Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Setiment Analysis of Public Opinion on The Go-Jek Indonesia Through Twitter Using Algorithm Support Vector Machine

Setiment Analysis of Public Opinion on The Go-Jek Indonesia Through Twitter Using Algorithm Support Vector Machine

H Syahputra[1*], L K Basyar[2], A A S Tamba[3]

[1,2,3]Department of Mathematics, FMIPA Universitas Negeri Medan

* hsyahputra@unimed.ac.id

Abstract. The development of technology and information, especially in Indonesia is very rapid so that social media is the most popular communication tool by the people of Indonesia today. One of these social media is Twitter. This also causes the public to tend to give opinions and assessments in the form of tweets to service companies, one of which is Go-Jek Indonesia. Public opinion and judgment on Twitter can be classified into 3 classes: negative, neutral, and positive. The purpose of this study is to analyze the sentiment of public opinion on Go-jek Indonesia on twitter using the Support Vector Machine (SVM) algorithm. The approach used were Multiclass One Vs Rest SVM with Univariate Chi Square feature selection to classify community tweets on Go-Jek Indonesia's services. Using testing data of 170 tweets, 31.2% of people with negative opinions were obtained, 24.1% were neutral and 36.5% were positive opinions and 5.9% failed to be classified. The results of sentiment analysis testing conducted provide a classification accuracy of 91.8%.

I. Introduction

In this modern era the demand for community mobility is higher and of course requires transportation facilities that can provide movement from one place to another quickly, even though the distance is far. Today there is a recent breakthrough, namely the innovation of transportation based on online applications that are supported by communication technology via smartphones. This online application-based transportation is a merger in terms of motorcycle taxi transportation services and communication technology. Of the many online application-based transportation available in Indonesia, Go-Jek is the most widely used by the public. In addition to transportation services, Go-Jek also provides other services such as Go-Food, and Go-Clean [2].

The development of technology and information has also caused social media to become the most popular communication tool, so that now the public tends to provide opinions, criticisms, and suggestions through social networking media and one of them is Twitter. Twitter according to statistics is the fastest growing social network since 2006. This social network, which is limited to 140 characters, sends 250 million tweets every day. According to the MIT Technology Review (2013), Indonesia is the third largest contributor to tweet contributing 1 billion, below the United States (3.7 billion) and Japan (1.8 billion) [10].

Tweets which are the status text of Twitter media account users in general can contain information about the user's identity, conversation, and feelings of the user. For example, with Go-Jek Indonesia's services, it is through this tweet that the public can express their opinions or assessments of the services provided by Go-Jek. Therefore, it can be used the application of machine learning methods, namely text mining to classify the polarity of the opinion. Text mining aims to extract useful information from data sources through the identification and exploration of interesting patterns in documents. The polarity found is a pattern in textual data that is not structured in a document [3]. One analysis in text mining is data sentiment analysis. Sentiment analysis data or also called opinion mining, is a field of study to analyze opinions, sentiments, evaluations, assessments, human attitudes, and emotions towards entities such as products, services, organizations, individuals, problems, events, topics, and their attributes [7].

Research on sentiment analysis has been done before. Some machine learning techniques that can be used include Naive Bayes Classifier, Decision Trees, and Support Vector Machine. Research on sentiment analysis using the dataset from Twitter was conducted by Nugroho [11]. In his research, he analyzes the sentiment of public opinion on the services provided by online motorcycle taxi services namely Go-Jek and Grab Indonesia using the Naive Bayes Classifier algorithm. The next research that became the author's reference in compiling this research is the research conducted by Athoillah [1]. In this study discuss the classification using the Support Vector Machine (SVM) algorithm to classify the image of the object of two-wheeled transportation with four or more wheeled transportation. Other research conducted by Vijayari [18] explained that the SVM algorithm has a better level of classification accuracy than the Naive Bayes algorithm.

In this study tweet sentiment analysis will be performed using SVM algorithm related to public opinion on the services provided by Go-Jek Indonesia on social media.

2. Basic of Theory

2.1. Data Mining

Data mining is a process of extracting or finding information from very large data. The process of extracting information is done by using computer learning techniques (machine learning) to find a pattern or information desired from the data [5]. Data mining aims to utilize data in a database by processing it and producing more useful information [12].

2.2. Text Mining

Text Mining can be interpreted broadly as a process where users interact with documents to find information that users want from these documents using analysis tools in data mining. The specific task of text mining is to group text and text categorization [3].

2.3. Sentiment Analysis

Sentiment Analysis is a combination of data mining and text mining, or a method used to process various opinions given by consumers or experts through various media, regarding a product, service or an agency. Sentiment analysis consists of 3 types of opinions, namely positive opinions, negative opinions and neutral opinions, so that the company or related agency sentiment analysis can find out the public response to a service or product, through public feedback or even experts. Sentiment refers to the focus of a particular topic, statements on a topic may be different meanings with the same statement on different subjects, therefore in some studies, especially on product reviews, preceded by determining the elements of a product being discussed before starting the Sentiment process Analysis [15].

