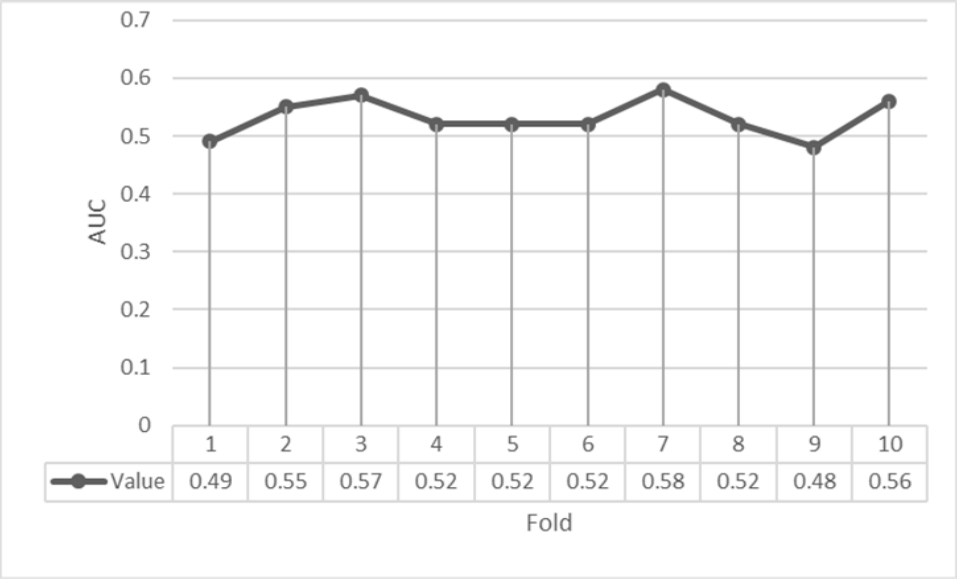

Figure 4: Selection the Best of Subset with NBC Method

Based on the Figure 4, it can be seen that the best AUC value is in the combination of the 7th fold which is 0.58. Note that the average 10-fold was 0.53. Hence, it can be seen that the model generated from the 7th fold is as in Table 5.

Table 5: Classification Model with NBC Method

Class	Model
Negative	$0.78 \times 0.00022^{(f_1)} \times \ldots \times 0.00050^{(f_{3687})} \times 0.00036^{(f_{3688})}$
Positive	$0.22 \times 0.00050^{(f_1)} \times \ldots \times 0.00116^{(f_{3687})} \times 0.00083^{(f_{3688})}$

From the model in Table 5, the classification can be performed on the next data. This is conducted by entering the frequency of words in f_1 to f_{3688} in each class. The class with the best score is selected.

The next sentiment analysis is using the SVM classification method, where there are two used kernels namely linear and RBF. In the linear kernel SVM method, the best average AUC value will be selected in the specified parameters using the cost (C) and starting from 10^{-2} to 10^2.

Figure 5 shows that the best C value is at 10^2, resulting in an AUC value of 0.65. This is proceeded with the SVM RBF kernel method, which has two parameters that must be determined, namely the cost parameter (C) and gamma. Figure 6 is a comparison of the average AUC values for each parameter C and gamma.

Based on Figure 6, it can be seen that the optimum average AUC value in the parameter cost is 10^2 and the parameter gamma is 10^{-1}. In this combination, the optimum average AUC is 0.64. Based on these, the kernel function is defined as follows.

$$K\left(\mathbf{x_1}, \mathbf{x_2}\right) = \exp\left(-0.1 \left\|\mathbf{x_1}, \mathbf{x_2}\right\|^2\right). \tag{8}$$

