Correlation Between Twitter Sentiment Analysis with Three Kernels Using Algorithm Support Vector Machine (SVM) Governor Candidate Electability Level

Dionisia Bhisetya Rarasati[1, a)] and Josef Cristian Adi Putra[2, b)]

[1, 2]*Faculty of Technology and Design, Universitas Bunda Mulia, Tangerang, Indonesia, 1543.*

[a)] Corresponding author: drarasati@bundamulia.ac.id
[b)] josefcristian57@gmail.com

Abstract. According to the expert, Twitter is a microblogging site that provides facilities for users to send a text message with a maximum length of 280 characters, and the text that we type on Twitter called 'tweet' we can write down what happened in various places followed by the author's emotions. One of the information obtained from Twitter is related to politics. Through the sentiment analysis, it can find emotions and correlation between the results of the statement analysis electability statistic of the candidate Paris for governor and deputy governor using sentiment analysis. The purpose of this paper is to get the best accuracy result by comparing the three kernels (Gaussian RBF Kernel, Linear Kernel, and Polynomial Kernel) using SVM and looking for the relationship between the sentiment analysis result and electability from multiple survey institutions using Pearson Product Moment Correlation Technique by using data from Twitter. Data from Twitter has been categorized into five emotions (love, happiness, sadness, fear, and anger). When three kernels are compared for accuracy, the Polynomial Kernel comes out on top with a score of 78.51%. The positive correlation for Basuki Tjahaja Purnama - Djarot Saiful Hidayat is -0.81, whereas the positive correlation for Anies Rasyid Baswedan - Sandiaga Salahuddin Uno is 0.98.

INTRODUCTION

Social media is a web-based online media that are currently familiar to society because, through social media, people can communicate with each other online and in real-time. Several social media are popular in society, namely Twitter, Blog, Facebook, and Instagram. Social media is the media in which users can easily participate to share and create the message, including blogs, social networks, online encyclopaedias/wiki, virtual forums, including virtual worlds (with 3D avatars/characters) [1].

Twitter is one of the most widely used social networking platforms and also can allow users to send and receive information in the form of text, where the maximum of the text is 280 characters. The text that we write on Twitter is called a 'Tweet'. Through the 'tweet' we can write down what is happening right now either in the environment around us or the things that are happening in various places or regions. The tweet that we type is of course followed by the author or user's emotions.

Recognition of the emotions contained in the tweet can be analysed using sentiment analysis. Through sentiment analysis, it can be used to find out what happens to the public or to find information about emotions that are currently happening. One of the information obtained from Twitter is related to politics. Recently, news about the election of Governor and Deputy Governor is being discussed in Indonesia, because the public response is very enthusiastic about matters related to politics, so the author can take the opportunity to see the responses of netizens or Twitter users related to the political issues. One of the algorithms that can be used to analyse sentiment is the Support Vector Machine (SVM). The Support Vector Machine (SVM) is a prediction technology that can be used in both classification and regression.

Regarding the data, the data used by the author is taken from Twitter related to the 2017 Governor Election which will be managed with the SVM Algorithm to get the result of sentiment analysis, then it can be seen its relationship with the electability data from several institutions related to the 2017 Governor Election by using correlation is an analysis of one the association or relationship measurement techniques. One of the correlations is the Pearson Product Moment Correlation, for this correlation technique we can find assumptions of normality and linearity with metric data or at least intervals. The Pearson Product Moment correlation used in this study is divided into 2 parts, namely Positive Correlation, and Negative Correlation, which correlates with the positive correlation namely the relationship between the Twitter data taken and electability of each pair, while for the Negative Correlation there is a relationship between the Twitter data taken and the electability data of each candidate pair.

Through the explanation above, the author intends to investigate the relationship between Twitter sentiment analysis and the Support Vector Machine (SVM) Algorithm with the electability level of the Governor and Deputy Governor Candidates in the 2017 DKI Regional Head Election by comparing the three kernels used by the Linear Kernel, Kernel Gaussian RBF and Kernel Polynomial.

LITERATURE REVIEW

In this section, the author will discuss the understanding of material related to Twitter data sentiment analysis by comparing three kernels and the Support Vector Machine (SVM) Algorithm.

