ASTES

Sentiment Analysis of Regional Head Candidate's Electability from the National Mass Media Perspective Using the Text Mining Algorithm

Ahmad Zainul Hamdi[1], Ahmad Hanif Asyhar[*,2], Yuniar Farida[2], Nurissaidah Ulinnuha[2], Dian Candra Rini Novitasari[2], Ahmad Zaenal Arifin[3]

[1]*Religious Studies Department, UIN Sunan Ampel Surabaya, Indonesia*

[2]*Mathematics Department, UIN Sunan Ampel Surabaya, Indonesia*

[3]*Mathematics Department, Universitas PGRI Ronggolawe, Indonesia*

A R T I C L E I N F O	A B S T R A C T
	Mass media plays an important role in leading public opinion, including in the election of regional head candidates. The tendency of mass media coverage can be used as a parameter to measure the strength of each regional head candidate. To analyze the tendency of media opinion, sentiment analysis is needed. In this study, text mining techniques were used to analyze opinion sentiments of a regional head election in East Java from the national media perspective. The researcher used the Support Vector Machine algorithm to build a sentiment analysis model. News documents about candidates for the regional head in East Java 2018 were taken from national mass media samples, namely JPNN, Kompas, Detik and Republika. From the test results, the model built on Khoffifah's data as a candidate for regional head number one has a value of precision, recall and AUC of 0.927, 0.931 and 0.902, respectively. Furthermore, the model built on Gus Ipul's data as a candidate for regional head number two has a value of precision, recall, and AUC of 0.940, 0.948 and 0.890 respectively. The models built on both data shows good performance with accurate estimation results. Based on the data obtained, the national media tends to have alignments to the regional head candidate number two namely Gus Ipul.

1. Introduction

The development of the internet in this globalization era is growing rapidly, the internet plays an important role in various scientific disciplines, such as science, psychology, sociology, marketing, communication, and politics too [1]. Human needs on the internet and information are always developing, causing rapid technological advancements. Most people use the internet to access media, both online and social media. Online mass media such as jpnn.com, kompas.com, detik.com, republika.co.id, and others are online media that continue to grow along with the rapid growth of people who are aware of information flows.

Recognized or not, opinions from national or local media can also lead to public opinion. Opinion in online media or social media plays a very large role in influencing habits or behavior towards the election of the head of state, regional head, or people's representatives both at the central and regional levels [2]–[5]. Online media is very accessible compared to print media because now everything is in your hands.

In 2018 is a political year in several regions including East Java province. In 2018, it will also be the end of the term of office of the Governor of East Java, so that regional head election needs to be held for the period 2018-2023. The East Java regional head election is also in the spotlight for the people of Indonesia in addition to the DKI Jakarta regional head election and also West Java and Central Java regional head elections because the election in these regions become political barometer in this country. The two Governor candidates are Khofifah and Gus Ipul.

The DKI Jakarta regional head election left an impression on how the votes for the pair Ahok and Djarot had to drop dramatically only due to religious sentiment and the issue of

*Corresponding Author: Ahmad Hanif Asyhar, UIN Sunan Ampel Surabaya, Indonesia, +62 82233250005, hanif@uinsby.ac.id

money politics on a few days before the voting. Here the role of media greatly influences candidates. The DKI Jakarta regional head election that has just finished raises an important note, namely the strong pulling of public opinion through the media. On national online media news in the pre-implementation and implementation of the DKI Jakarta regional head election, there were various opinions and responses with positive and negative sentiments. The problems that arise when analyzing all the results of sentiment and the classification of social media news or national online media manually, it takes a lot of time and effort in its implementation. From these problems, an opinion classification system is needed in the form of sentiment text on national mass media to predict the strength of the pair governor and deputy candidate early quickly and accurately [6]. Media news that available online can be used as a data source to predict the results of political elections. The influence of media opinion that has a major impact on public opinion, as happened in the case of elections in DKI Jakarta, needs to be considered in order to make predictions on the electability of governor elections.

