MATEMATIKA, 2018, Special Issue, 91–102

Linking Twitter Sentiment Knowledge with Infrastructure Development

Zakya Reyhana*, **Kartika Fithriasari, Moh. Atok and Nur Iriawan**
Institut Teknologi Sepuluh Nopember
Statistics Department, Surabaya 60111, Indonesia
*Corresponding author: zakyareyhana@gmail.com

Article history
Received: 22 August 2018
Received in revised form: 6 December 2018
Accepted: 17 December 2018
Published on line: 31 December 2018

Abstract Sentiment analysis is related to the automatic extraction of positive or negative opinions from the text. It is a special text mining application. It is important to classify implicit contents from citizen's tweet using sentiment analysis. This research aimed to find out the opinion of infrastructure that sustained urban development in Surabaya, Indonesia's second largest city. The procedures of text mining analysis were the data undergoes some preprocessing first, such as removing the link, retweet (RT), username, punctuation, digits, stopwords, case folding, and tokenizing. Then, the opinion was classified into positive and negative comments. Classification methods used in this research were support vector machine (SVM) and neural network (NN). The result of this research showed that NN classification method was better than SVM.

Keywords Classification; Support Vector Machine; Neural Network; Text Mining; Twitter.

Mathematics Subject Classification 03C45, 62M45, 68T20.

1 Introduction

Surabaya city is the second largest city in Indonesia. The problem that most often occurs in urban city is congestion and several infrastructure problems. Mrs. Tri Rismaharini as the mayor of Surabaya had said that the city government would put forward proposals related to accessibility for infrastructure, the economy, education, and health [1]. Surabaya citizen dispenses their opinion about Surabaya's accessibility through social media platform. Twitter is a social networking and microblog service that allow users to send and read text-based messages up to 140 characters, known as tweets. From citizen's tweet, the aims to classify implicit contents can be done by sentiment analysis method. Sentiment analysis is a computing-based detection and learning of opinions (sentiments), emotions, and subjectivity in the text. As a special text mining application, sentiment analysis is related to the automatic extraction of positive or negative opinions from the text [2]. Tweet data is classified into positive and negative opinions using the text classification method. SVM is a technique suitable for negative

positive (binary) classification cases [3]. Besides SVM, another method that can be used is the neural network (NN). The NN method can be used to model complex relationships between input and output to find patterns in data. The NN method has been widely applied in various fields, including time series [4], regression [5], and classification [6]. According to Habibi [7], the percentage value of performance evaluation produced in his research shows that the classification of sentiment analysis with NN method with backpropagation algorithm gives good results.

The tweet data classification researches use the SVM method to classify negative and positive opinions. However, these studies focus more on discussing sentiments within certain companies or political parties. Whereas sentiment analysis can also be done on other topics that have much sentiments and opinions. The method used is the traditional method. So, this study formulated a problem, how to analyze sentiment on Twitter data about the infrastructure development of Surabaya City using the support vector machine and neural network methods. As for the performance measurement, text classification borrows from Information Extraction (IE) which initiated the use of Machine Learning in processing and understanding automatic text. Therefore, the criteria for comparing evaluation measures used in this study are accuracy, precision, recall, F1 score, and AUC.

2 Literature Review

Text mining is the mining of data in the form of text obtained from unstructured data in the form of sentences in the document. Then from the document are searched words that can represent the contents of the document to be able to analyze the core of the document. Text mining represents the ability to take unstructured languages in large numbers and quickly in extracting insights that are useful for decision making. These things are done without forcing someone to read the entire body of the text [8]. Sentiment analysis, also called opinion mining, is a field of study that analyzes people's opinions, sentiments, judgments, attitudes, and emotions towards the entity and is expressed in written texts. Entities from sentiment analysis can be products, services, organizations, individuals, events, issues, or topics [9]. The basic task in sentiment analysis is to classify the polarity of the text in the document or sentence. Working on text analysis often involves more processes than statistical analysis or machine learning. All machine learning algorithms, supervised or unsupervised techniques, usually begin with preprocessing before processing and analyzing data. The purpose of preprocessing is to eliminate noise, homogenize words and reduce the volume of words. The preprocessing stage consists of cleaning, case folding, parsing and filtering [10]. The information retrieval process starts when the user enters a query into the system. A reasonable valuation mechanism is to calculate a score which is the number of queries between terms (terms) in queries and documents. Determination of the weighting of the terms in each document which depends on the number of terms appearing on the document is carried out in the assessment mechanism, using TF-IDF, short of term frequency-inverse document frequency [11].