Block Diagram

Below is a block diagram of stages of the research to be carried out: or stages of the research to be carried out:

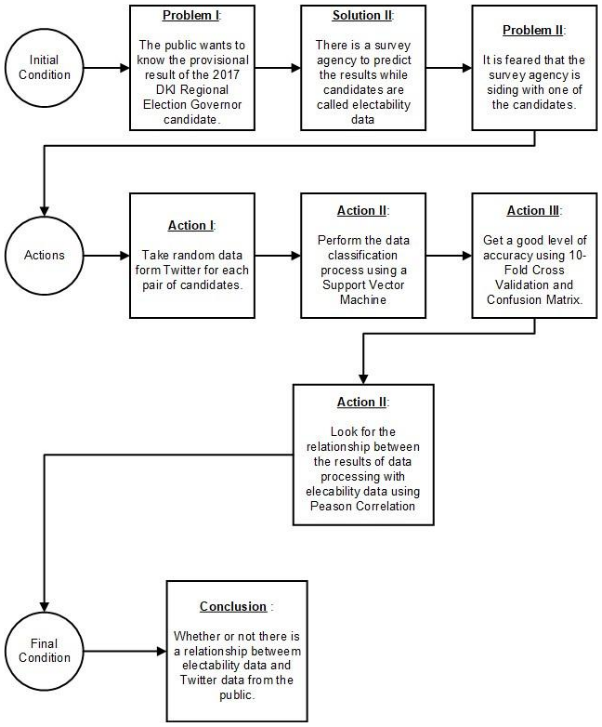

Initial Condition
Problem I: The public wants to know the provisional result of the 2017 DKI Regional Election Governor candidate.
Solution II: There is a survey agency to predict the results while candidates are called electability data
Problem II: It is feared that the survey agency is siding with one of the candidates.
Actions
Action I: Take random data form Twitter for each pair of candidates.
Action II: Perform the data classification process using a Support Vector Machine
Action III: Get a good level of accuracy using 10-Fold Cross Validation and Confusion Matrix.
Action II: Look for the relationship between the results of data processing with elecability data using Peason Correlation
Final Condition
Conclusion : Whether or not there is a relationship betweem electability data and Twitter data from the public.

FIGURE 1. Block Diagram

In the diagram above, various actions will be taken by researchers related to the problems that have been described. Actions include performing the data classification process using the SVM (Support Vector Machine) algorithm, which is followed by calculations using 10-Fold Cross and Confusion Matrix to get a result where the accuracy level is even higher. By using these two algorithms, we can find out that the actual data that exist can determine or can help us to determine the electability of the candidate pair of the Governor of DKI Jakarta and from the data taken from the Twitter platform, it can be concluded that the electability data has continuity with the data on Twitter. SVM is a prediction technique that can be used for both regression and classification [2]. The SVM technique was used to find the best hyperplane function for separating observations with different target variable values [3]. There are 3 types of techniques in the SVM algorithm, here are the examples:

Linear Kernel: Support Vector Machine (SVM) which can be obtained by the equation below [4]:

$$[(w^t . x_i) + b] \geq 1 \; for \; y_i = +1$$

$$[(w^t . x_i) + b \leq -1 \; for \; y_1 = -1 \qquad (1)$$

Description:

x_i = training dataset

i = 1, 2, …, n

y_i = label from x_i

Kernel Gaussian RBF:

$$K(x, x') = exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right) \qquad (2)$$

Description:

$\|x - x'\|^2$: Euclidean Distance

σ : an independent parameter set to determine the decay rate of $k(x, y)$ towards zero.

Kernel Polynomial:

$$K(xi, xj) = \left((x_{i,} . x_j) + c\right)^d \qquad (3)$$

Description:

$x_{i,} . x_j$: pair of two training data

$c, d > 0$: constant

Sentiment Analysis

Sentiment Analysis, also known as Opinion Mining is the process of automatically understanding, extracting, and processing textual data to obtain sentiment information contained in an opinion sentence. Sentiment analysis is used to determine someone's opinions or opinion tendencies toward a problem or object, such as whether they have a negative or positive opinion [5]. The positive emotion vocabulary group consist of two basic emotions, namely the emotions of love and pleasure. The negative emotion vocabulary group consist of three basic emotions namely anger, fear, and sadness [6].

Text Pre Processing

Text mining is defined as the mining of data in the form of text, where the data source is typically obtained from documents and the foal is to find words that can represent the contents of the document so that an analysis of the connectivity between documents can be performed [7]. In-text mining is accomplished through text preprocessing, which is a process applied to text data to produce numerical data. The stages in this research process are stop words removal/filtering, tokenizing, word merging based on synonyms, and stemming.