To analyze media opinion, sentiment analysis is needed [7]. Sentiment analysis is a branch of learning science in the text mining domain that studies analysis of the opinion, sentiment, emotion, attitude, an evaluation which is poured into textual form [7]. The use of sentiment analysis techniques is often used for product reviews, reputation management, analysis of a topic and so on[2]. Research that analyzes mathematically or computationally about online media sentiments about the election is rarely found. This is because the collection of news data through online media requires extra hard work compared to twitter data collection that can be easily accessed via the Twitter API.

There are two main problems in this study. The first problem is because of the lack of negative data on media data, there will be an imbalance class. The problem obstacle is that there are so many features on news data that feature selection will be needed before going into sentiment analysis. In this research, by considering this problem, the classification of news sentiment in online mass media will be applied in politics, in a case study of regional head elections in East Java in 2018.

2. Related Works

In recent years, sentiment classification has become a principle problem in sentiment analysis research. Sentiment classification aims to determine the polarity of sentiment documents. The sentiment classification mostly focuses on English. Moreover, research focuses on film reviews, products, and others. In Mullen's research and also the research conducted by Wilson and the team [8], the supervised learning method was applied in the classification of film reviews. In the research conducted by Cui and the team [9], as well as the research conducted by Dave and the team [10], the sentiment classification was applied to product reviews. Sentiment research on Indonesian-language documents, especially in news article documents, was not found. In addition, sentiment analysis of documents is a problem because of the many ways opinions are developed in one document. News articles present a bigger challenge because news writers usually avoid overt attitude

indicators. However, despite real neutrality, news articles can still bear polarity if they describe events that are objectively positive or negative.

Today, sentiment analysis of the textual data is one of the interesting topics. Many researchers are working on the automated techniques of extraction and analysis of a huge amount of user-generated data, which is available in social media and online mass media. In [11], the authors proposed a way to get the pre-labeled data from Twitter which can be used to train an SVM classifier. They used the Twitter hashtags to judge the polarity of the tweet. To analyze the accuracy of the proposed technique, a test study on the classifier was conducted which showed the result with the accuracy of 85%. Support Vector Machine (SVM) method resulted in a relatively higher accuracy compared to other machine learning methods but was greatly influenced by the number of data sets, training data, testing data and the number of positive data and the negative. In this study, it was explained that the use of the SVM method successfully classifies documents. This is indicated by the high level of accuracy of the method used. Similarly, Priyavrat's research concluded that the Support Vector Machine (SVM) method also has a better level of accuracy compared to other machine learning methods, namely Naive Bayes, Decision Tree and Neural Network [12]. Based on some of the studies that have already been conducted, in this sentiment analysis study, the classification method that will be used is SVM. This is because the SVM method proved to be more resilient to noisy data that is relatively high compared to other methods.

3. Proposed System

In this study, sentiment classification will be carried out with text mining processing. The construction of the sentiment model uses the machine learning method, Support Vector Machine (SVM). There are two additional processes that aim to improve the performance of the SVM algorithm. The first process is to select features with the information gain method. Feature selection is very important because the news text domain has several tens of thousands of features. The second process is boosting to deal with class imbalances. The classification of data with unbalanced class divisions can cause a significant decrease in performance. The method used in the second process is adaptive boosting. The best SVM model is not only tested through kernel selection but also involves feature selection and boosting. This aims to determine whether feature selection and boosting can improve the performance of models in sentiment classification.