2.1 Support Vector Machine

The classification algorithm is a supervised machine learning algorithm that is used to classify or label data points based on what has been previously observed. There are many types of classification algorithms, but effective algorithms for text classification include Support Vector

Machine (SVM) and Naive Bayes. In addition to these two algorithms, several other algorithms are known, namely logistic regression, decision tree, and NN[12].

The SVM is a learning algorithm for classification. It tries to find the optimal separating hyperplane such that the expected classification error for unseen patterns is minimized. For linearly non-separable data, the input is mapped to high-dimensional feature space where they can be separated by a hyperplane. This projection into high-dimensional feature space is efficiently performed by using kernels. More precisely, given a set of training samples and the corresponding decision values, SVM aims to find the best separating hyperplane given by the equations that maximizes the distance between the two classes[13].

The main task in training SVM is to solve the following problem (1):

$$\max_{\alpha} L_D = \max \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathrm{K} \left(\mathbf{x_i} \mathbf{x_j} \right) \right) \tag{1}$$

subjected to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, n,$$

where C is the penalty parameter and K is kernel function.

The kernel function will map data into space with a higher dimension in order to be able to split the data. SVM is supported by the kernel function to separate data with very complex constraints. The SVM model will be formed by the equation (2).

$$f\left(x_k\right) = \sum_{i=1}^{n} \alpha_i y_i \mathrm{K} \left(\mathbf{x_i} \mathbf{x_j} \right) + b. \tag{2}$$

There exist many popular kernel functions that have been widely used for classification.

1. Linear Kernel
$$\mathrm{K} \left(\mathbf{x_i}, \mathbf{x_j} \right) = \mathbf{x}_i^T \mathbf{x}_j.$$
2. Polynomial Kernel
$$\mathrm{K} \left(\mathbf{x_i}, \mathbf{x_j} \right) = \left(\delta \mathbf{x}_i^{\mathbf{T}} \mathbf{x} + r \right)^{\mathbf{P}}, \ \delta > 0.$$
3. Radial Basis Function (RBF) Kernel
$$\mathrm{K} \left(\mathbf{x_i}, \mathbf{x_j} \right) = \exp \left(-\frac{1}{2\sigma^2} \left\| \mathbf{x_i} - \mathbf{x}_j \right\|^2 \right).$$
4. Sigmoid Kernel
$$\mathrm{K} \left(\mathbf{x_i}, \mathbf{x_j} \right) = \tanh \left(\mathbf{x}_i^T \mathbf{x}_j + r \right).$$

SVM performance is closely related to the selection of kernel functions. For the most methods are based on simple heuristics based on knowledge about data input. There is no standard method for getting the best kernel. Therefore, the optimal choice of the kernel has been reduced to a trial and error procedure [13], or tuning the parameter with grid search [14].

2.2 Neural Network

Text classification as a problem that cannot be solved sequentially or with sequential algorithms, NN provides a better solution. A very useful method in recognizing complex patterns and performing non trivial mapping functions of NN is Backpropagation [15]. Neural network model with backpropagation network is one of methods with the characteristic of multiple layer networks. The characteristic is that it has three types of layers, namely the input layer, hidden layer, and output layer. Backpropagation model uses a supervised learning method.

This method works by estimating the non-linear relationship between input and output by adjusting the value according to the minimum value of the error function so as to allow the network to center on a stable state and provide an appropriate output when receiving input that is not included in the training data pattern. Backpropagation architecture is illustrated as in Figure 1.