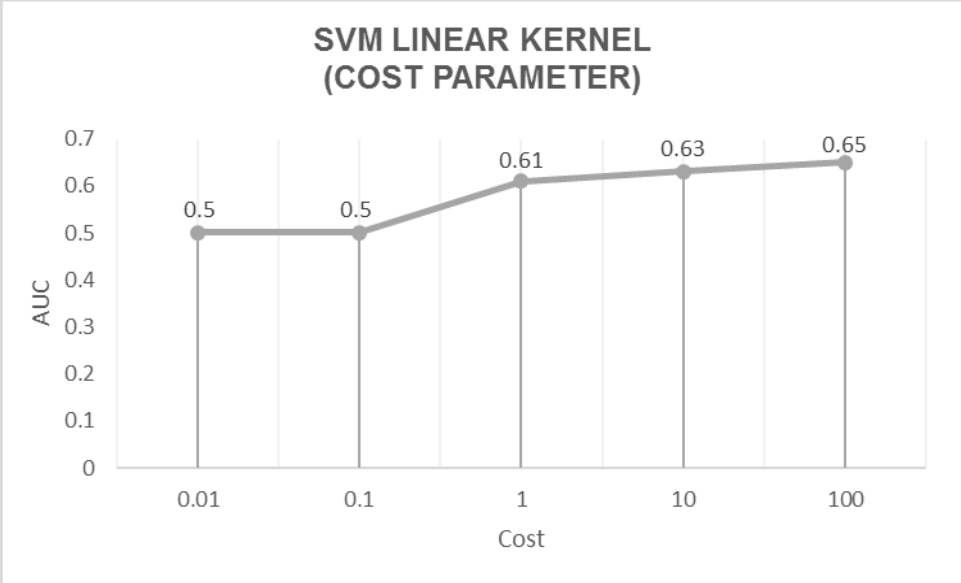

Figure 5: Selection the Best of Parameter Cost SVM Linear

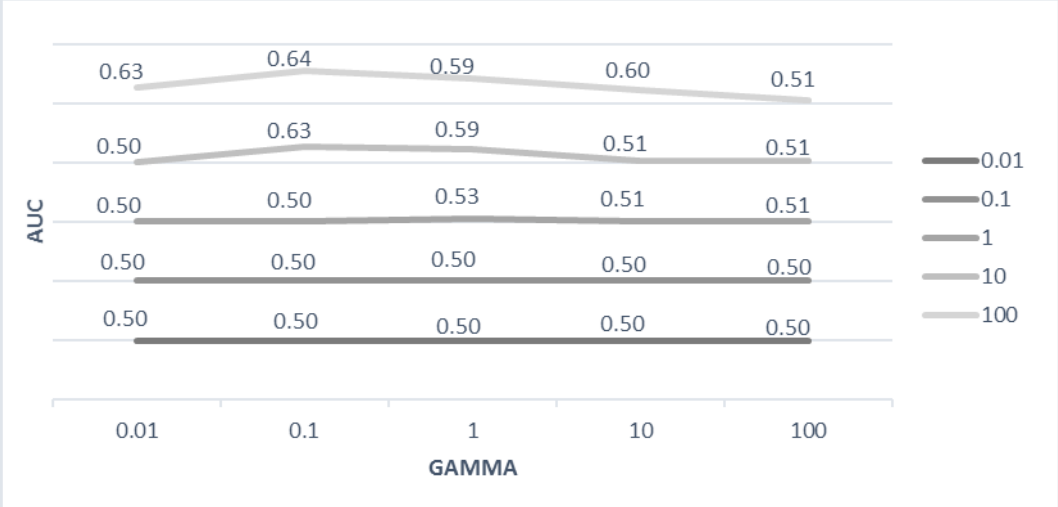

Figure 6: Selection the Best of Parameter Cost SVM RBF

The last classification analysis is logistic regression, just like the previous that splits training and testing using 10 fold cross-validation with the results as illustrated in Figure 7.

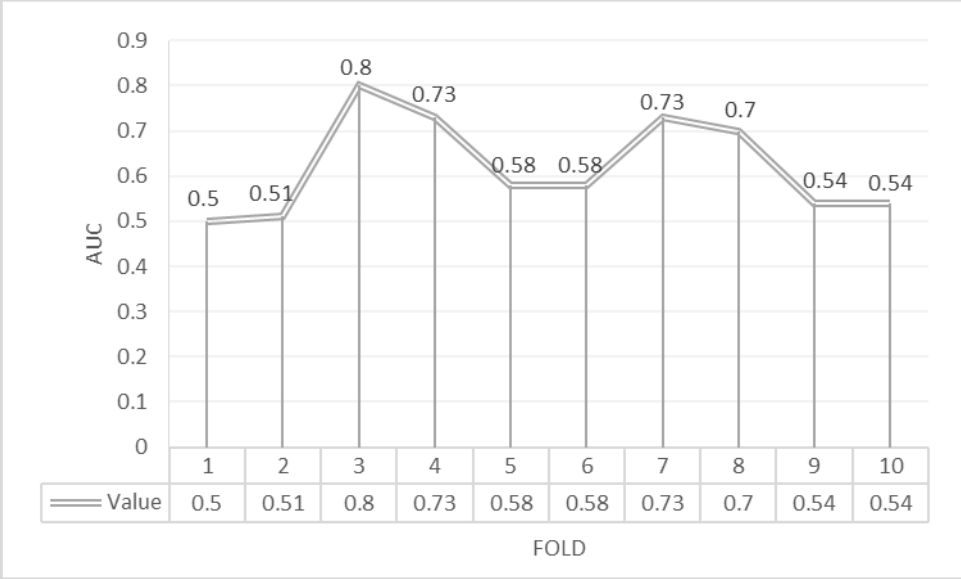

Figure 7: Selecting the Best Subset using Logistic Regression Method

Based on the picture in Figure 7, it can be seen that the best AUC value of 0.80 is at the 3rd fold with an average of 0.62. As in equation (3), a model from 3rd fold can be performed as shown in equation (9) below.

$$\hat{\pi}(x) = \frac{\exp\left(-8.69 \times 10^{34} - 2.05 \times 10^{34} X_1 + \ldots + 8.58 \times 10^{39} X_{3689}\right)}{1 + \exp\left(-8.69 \times 10^{34} - 2.05 \times 10^{34} X_1 + \ldots + 8.58 \times 10^{39} X_{3689}\right)}. \tag{9}$$

After carrying out all classification analysis and determining the average value for each matrix, the best classification method can be obtained.