3.1. Text Mining

Text mining is mining data in the form of text where data sources are usually obtained from documents. The purpose of text mining is to get useful information from a set of documents. The specific tasks of text mining include categorizing text and grouping text (text clustering). Based on the irregularity of the text data structure, the process of text mining requires several initial stages which in essence is to prepare so that the text can be changed to be more structured. Text mining can be considered as a two-stage process that begins with the application of structures to text data sources and continues with the extraction of relevant

information and knowledge from structured text data using the same techniques and tools like data mining. Therefore, it is not surprising that text mining and data mining will be at the same architectural level. The stages of text mining are shown in Figure 1.
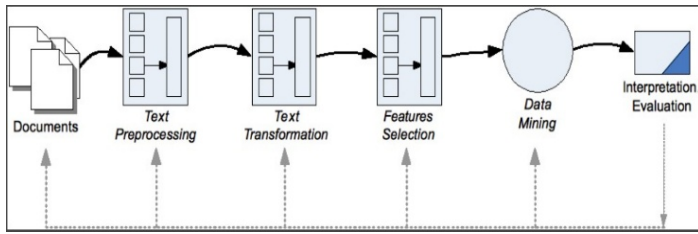

Figure 1: Text mining stages [13]

The stages are (1) Document collection (2) Preparation (3) Text Transformation (4) Feature Selection (5) Mining Data Process (6) Interpretation/Evaluation [13]. In the text transformation stage, several studies have applied a combination of TF and IDF combinations, namely multiplying local and global weights (TF-IDF) [14], [15]. The TF-IDF method combines two concepts, namely the frequency of occurrence of a word in a document and the inverse frequency of the document containing the word. The IDF formula is shown by equation 1, while the TF-IDF formula is indicated by Equation 2.

$$idf = \log\left(\frac{N}{df}\right) \quad (1)$$

$$w(t,d) = tf(t,d) * idf \quad (2)$$

where $tf(t, d)$ is the appearance of the word t in document d, N is the number of documents in the collection of document, and df is the number of documents containing term t.

3.2. Feature Selection

In general, sentiment analysis becomes more difficult when the volume of text increases, due to the complex relationship between words and phrases. This is mostly because the text domain has several tens of thousands of features. Feature selection can make classification more effective by reducing the amount of data analyzed, as well as identifying features that are suitable for consideration in the process of making learning algorithms so that they can run in a faster and more efficient way.

Information gain is often superior to other feature selection methods [16]. Information gain measures how much information is present and the absence of a role word to make correct classification decisions in any class. Information gain is one filter approach that is successful in classifying text.

3.3. Support Vector Machine

Classification is the process of finding a model from a set of data [17], [18]. The purpose of classification is to make a decision by predicting a case based on the results of the classification obtained [19]. In the classification process, there are two processes carried out, namely the training process and the testing process. The classification method used is Support Vector Machine classification. The advantage of using SVM compared to other classification algorithms such as K-Means or Naive

Bayes is that SVM finds decision boundaries that maximize the distance from the closest data points of all classes. SVM not only finds the limits of decisions but also finds the most optimal decision limits.
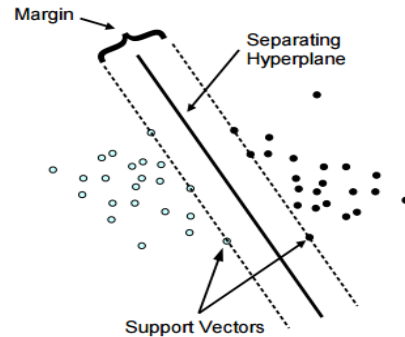

Figure 2: Illustration SVM[20]

In the illustration, there are several data points that touch the highest margin line. These data points are referred to as support vectors. The data used in the SVM method is denoted by $x_i \in R^d$ and the label is denoted by $y_i \in \{1,2\}$ to $i = 1,2,...,l$ where l is the amount of data.

There are several kernel functions [21] that will be tested in this study, they are:

(1) The function of Linear Kernel

The linear Kernel function is defined as:

$$k(x_i, x_i) = x_i^T x_i \quad (3)$$

(2) The function of the Polynomial Kernel

Polynomial Kernel functions with degrees of d, where r and d are parameters defined as follows:

$$k(x_i, x_j) = (\gamma x_i^T x_j + r)^a, \gamma > 0 \quad (4)$$

(3) Kernel Radial base Function (RBF) function

The function of the RBF Kernel is also called the Kernel-Gaussian function. The RBF Kernel function is defined as

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (5)$$

where γ is a positive parameter to set distance.