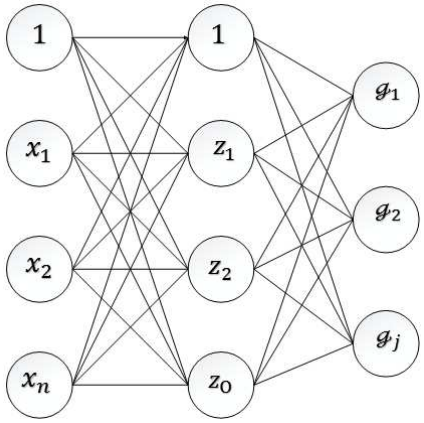

Figure 1: Backpropagation Neural Network Architecture

Discipline work of backpropagation network is divided into two stages which are feedforward propagation and backward propagation. For example, The first step of the learning is initial hypothesis. The next step to do after initialising the model at random, is to check its performance. This step is called forward-propagation. At this stage, we have the actual output of the randomly initialised neural network. The further step is differentiation. Basically it deals with the derivative of the loss function. Hereafter, we do backward-propagation. We have the starting point of errors, which is the loss function, and we know how to derivate it, and if we know how to derivate each function from the composition, we can propagate back the error from the end to the start. Afterwards, we update the weights. At last, we iterate it until convergence. The weight change cycle is performed on each set of training until it reaches the desired amount of weight or until the specified threshold value is exceeded [16]. The weights were adjusted as equations in (3) and (4):

$$z\_in_o = v_{0o} + \sum_{i}^{I} x_i v_{io}. \tag{3}$$

$$g \in_j = u_{0j} + \sum_{k}^{K} z_o u_{oj}, \tag{4}$$

where z indicates the hidden layer and g indicates the output layer.

2.3 Performance Measures

Cross-validation is a statistical method used to study and compare model algorithms by dividing data into two parts: one for training the model and the other for testing. There are two purposes of cross-validation. First, to measure the performance of the training model. The second one, to compare two or more different performances and find out the best model algorithm. In k-fold cross-validation, the dataset is randomly split into k mutually exclusive subsets of approximately equal size. In stratified cross-validation, the folds are stratified so that the fold contains approximately the same proportions of labels as the original dataset [17].

Some calculations that determined the performance of model predictions in text classification were: accuracy, precision, sensitivity, and F1 score. Classification accuracy can be evaluated by calculating the number of sample classes that are correctly recognized (True Positive or TP), the number of samples that are correctly recognized but not classified in class (True Negative or TN), the sample incorrectly classifying (False Positive or FP) or which is not recognized as a class sample (False Negative or FN).

Figure 2 shows a confusion matrix and there are several common metrics that can be calculated. Common metrics that are calculated from the confusion matrix among others are accuracy, precision, recall,F1 score [18], and AUC [19]. As said in this paper, AUC is proved to be better than precious measure [20]. This paper suggest that AUC should replace accuracy in measuring and comparing classifiers, as AUC is better measure in general [21]. Another result in the paper suggest that, in real-world applications of machine learning, we sgould use learning algorithms to optimize AUC instead of accuracy [22].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}, \tag{8}$$

$$\text{AUC} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{9}$$

3 Methodology

3.1 Data Source

The conducted research is about sentiment analysis of Twitter data. The collected tweets are the ones that concerning infrastructure development in Surabaya. Data sourced is from residents of Surabaya. Then, the data is taken with the keywords on Twitter accounts that

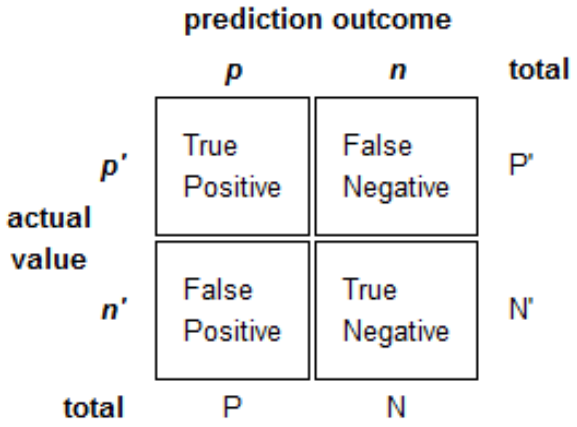

Figure 2: Confusion Matrix For Binary Classification.

move under the City of Surabaya's government. These accounts are @sapawargaSby and @ e100ss. The @sapawargaSby account is one of the official social media accounts owned by the city government, while @e100ss is a Twitter account owned by Radio Surabaya on the 100.00 FM frequency, the most likeable streaming radio of Surabaya citizen.