Future Extraction

Future extraction is a stage that gives an index to each document. The following information is related to feature extraction namely word weighting and Z-Score normalization.

SVM Algorithm

Boser, Guyon, and Vapnik invented the Support Vector Machine (SVM). The term was first used at the Annual Workshop on Computational Learning Theory in 1992 [8]. SVM is a prediction technique that can be used for both regression and classification [9]. The SVM technique is used to find the best hyperplane function for separating observations with different target variable values. SVM is predicting technique that can be used for both classification and regression [10].

K-Fold Cross-Validation

The Cross-Validation method will avoid overlapping the testing data. There are two stages related to K-Fold Cross Validation [11]. The following is how K-Fold Cross Validation works:

1. The total instance is divided into N sections
2. The first fold is when the first part becomes test data (testing data) and the rest becomes training data. Next, compute the accuracy based on the data segment. The following formula is used to calculate the accuracy:

$$Accuracy = \frac{\sum \quad classification\ correct\ test\ data}{\sum \quad total\ test\ data}\ x\ 100\%$$

3. The second fold occurs when the second portion becomes test data (testing data) and the remainder becomes training data. Then, based on that portion of data, compute the accuracy.
4. Continue in this way until you reach the K-Fold. Calculate the average accuracy of the 'K' pieces above. This average accuracy becomes the final accuracy.

Confusion Matrix

Confusion Matrix contains actual and predictable information and the data in the matrix can be used to evaluate system performance. The confusion matrix for the two classes is shown in the table below [12]:

TABLE 1. Confusion Matrix

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

Accuracy is the number of correct predictions, which is determined by the equation (5):

$$AD = \frac{a+d}{a+b+c+d} \qquad (5)$$

Pearson Correlation

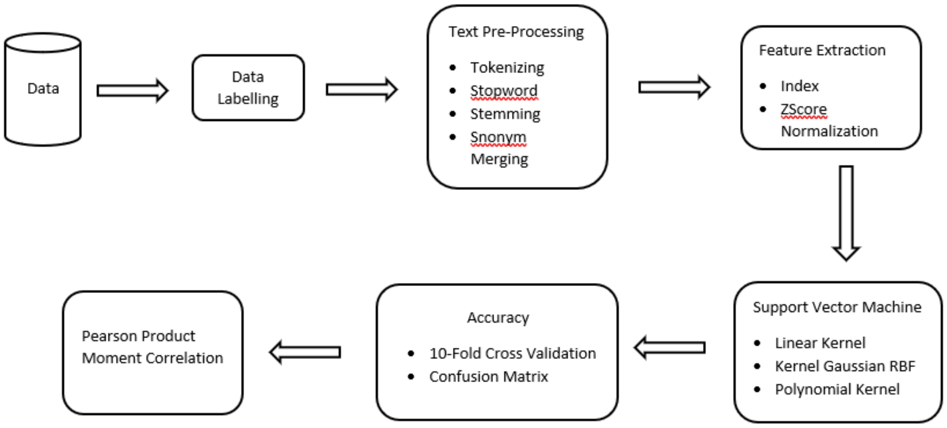
Correlation can be used to assess the strength of a relationship between two variables (sometimes more than two variables) with a certain scale, for the case of using the Pearson Correlation on an interval or ratio scale, with a distance of 0 to 1. It can be said that the correlation is in the same direction if the correlation coefficient value is found to be negative [13]. To test the validity, one of them is using the Pearson Product Moment Correlation with the formula below [14]:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2 - \Sigma X)^2\}\{N\Sigma Y^2 - (\Sigma Y)^{2\}}}}$$

Description:

r_{xy}	: Item correlation coefficient
$\sum X$	: Total score of each item
$\sum Y$	: Item total score
$\sum X^2$	: The sum of the X scores squared
$\sum XY$	: The sum of the multiplications of X and Y
N	: Number of samples

RESEARCH METHODOLOGY



Data on Twitter

The data used comes from tweets in Indonesian, which are divided into two emotions (positive and negative). The data used is Twitter data taken from two keywords, namely related to the names of the Governor Candidates ad Deputy Governor candidates in the second round of the 2017 DKI Jakarta Governor Election. The first search category is the governor and deputy governor candidate "Basuki Tjahaja Purnama – Djarot Saiful Hidayat" with a total of 1.165 pieces of data obtained and the second search category is the governor and deputy governor candidate "Anies Rasyid Baswedan – Sandiaga Salahuddin Uno" with a total the data obtained are 1,138 data.