Based on the comparison of Table 6, it can be seen that the method that provides the best classification accuracy is SVM RBF kernel. This can be seen from the value of AUC that the SVM Linear kernel method which is higher than the other methods. Based on the interpretation of the AUC Values in Table 2, which were in the range of 0.5 - 0.7, it can be concluded to being in the fair category. The next stage that will be carried out is word cloud visualization.

Table 6: Best Classification Method

Method	AUC
NBC	0.53
SVM Linear	**0.65**
SVM RBF	0.64
Logistic Regression	0.62

Data visualization using the word cloud is used to determine predictor variables (words) that often appear on tweets. This visualization is carried out on data that has been preprocessing, with the following results as in Figure 8.

(a) (b)

Figure 8: Word Cloud Visualization

The visualization of Word cloud in Figure 8(a) shows that it contains positive sentiments, while Figure 8(b) illustrates data classified as negative. It can be seen that many people appreciate the Surabaya City Government owing to the outstanding performance of the police with regards to deciphering jam-traffic and regulating traffic. The negative comments are tailored at road traffic in Surabaya.

5 Conclusion

Based on the analysis discussed earlier, it can be concluded that the community is more active in expressing their opinions to the Radio Suara Surabaya twitter account (@e100ss) than the official Surabaya City Government account (@SapawargaSby). Furthermore, their opinions are expressed by complaints or negative sentiments. Classification performance result, of the SVM RBF kernel is better than the Naïve Bayes Classifier, SVM Linear kernel, and Logistic Regression. From all the obtained data, it can be seen that the community highly appreciates the police in road arrangements, but there are still many people who complained about traffic jams in Surabaya. However, the obtained results, can be used as an input medium by the Surabaya city government to fix traffic congestions.

Acknowledgment

The Authors are grateful to the Directorate for Research and Community Service (DRPM) Ministry of Research, Technology, and Higher Education Indonesia for supporting this research under the PDUPT (Penelitian Dasar Unggulan Perguruan Tinggi) research grant no. 907 / PKS / ITS / 2018 dated February 1, 2018.

References

[1] Salim, A. *Optimization of Naïve Bayes and Logistic Regression Using Genetic Algorithm for Classification Data*. Master's Thesis. Institut Teknologi Sepuluh Nopember. 2017.

[2] Asiyah, S. N. and Fithriasari, K. Klasifikasi berita online menggunakan metode support vector machine dan k-nearest neighbor. *Jurnal Sains dan Seni ITS*. 2016. 5(2): 132–147.

[3] Dio, A. and Fithriasari, K. Klasifikasi berita indonesia menggunakan metode naïve bayesian classification dan support vector machine dengan confix stripping stemmer. *Jurnal Sains dan Seni ITS*. 2016. 4(2): 157–165.

[4] Suryaningtyas, W., Iriawan, N., Fihriasari, K., Ulama, B. S. S., Susano, I. and Pravitasari, A. A. On the Bernoulli mixture model for bidikmisi scholarship classification with Bayesian MCMC. *Journal of Physics: Conf. Series*. IOP Publishing. 2018. 1090(1): 12–32.

[5] Weiss, S. M., Indurkhya, N., Zhang, T. and Damerau, F. J. *Text Mining Predictive Methods for Analyzing Unstructured Information.* New York: Springer Science+Business Media. Inc. 2005.

[6] Gokgoz, E. and Subasi, A. Comparison of decision tree algorithms for EMG signal classification using DWT. *Biomedical Signal Processing and Control*. 2015. 18: 138–144.

[7] Feldman, R. and James, S. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press. 2007.

[8] Nugroho, A. S., Witarto, A. B. and Handoko, D. Support vector machine: teori dan aplikasinya dalam bioinformatika. *Kuliah Umum Ilmu Komputer.com*. 2007.

[9] Hosmer, D. W. and Lemeshow, S. *Applied Logistik Regression*. New York: John Wiley & Sons, Inc. 2000.

[10] Hotho, A., Nurnberger, A. and Pass, G. Brief survey of text mining, LDV forum. *GLDV Journal of Computational Linguistic and Language Technology* 2005. 20: 19–62.

[11] Castella, Q., and Sutton, C. Word storms: multiples of word clouds for visual comparison of documents. *23rd International Conference on world wide web WWW*. 2014. 14: 665–676.