Tweet data is collected from 29 August to October 2017. The total collected data is 44,684 tweets. However, the raw data underwent a manual sorting process based on the sentiment content of the Surabaya City infrastructure so that data is obtained as much as 1,500 data.

3.2 Research Variable

Data from each account name search was combined into one file. The data structure that was ready to be processed is shown in Table 1.

Table 1: Data Structure

***Tweet* no-**	x_1 **= Term 1**	x_2 **= Term 2**	$\cdots$	x_p **= Term P**
1	the frequency x_1 of tweet 1	the frequency x_2 of tweet 1	$\cdots$	the frequency x_p of tweet 1
2	the frequency x_1 of tweet 2	the frequency x_2 of tweet 2	$\cdots$	the frequency x_p of tweet 2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
n	the frequency x_1 of tweet n	the frequency x_2 of tweet n	$\cdots$	the frequency x_p of tweet n

3.3 Analysis Steps

The steps of analysis that carried out in this research are presented as follows.

Step 1. Crawl tweet data via Twitter API.
Step 2. Doing preprocessing as it is needed to avoid data that is not ready yet.
Step 3. Make a document-term matrix and give TF-IDF weighting.
Step 4. Classification of sentiment data using the SVM method and the NN method.
Step 5. Draw interpretation and conclusions.

4 Results and Discussion

Sentiments in this study basically refer to the contextual polarity of the text or documents that contain emotional effects on the reader to be conveyed by the author about the subject or object related to Surabaya infrastructure development. Sentiments are subjective and depend on personal morals, moral values, and one's beliefs. Sentences or positive words have positive sentiments attached to text or documents. For example, when some texts show happiness, enthusiasm, kindness, etc. Similarly, negative sentences or words have negative sentiments attached to text or documents. For example, when the text shows sadness, hatred, violence, discrimination, etc. If there are no emotions implied, then the text or document is classified as neutral.

Data extracted from the official account of the city administration @sapawargaSby and the official Twitter account Radio Surabaya 100.00 FM named @ e100ss. Tweet data that contains sentiments about the collected infrastructure amounted to 1500 data. Based on this type of sentiment, the data was divided into two types: negative sentiment and positive sentiment. Negative sentiment amounted to 1200 data and positive sentiments amounted to 300 data. The expressed opinions mostly were dominated by negative sentiments as shown as Pie Chart in Figure 3.

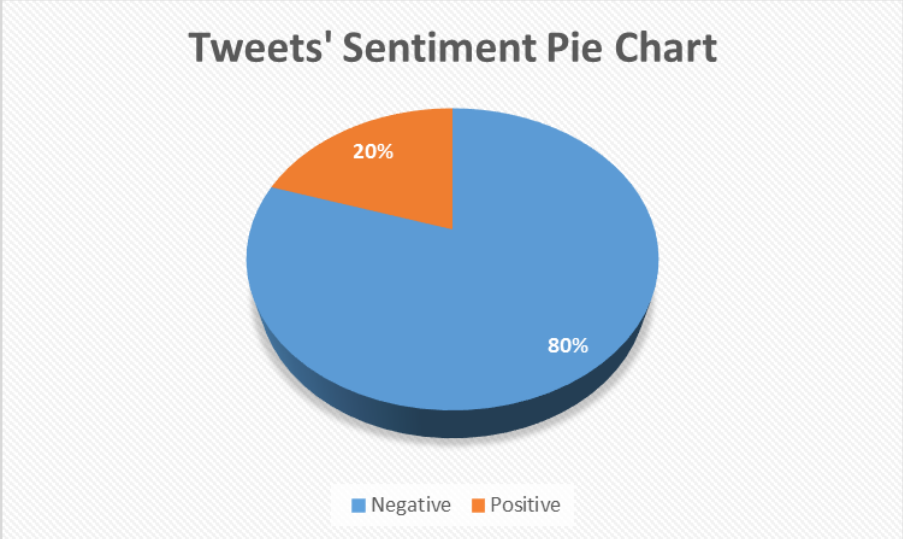

Figure 3: Tweets' Sentiment Pie Chart

The purpose of preprocessing the data was to remove noise, clarifying features, and convert the original data so that the data is processed according to the needs of research. The taken preprocessing steps in this research are tabulated in Table 2.