Text Preprocessing

In the pre-processing stage, the system performs tokenizing, removing stop word, and stemming stages. The system also performs some special treatment on the data used because tweets contains a lot of noise. The system will remove URL links, usernames, retweet marks, and various other noises. The system will change non-standard words or shortened words into standards words. The system will also take words to take words that begin with a hash mark.

Feature Extraction

Feature Extraction also aims to reduce the amount of data with the following steps below:

- Weighting Stage
  At the weighting stage, the system will represent the tweet as a vector with the weight value of each term. Calculation of term weights using the tf-df weighting method.
- Z-Score Normalization Stage
  In the normalization stage, the method used is Z-Score Normalization. The normalized term weight value is only the dominant term weight. Dominant weight is a weight that is worth more than a certain threshold.

Support Vector Machine

At this stage, the system will group tweets into two clusters namely positive and negative. Each tweet will be grouped by a hyperplane. There are three kernels used to find the best hyperplane namely the Gaussian RBF Kernel, Linear Kernel, and Polynomial Kernel. Therefore, from the three kernels, it can be seen that the best kernel is based on the higher accuracy, so that later it will be used the Pearson Product Moment Correlation stage as the Result of Sentiment Analysis.

Calculation Accuracy

To use the best number of folds for validity testing, it is recommended to use 10-Fold Cross-Validation in the model. As a result, the author is in the stage of accuracy using 10-Fold Cross-Validation. The following is how 10-Fold Cross-Validation works:

1. Total instance divided into 10 parts.
2. The first fold occurs when the first part of the data becomes test data and the remainder becomes training data. Then, based on that portion of the data, compute the accuracy.
3. The second fold occurs when the second part becomes test data and the remainder becomes training data. Then, based on that portion of data, compute the accuracy.
4. And so on until the tenth fold is reached. Calculate the average accuracy of the ten accuracy pieces listed above. This average accuracy is used to calculate the final accuracy.

The second method the author used is the Confusion Matrix, with the calculation of accuracy (AD) where accuracy is calculated by adding the number of the correct predictions for which the sample is negative to the number of correct predictions for which the sample is positive and dividing by the total number of predictions. So, it is expected that the two accuracy calculation methods have the same level of accuracy.

Pearson Product Moment Correlation

The Pearson Product Moment Correlation is used to analyze the validity test, wherein this study the Pearson Correlation is used to compare the result of Twitter sentiment analysis with the electability data of the Governor and Deputy Governor Candidates in the second round.

The Pearson Correlation used by the author is made up of two sentiments, namely positive and negative which can be seen in the form of a graph for each pair of Governor Candidates. So, through the Pearson Correlation, it can be seen that there is a link between the result of sentiment and electability data and it can be concluded that tweets form the public related to the election of the Governor and Deputy Governor are by the electability data sources from several institutions in Indonesia. The electability data obtained is in the form of a percentage (%).

The data from the result of the sentiment on Twitter can be compared with the data of electability, the data from the result of the sentiments of Twitter is changed in percentages. For example, the total of data is 1000, consisting of 600 positive sentiment divided by total data number multiplied by 100% so that the percentages of each sentiment result are obtained.

RESULT AND DISCUSSION

All of the data that successfully carried out the pre-processing process resulted in a total of 592 words where 2.000 data were used for tweet document data input. The classification process used is the Support Vector Machine Algorithm with 3 kernels and the accuracy calculation uses 10-Fold Cross-Validation. After all, kernels have displayed accurate results, below is a description of the accuracy result that has been obtained.

TABLE 2. Accuracy Result Table

Kernel	Accuracy
Linear	85.87%
Gaussian RBF	90.58%
Polynomial	78.51%

Based on the result of the program the candidate for governor Basuki Tjahaja Purnama – Djarot Saiful Hidayat the author summarizes into a table form consisting of the result of Positive Sentiment, Negative Sentiment, and the average monthly electability data.

TABLE 3. Sentiment and Electability Result Basuki T.P. - Djarot S.H.

Month	Negative Sentiment	Positive Sentiment	Data Electability
January	62.14	37.86	31.24

February	68.63	31.37	35.2
March	75.52	28.48	43.7
April	70.24	29.76	46.12

Below is a summary of the result of the Pearson Correlation of Basuki Tjahaja Purnama – Djarot Saiful Hidayat.

TABLE 4. Correlation Pearson Results Basuki T.P. - Djarot S.H.