(4) The function of the Sigmoid Kernel

The function of the sigmoid kernel is defined as follows:

$$k(x_i, x_j) = \tanh(\gamma x_i, x_j + r) \quad (6)$$

where $\tanh(a) = 2\sigma(a) - 1$ and $\sigma(a) = \dfrac{1}{1 + \exp(a)}$

3.4. Boosting Algorithm

Boosting algorithm is an iterative algorithm that gives different weights to the distribution of training data in each

iteration. Each boosting iteration adds weight to examples of classification errors and decreases the weight in the correct classification example, so it can change the distribution of training data effectively [22]. The Boost algorithm (adaptive boosting) is an algorithm that builds strong classifiers by combining a number of simple (weak) classifiers. The first model was built according to the classification method used. Each instance in the training dataset is weighted and the weight is updated based on the accuracy of the entire model and whether an instance is correctly classified or not. Subsequent models are trained and added until minimum accuracy is reached or no further improvement is possible. Each model is given a weight based on its skills and this weight is used when combining predictions from all models in the new data.

3.5. Analysis Method

There are two analyzes used in analyzing sentiments results. The first is the accuracy of model analysis, and the second is an analysis of patterns of media trends based on news topics.

3.5.1. Evaluation method

The accuracy of the model analysis can be seen from the value of precision and recall resulting from the confusion matrix. The more precision and recall values are close to 100, it indicates the more accurate this model in analyzing news sentiments, and vice versa. To find out the strength of the model in overcoming imbalance data, Area Under Curve (AUC) measurements were carried out. If the AUC value closer to 1, indicates a good model performance in classifying data [23] and vice versa. This study will evaluate the performance of SVM model through the accuracy based on precision, recall and AUC by doing: 1) testing four SVM kernels namely linear, polynomial, RBF and sigmoid, 2) setting the number of features used in classification through the feature selection process and 3) adding the boosting process and vice versa into the construction of the model.

3.5.2. News Topic Analysis

Analysis of news topics will be shown by graphs compiled of word patterns in the news. The more words appear together in one document, the stronger the pattern of relationships will be. Based on the graph, media trends can be analyzed against two pairs East Java regional head candidates, namely Khofifah–Emil and Gus Ipul–Puti.

4. Result and Discussion

4.1. Data Presentation

The news data relating to the candidates for governor of East Java 2018 is taken from the period of January 1, 2018, to June 26, 2018. To obtain these data, carried out by accessing directly the data on news portals available online with national media samples, namely JPNN, Kompas, Detik and Republika. To find data relating to candidate governor with number one, Khofifah Indar Parawansa, the keyword "khalifah" was used. From the process of collecting data in the four national media, there were 998 data obtained.

Title	Content	Polarity	Tanggal	Resource
Kakek Detektif, Penebar	Akun Twitter @KakekDetektif menyita perhatian publik lan	NEGATIF	25/06/2018	detik.com
Ini Cagub Jatim yang Kuasai	Calon Gubernur Saifullah Yusuf dan Cawagub Puti Guntur S	NEGATIF	25/06/2018	detik.com
Debat Pilgub Jatim Terakhir	Sekjen DPP PDI Perjuangan (PDIP) Hasto Kristiyanto menga	NEGATIF	23/06/2018	detik.com
Hadiri Doa Bersama Guyub	Cagub Cawagub Jatim Khofifah Indar Parawansa-Emil Dard	POSITIF	26/06/2018	detik.com
Pilkada Rasa Pilpres, AHY: P	Partai Demokrat (PD) menargetkan perolehan suara melar	POSITIF	26/06/2018	detik.com

Figure 3: Sample data for Khofifah

To find data relating to candidate governor number two, Saifullah Yusuf, the keyword "gus ipul" was used. From the data collection process in the four national media, 1082 data were obtained.