Table 2: Preprocessing Stages

Preprocessing Steps	Annotation
Selecting the data according to the infrastructure topic.	Deleting data that the topic is not about infrastructure and does not contain sentiments.
Remove “RT”.	Delete the word ”RT” which stands for retweet.
Remove username.	Delete username.
Remove punctuation.	Remove punctuation and symbols, such as ‘˜!@#$%ˆ&*()_+-={}[]|:;”’<>,.
Case folding.	A process which identified the sequence of characters as non-uppercase
Remove stopwords.	Delete the word in the tweet text contained in the stopwords.
Remove number.	Delete a digit number.
Remove whitespace.	Delete any section of a document that unused or space around the subject.
Tokenizing.	A process of chopping sequence word into pieces.

The words of the term collected on 1500 tweets that have been pre-processed amounted to 2824 words. Broadly speaking, the frequency of data after pre-processing was shown in Figure 4.

According to Figure 4, the highest word frequency was owned by the word “listrik” <electricity>. It can be said that the word ”listrik ”<electricity>amounts to 241 words in 1500 datasets tweets. The next highest frequency position is followed by the word “tol” <toll>amounts to 229 words and the word “padam” <extinguished>as many as 181 words. Therefore, the main concern topics of Surabaya citizen who share their opinion in Twitter mentioned to @e100ss and @sapawargaSby were electricity and toll as highway roads.

The step after preprocessing was to create a document-term matrix (DTM). DTM is a two-dimensional matrix $n \times p$ size with document observations as rows and terms as columns while the elements were the number of terms in a document observation. The majority of 99.7% elements the DTM were zero. Therefore, this sparse was then reduced using a computational technique named Remove Sparse Term. The formation of the DTM matrix in the tweet data results in 110 terms with 99% sparsity. Data obtained from DTM is then weighted with term frequency-inverse document frequency, which is often referred to as TF-IDF. The TF-IDF value increases proportionally according to the number of times a word appears on the document and is balanced by the frequency of words in the corpus with the following results in Table 3.

The data was known as a nonlinear data by its nature. Nonlinear cases can be overcome by the presence of Sigma parameter owned by Radial Basis Function (RBF) Kernel. The Grid Search technique was used to determine the value of Cost parameter and the sigma parameter. Grid Search is the process of scanning data to configure the optimal parameters for a given model.

Based on Figure 5, it appears that the optimal parameters obtained based on the grid search technique were 0.01 and 100 for C and Sigma respectively. The sigma value itself substituted

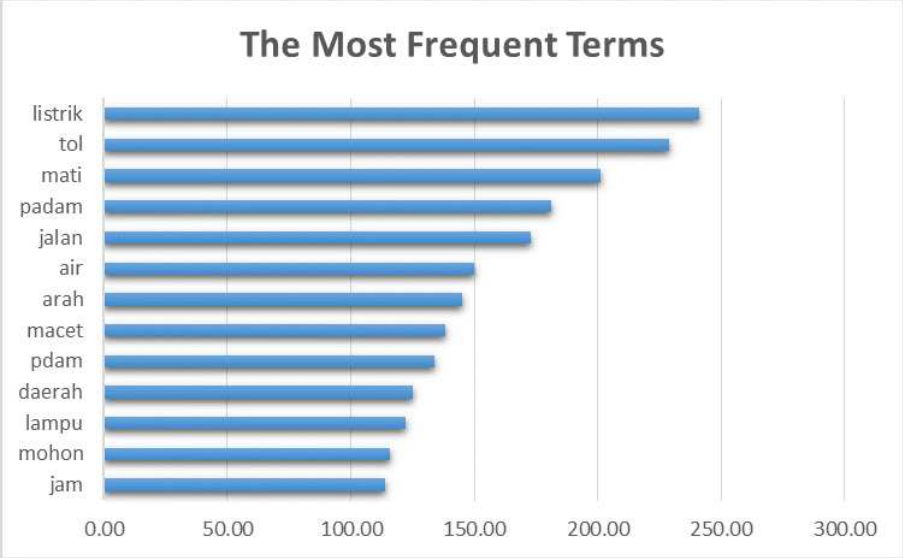

Figure 4: Barchat of Preprocessed Data

Table 3: Document-Term Matrix with TF-IDF Weighting

Terms / Docs	"Listrik" <electricity>	"Padam" <die out>	...	"Hati" <heart>
1	1.3463	1.5334	...	0
⋮	⋮	⋮	⋱	⋮
14	0	0	...	0
15	1.3463	0	...	0
⋮	⋮	⋮	⋱	⋮
1500	0	0	...	1.5067

in the RBF Kernel equation to get kernel function. Then, the kernel function is used to form get the hyperplane function by substituting kernel functions. Thus, the hyperplane function of the training data obtained is as follows.