Correlation Pearson Results	
Positive Sentiment Correlation	-0.811
Negative Sentiment Correlation	0.811

Based on the results of the program the candidate for governor Anies Rasyid Baswedan – Sandiaga Salahuddin Uno the author summarizes into a table form consisting of the result of Positive Sentiment, Negative Sentiment, and the average monthly electability data.

TABLE 5. Sentiment and Electability Result Anies R.B. - Sandiaga S.U.

Month	**Negative Sentiment**	**Positive Sentiment**	**Data Electability**
January	67.88	32.12	24.6
February	57.61	42.39	38.1
March	52.45	47.55	48.8
April	53.93	46.07	48.35

Below is a summary of the results of the Pearson Correlation of Anies Rasyid Baswedan – Sandiaga Salahuddin Uno.

TABLE 6. Correlation Pearson Result Anies R.B. - Sandiaga S.U.

Correlation Pearson Results	
Positive Sentiment Correlation	0.987
Negative Sentiment Correlation	-0.987

SUMMARY

The conclusions are:

1. The best kernel used in this study is the Gaussian RBF Kernel with an accuracy of 90.58% while the lowest accuracy is the Polynomial Kernel with an accuracy of 78.51%
2. In the candidate pair for governor Basuki Tjahaja Purnama – Djarot Saiful Hidayat, the Correlation of Positive Sentiment Correlation Result is -0.811 and Negative Sentiment is 0.811 meaning the conclusion obtained is a very strong correlation results are negative (-0.811) then in this candidate pair the higher the positive sentiment, the lower the electability.
3. In the candidate pair for governor Anies Rasyid Baswedan – Sandiaga Salahuddin Uno, the Correlation of Positive Sentiment Correlation Result is 0.987 and Negative Sentiment is -0.987 meaning the conclusion obtained is very strong correlation and has a unidirectional relationship because the positive sentiment correlation results are positive (-0.987) then in this candidate pair the higher positive sentiment, the higher the electability.

REFERENCES

1. Mayfield A. What is social media? image: weather project bw 01 by [Internet]. Icrossing. 2008. 1–36 p. Available from:

http://www.icrossing.co.uk/fileadmin/uploads/eBooks/What_is_Social_Media_iCrossing_ebook.pdf

2. Fachrurazi. Penerapan Pembelajaran Berbasis Masalah Untuk Meningkatkan Kemampuan Berpikir Kritis Dan Komunikasi Matematis Siswa Sekolah Dasar. J Penelit Pendidik UPI [Internet]. 2011;Edisi Khus(1):76–89. Available from: http://jurnal.upi.edu/penelitian-pendidikan/view/637/
3. Hearst M, Dumais ST, Osman E, Platt J, Scholkopf B. Support Vector Machines. IEEE Intell Syst Their Appl. 1998;13(4):18–28.
4. Cortes C, Vapnik V. Support-Vector Network. Mach Learn. 1995;20(3):273–97.
5. Rozi I, Pramono S, Dahlan E. Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi. J EECCIS. 2012;6(1):37–43.
6. Shaver P., Murdaya U, Fraley R. Structure of The Indonesian Emoticon Lexicon. Asian J Soc Psychol. 2001;4(3):201–24.
7. Ch. MH. Text Mining. 2006.
8. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. Proc Fifth Annu ACM Work Comput Learn Theory. 1992;(August):144–52.
9. Fachrurrazi S, Burhanuddin. Penggunaan Metode Support Vector Machine Untuk Mengklasifikasi Dan Memprediksi Angkutan Udara Jenis Penerbangan Domestik dan Penerbangan Internasional di Banda Aceh. J Sist Inf. 2018;2(2):1–10.
10. Santosa B. Data mining:Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu-Bisnis.Edisi Pertama. Data miningTeknik Pemanfaat Data untuk Keperluan Bisnis Yogyakarta Graha Ilmu-BisnisEdisi Pertama. 2007;33(4):365–73.
11. Murfi H. Evaluasi.
12. Joa Y. Intermedia Issue Agenda and Attribute Agenda Setting in Online. 2017;(August).
13. Sarwono J. Prosedur-Prosedur Populer Statistik Untuk Mempermudah Riset Skripsi. Angew Chemie Int Ed 6(11), 951–952. :1–19.
14. Arikunto S. Prosedur Penelitian Suatu Pendekatan Praktik. Rev. Jakarta: Rineka Cipta; 2010.