Title	Content	Polarity	Tanggal	Resource
Debat Langsung Panas, Khofifah	Debat terakhir cagub Jatim berlangsung cukup seru. E	NEGATIF	23/06/2018	detik.com
Poltracking: Khofifah-Emil Unggu	Poltracking Indonesia melakukan survei elektabilitas p	NEGATIF	23/06/2018	detik.com
Survei RTK: Khofifah-Emil Unggul	Roda Tiga Konsultan merilis hasil survei terhadap elek	NEGATIF	22/06/2018	detik.com
Survei SMRC: Khofifah-Emil Ungg	Saiful Mujani Research and Consulting (SMRC) melaku	NEGATIF	22/06/2018	detik.com
Hadiri Doa Bersama Guyub Ruku	Cagub Cawagub Jatim Khofifah Indar Parawansa-Emil	POSITIF	26/06/2018	detik.com

Figure 4: Sample data for Saifullah Yusuf

4.2. Data Processing

The data obtained is labeled according to its tendency, whether it is positive sentiment or negative sentiment. Not all data was taken to construct the sentiment classification model. This is because the number of documents with positive sentiments is much greater than the number of documents provided with negative sentiments with a ratio of 52:1. Therefore document retrieval from positive sentiments is randomly around 30% of the positive sentiment document database with a ratio of 14:1 to the number of documents negative sentiment. Table 3 shows the ratio of the overall data and sample data to Khofifah and Gus Ipul database documents. It can be noted that all negative news population data are all sampled to build a classification model.

Table 1: Sentiment ratio of overall data and sample data

Database	Positive Population	Negative Population	Positive samples	Negative Samples
khofifah	949	49	344	49
gus ipul	1059	23	320	23

The next step is pre-processing text, at this stage carried out of cleansing, folding case, stopword removal. After that, the tokenization and weighting of the tokenization data were carried out to partition the documents in the form of sentences. This is done so that the words in the news document can be given weights for each word. The word weighting method used is TF-IDF which is a development of the TF method. After TF shooting, the weight of each word is also multiplied by the IDF or the number of occurrences of the word in the relevant document. The less often a word appears in a document, the greater IDF value generated.

After the word weighting is carried out, then the data is analyzed. To analyze data, two different scenarios were carried out, namely SVM without combination (pure SVM) and SVM method combined with the selection of information gain feature. The comparative process is carried out on the value of precision and recall generated when not using the feature selection method, and when using the feature selection method. Total features tested

with feature selection in the system included 500, 1000, 2000 and 3000 features where the total features of Khofifah's data were 4764 features, and Gus Ipul's data were 4381 features

4.3. Implementation of Text Mining Algorithm

Comparison of SVM performance as an algorithm used in text mining with four different kernels namely linear, polynomial, RBF and sigmoid as shown in table 6. Based on the results of sentiment classification, it can be observed that in Khofifah's data, the best SVM model is in a linear kernel using feature selection with as many as 500 features The model also has the fastest execution time among other kernel models. Whereas in Gus Ipul's data, the best SVM model is in the RBF kernel that uses feature selection with 500 features. Even though it has good performance, the model has the slowest execution time among other kernel models. Even though in the polynomial kernel, feature selection does not affect SVM performance, both in Khofifah's data and Gus Ipul's data, the best performance is obtained when boosting and selection features with as many as 500 features. This shows that boosting and feature selection can improve the SVM method classification performance.

Table 2: Comparison of SVM performance with four different kernels

Can-didate	Type of Boosting	Kernel	N attri-butes	Presisi	*Recall*	AUC	Time (s)
Khofifah	with *boosting*	Linear	500	0,927	0,931	0,902	0,58
	with *boosting*	Polino-mial	4764	0,893	0,878	0,505	1,73
	with *boosting*	RBF	500	0,930	0,934	0,899	3,12
	with *boosting*	Sigmoid	500	500	0,863	0,885	1,65
Gus Ipul	with *boosting*	Linear	500	0,931	0,942	0,793	0,71
	-	Polino-mial	-	0,870	0,933	0,500	0,33
	with *boosting*	RBF	500	0,940	0,948	0,890	3,43
	with *boosting*	Sigmoid	3000	0,940	0,936	0,865	1,07