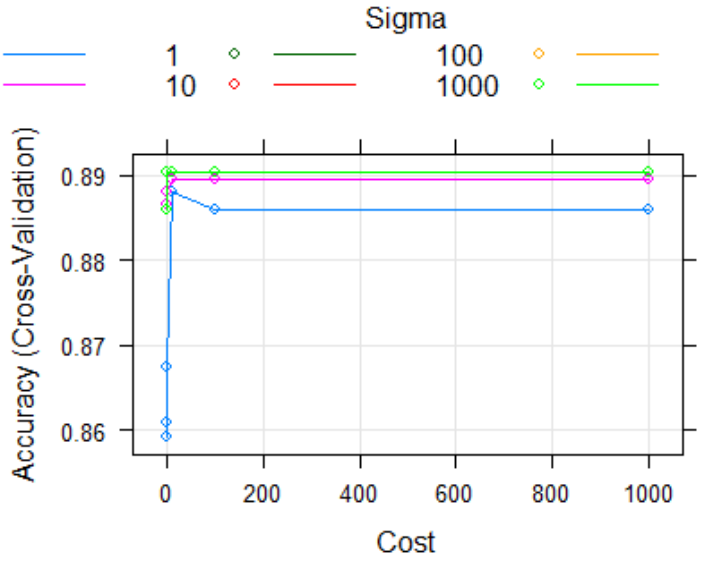

Figure 5: Determination of the parameter C and Sigma using Grid Search Technique

As a description, α_i is Lagrange multiplier from support vector size (854 $\times$1) or it can also be called a vector coefficient. The variables are the class label of the sentiment content marker whose value is -1 for negative sentiment and +1 for positive sentiment. The variables are the term input value and the values are the bias value. Non-linear SVM method with RBF Kernel uses parameter C was 0.01 and the sigma was 100 issues an accuracy value of 71.60%. Precision value means that the level of accuracy between the information requested by the user and the results provided by the system is exactly 84.75%. Recall value of 80.71% is the accuracy of the system to classify data correctly. The F1 score is a measurement of the accuracy of the dataset calculated by taking the harmonic average of precision and recall was amounted to 82.68%. Lastly, one that focus to avoid false classification called AUC was amounted to 52.53%.

Back propagation of NN algorithm was used as a comparison method. The network consists of 110 inputs, 1 hidden layer, and 2 output layers that were negative and positive class. In the hidden layer, trial-error was conducted to obtain the optimal number of neurons. The model was evaluated by 10-fold cross-validation. Accuracy value that obtained from the model were used to determine the optimal model of neural network. The obtained optimal model with one hidden layer and 71 neurons provided an accuracy of 80.00%. While the value of precision, recall, F1 scores, and AUC were 99.58%, 80.02%, 88.84%, and 65.10% respectively.

The final step after getting the performance results of SVM and NN is comparing the classification accuracy. The compared accuracy of the classification is shown in Table 4. The performance of the two classification methods shows that the backpropagation NN with 71 neurons in one hidden layer gave better performance than SVM with kernel RBF.

Table 4: Accuracy Comparison of SVM and NN

Method	Accuracy	Precision	Recall	F1 Score	AUC
SVM	71.60%	84.75%	80.71%	82.68%	52.23%
NN	80.00%	99.58%	80. 02%	88.84%	65.10%

5 Conclusion

Based on the results and discussion earlier, the citizen of Surabaya generally tend to express negative sentiments on Twitter. The main concern topics were about electricity and toll as highway roads. In this research, the data were classified using the SVM and neural network results. Based on the analysis that has been done, the neural network and SVM were successfully applied to classify the data tweet. Between the two model employed, the most appropriate classifier model is a NN. This model is able to give classification accuracy and AUC higher than SVM. It can be concluded that the NN method has a greater performance than SVM.