4.4. Evaluation of the Text Mining Model

Model evaluation was carried out by looking at the value of precision, recall, and AUC. The more value of precision, recall, and AUC is close to 1, the model is getting closer to the best performance. The best performance of the model in Khofifah's data achieved a precision value of 0.927, recall of 0.931 and AUC of 0.902 using SVM linear kernel. Whereas in Gus Ipul's data, the best performance of the model reached a precision value of 0.940, recall of 0.948 and AUC of 0.890 using the RBF kernel. Precision, recall and AUC scores above 0.8 in both the Khofifah and Gus

Ipul data indicates that the SVM model that is enhanced by feature selection and boosting can work well in classifying news sentiment data. From collecting news documents to building SVM models, various problems were found that made this research work not optimally, such as data imbalance, inappropriate polarity sentiment reading, and document sentiment not reflected in all sentences.

4.5. Results of National media Trends

Relationships that exist in the data of two candidates for the regional head in the four major national media in Indonesia namely JPNN, Kompas, Detik and Republika are visualized in figure 5. The names of the pair Gus Ipul – Puti are more frequently mentioned that the names of the pair Khofifah – Emil. It means that Gus Ipul is more dominant in the four online media. In addition, the name "Sukarno" which is the grandfather of Puti is also often referred to. This is one of the things that can strengthen the electability of the pair Gus Ipul and Puti. In addition, the names "megawati" and "pdip" are also often referred to as the main supporters of the pair Gus Ipul and Puti. There is also the name "risma" which when called is almost always related to Gus Ipul and Puti. This shows that Risma was one of the leaders who supported Gus Ipul and Puti.
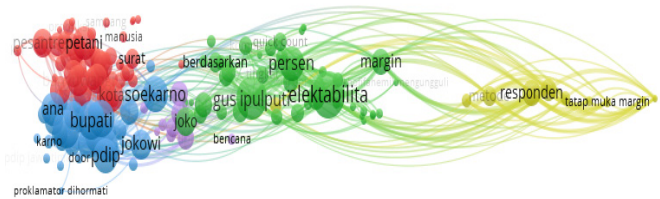

Figure 5: The results of the Khofifah-Gus Ipul data visualization on all four media

Even so, in the majority of the news in the four online media, it was said that according to the survey institute, the electability of the pair Khofifah and Emil Dardak was still superior to Gus Ipul and Puti. In addition, in some news also stated that Khofifah and Emil had a target to gain a lot of votes from undecided voters. This is also supported by the fact that the pair Khofifah and Emil are considered as representations across generations x and y because of Khofifah's experience and the youth spirit brought by Emil. This is certainly one of the hallmarks of Khofifah and Emil in gaining the voice of undecided voters, the majority of which are from the younger generation. It was also stated that the way Khofifah-Emil campaign directly touched the lowest layer of voters. This makes the relationship and political identification between Khofifah-Emil and voters direct.

4.6. Discussion

Based on the news gathered from four major national media in Indonesia, namely in JPNN, Kompas, Detik and Republika, the number of positive news sentiments of Gus Ipul is more than the number of positive sentiments of Khofifah with a ratio of 53 %: 47%. Negative news sentiment on Gus Ipul tends to be less than

Khofifah with a ratio of 32 %: 68%. From all the data obtained, it can be concluded that the number two governor candidate, Gus Ipul is superior to the national online media sentiment version rather than the number one candidate governor, Khofifah. This conclusion can be obtained based on the fact that: (1) candidate number two governor gets more attention from national media coverage with more evidence of data that mentions governor candidate number two, (2) percentage of news with positive sentiment towards Gus Ipul – Puti is more than Khofifah – Emil and the percentage of negative sentiment towards Gus Ipul – Puti is less than Khofifah – Emil.

5. Conclusion

Text mining method is a method that can be used to classify text documents, especially the classification of news document sentiments. To obtain an optimal architecture model and an accurate estimation, an experiment is performed. Model evaluation was carried out by looking at the value of precision, recall, and AUC. The best performance of the model in Khofifah's data reached a precision value of 0.927, recall of 0.931 and AUC of 0.902 using a linear kernel. Whereas in the Gus Ipul's data, the best performance of the model reached a precision value of 0.940, recall of 0.948 and AUC of 0.890 using the RBF kernel. The values of precision, recall, and AUC obtained in both Gus Ipul and Khofifah's data are above 0.8, which indicates that the model has a good performance in classifying news sentiments. Based on the data obtained and processed from the four media, both JPNN, Kompas, Detik and Republika, the media more often mentioned the pair Gus Ipul and Puti than the Khofifah-Emil pair.

Acknowledgment

References

[1]   T. Peng, L. Zhang, and Z. Zhong, "Mapping the landscape of Internet Studies: Text mining of social science journal articles 2000–2009," 2013.

[2]   M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4241–4251, 2013.

[3]   X. Bai, "Predicting consumer sentiments from online text," *Decis. Support Syst.*, vol. 50, no. 4, pp. 732–742, 2011.

[4]   M. Eirinaki, S. Pisal, and J. Singh, "Journal of Computer and System Sciences Feature-based opinion mining and ranking," *J. Comput. Syst. Sci.*, vol. 78, no. 4, pp. 1175–1184, 2012.

[5]   H. Savigny, "Public Opinion, Political Communication and the Internet," *Politics*, vol. 22, no. 1, pp. 1–8, 2002.

[6]   A. R. T. Lestari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.

[7]   B. Liu, *Sentiment Sentiment Analysis Analysis and and Opinion Opinion Mining Mining*. 2012.

[8]   P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis."

[9]   M. Cui, hang., Mittal, Vibhu., Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," *J. Comput. Sccience*, pp. 1265–1270, 2006.

[10] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. 12th Int. Conf. World Wide Web (WWW '03)*, pp. 519–528, 2003.

[11] W. A. Zgheib and A. M. Barbar, "A Study using Support Vector Machines to Classify the Sentiments of Tweets."

[12] A. J. Singh, "Sentiment Analysis : A Comparative Study of Supervised Machine Learning Algorithms Using Rapid miner," vol. 5, no. Xi, pp. 80–89, 2017.

[13] B. Susanto, *Text Mining*. Surabaya: Graha Ilmu, 2015.

[14] R. Intan, A. Defeng, J. T. Informatika, F. T. Industri, U. Kristen, and P. Surabaya, "Hard : Subject-Based Search Engine Menggunakan Tf-Idf Dan Jaccard ' S Coefficient," pp. 61–72.

[15] A. F. Hidayatullah *et al.*, "Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada twitter," vol. 2014, no. semnasIF, pp. 115–122, 2014.

[16] A. K. Uysal, "An improved Global feature Selection Scheme for Text Classification.," *Expert Syst. With Appl*, vol. 43, pp. 82–92, 2016.

[17] Suyanto, *Data mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.

[18] D. K. S. Putra, *Media dan Politik*. Surabaya: Graha Ilmu, 2012.

[19] R. Primartha, *Belajar Machine Learning Teori dan Praktik*. Bandung: Informatika, 2018.

[20] Muthu krishnan, "Understanding Support vector Machines using Python," 2018.

[21] C. Hsu and C. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," vol. 13, no. 2, pp. 415–425, 2002.

[22] P. Harrington, *Machine Learning in Action*. New York: Manning Publication Co, 2012.

[23] F. Shabani, L. Kumar, and M. Ahmadi, "Assessing Accuracy Methods of Species Distribution Models: AUC, Specificity, Sensitivity and the True Skill Statistic," *Glob. J. Hum. Soc. Sci.*, vol. 18, no. 1, 2018.