Acknowledgment

The authors would like to thank the Directorate for Research and Community Service Ministry of Research, Technology and Higher Education Indonesia for supporting this research under the guidance of PDUPT (Penelitian Dasar Unggulan Perguruan Tinggi) research grant according to a contract number: 907/PKS/ITS/2018 on February 1st, 2018.

References

[1] Yulistiani, N. K. *Muresbang Surabaya 2017 Fokus Percepatan Pembangunan Infrastruktur Berwawasan Lingkungan: Draft out in April*. Bangsa Online. 2017.

[2] He, W., Wu, H., Yan, G., Akula, V., and Shen, J. A novel social media competitive analytics framework with sentiment benchmarks. *Information and Management*. 2015. 52(7): 801-812.

[3] Auria, L., and Moro, R. *Support Vector Machine (SVM) as a Technique for Solvency Analysis*. Berlin: German Institute for Economic Research. 2008.

[4] Fithriasari, K., Iriawan, N., Ulama, B., and Sutikno. On the multivariate time series rainfall modeling using time delay neural network. *International Journal of Applied Mathematics and Statistics*. 2013. 44(14): 193-201.

[5] Bataineh, M., and Marler, T. Neural network for regression problems with reduced training sets. *The Official Journal of the International Neural Network Society*. 2017. 95: 1-9.

[6] Mandal, S., and Banerjee, I. Cancer classification using neural network. *International Journal of Emerging Engineering Research and Technology*. 2015. 172-178.

[7] Habibi, R. Analisis sentimen pada Twitter mahasiswa menggunakan metode backpropagation. *INFORMATIKA*. 2016. 12(1).

[8] Kwartler, T. *Text Mining in Practice with R.* New Jersey: John Wiley & Sons Ltd. 2017.

[9] Liu, B. *Sentiment Analysis Mining Opinions, Sentiments, and Emotions.* United State of America: Cambridge University Press. 2015.

[10] Putranti, N. D., and Winarko, E. Analisis sentimen twitter untuk teks berbahasa Indonesia dengan maximum entropy dan support vector machine. *IJCCS*, 2014. 91-100.

[11] Manning, C. D., and Raghavan, P. *Introduction to Information Retrieval.* Stuttgart: Cambridge University Press. 2008.

[12] Sarkar, D. *Text Analytics With Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data.* Bangalore, Karnataka: Apress. 2016.

[13] Rahman, Md. H., Chowdhury S., and Bashar, M. A. An Automatic face detection and gender classification from color images using support vector machine. *Journal of Engineering Trends in Computing and Information Sciences.* 2013. 4(1): 5-11.

[14] Lameski, P., Zdravevski, E., Mingov, R., and Kulakov, A. SVM parameter tuning with grid search and its impact on reduction of model over-fitting. In Y. Yao *et al.* (Eds). *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing.* Switzerland: Springer. 2015. 464-474.

[15] Ramasundaram, S., and Victor, S.P. Text categorization by backpropagation network. *International Journal of Computer Applications.* 2010. 8(6):1-5.

[16] Widhianingsih, T. D. *Text Mining Application for Automatization of Articles Classification on Female Online Magazine Using Naive Bayes Classifier (NBC) and Artificial Neural Network (ANN).* Ph.D. Thesis. Institut Teknologi Sepuluh Nopember. 2016.

[17] Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* 1995. 2(14): 1137-1143.

[18] Fawcett, Tom. ROC graphs: notes and practical considerations for researches. *Machine Learning.* 2004. 31(1): 1-38.

[19] Sokolova, M., and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing and Management.* 2009. 45(4): 427-473.

[20] Ferri, C., Flach, P., and Orallo, J.H. Learning decision trees using the area under the ROC curve. In *Proceedings of the 19th International Conference Machine Learning.* 2002. (2): 139-146.

[21] Huang, J., and Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. In *Proceedings of the IEEE Transactions on Knowledge and Data Engineering.* 2005. 17(3): 299-310.

[22] Ling, C. X., and Zhang, H. Toward Bayesian classifiers with accurate probabilities. In *Proceedings of the Sixth Pacific-Asia Conference Knowledge Discovery and Data Mining.* 2002. 123-134.