2.4. Twitter

Twitter is one social media that allows users to send messages that are limited to 140 characters, known as tweets. Twitter users discuss many different topics in their tweets. The topic can be in the form of responses to an event, product, figure, political campaign, and others [13].

2.5. Text Pre-Processing

Pre-processing is the initial stage of text mining to change data in accordance with the required format. This process is carried out to explore, process and manage information and to analyze the textual relationship of structured and unstructured data.

1) Case Folding.

Case folding is the initial stage in Pre-processing which aims to change every word form to be the same. This is done by changing words into lowercase letters.

2) Data Cleaning.

Data Cleaning is the process or cleaning of words by eliminating comma delimiter (,), period (.), And other punctuation.

3) Language Normalization.

At the Pre-processing stage, language normalization is performed on nonstandard words. This stage aims to restore the form of writing of each word in accordance with the Big Indonesian Dictionary (KBBI).

4) Stopword Removal.

Stopword is a list of common words that have no significance and are not used. In this process common words will be deleted to reduce the number of words stored by the system.

5) Stemming.

Stemming is a process to find the stem (basic word) from the word stopword removal (filtering). There are two rules for doing stemming, namely the dictionary approach and the rule approach.

6) Tokenisation.

Tokenisation is the process of cutting a document into small pieces which can be in the form of chapters, sub-chapters, paragraphs, sentences, and words (tokens) [8].

2.6 TF-IDF weighting

This method is the process of calculating the weight of the number of words contained in the text in accordance with existing features. This method is called Term Frequency (TF). TF is the frequency of occurrence of a word in the document concerned. The greater the number of occurrences of a word (high TF) in the document, the greater the weight.

In the TF stage, weighting of tweet data is as follows,

$$tf_j(d) = w_j \qquad (1)$$

w_j represents the number of occurrences of the word t_j in the document d.

Followed by the Inverse Document Frequency (IDF) method. IDF (Inverse Document Frequency) is a calculation of how words are widely distributed in the document concerned. The IDF equation is written as follows:

$$idf_j = log\frac{N}{df_j} \qquad (2)$$

Where N is the number of all documents in the collection while df_j is the number of documents containing the word (t_j). So the TF-IDF equation can be written as follows:

$$w_{ij} = tf_{ij} \times idf_j \qquad (3)$$

$$w_{ij} = tf_{ij} \times log\frac{N}{df_j} \qquad (4)$$

where w_{ij} is the word weight or term (t_j) of the document (d_i), tf_{ij} is the number of occurrences of words or terms (t_j) in the document (d_i), N is the sum of all documents in the database, and df_j is the number of documents containing term (t_j) [9].

2.7 Scaling Feature

Scaling features is one of the important things before the data is processed with the support vector machine algorithm. The purpose of scaling features is to avoid features that have a large range value that is more dominant than features that have a smaller range value. Besides scaling features can also be used to avoid numerical difficulties during the calculation process [6].

The method used in scaling features is the Min-Max Scaler method. Min-Max Scaler is a method that scales features that depend on the maximum weight of the feature and the minimum weight of the feature of a numeric vector. Following the Min-Max Scaler method:

$$X_j' = \frac{X_j - X_{min}}{X_{max} - X_{min}} \qquad (5)$$

With X_j' the jth feature weight that has been scaled by the Min-Max Scaler method, X_j is the jth feature weight that has not been scaled, X_{max} is the weight of the feature that has the largest weight value, and X_{min} is the weight of the feature has the smallest weight value [13].

2.8 Feature Selection

Univariate Chi Square is a method of feature selection using a statistical approach. The Chi Square Univariate method uses Chi Square in selecting features. The Univariate method looks for a feature that has an influence with the numeric data that has been generated. The following models are given:

$$X^2(t,c) = \frac{N(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \qquad (6)$$

where, A is the number of tweets in class c that contain the word t, B is the number of tweets that are not in class c but contains the word t, C is the number of tweets in class c but does not contain the word t, D is the number of tweets that are not in class c and does not contain the word t, and N is the total number of documents [16].

The feature will be selected based on the results made in the above equation. The greater the value generated will oppose the value of H0, in other words the feature in question has a close relationship with many tweets. In the Univariate method, we will look for the number of features with the highest Chi Square value [13].

2.9 Support Vector Machine

Support Vector Machine (SVM) was developed by Boser, Guyon, and Vapnik in 1992. SVM is a technique for making predictions, both in the case of classification and regression. The concept of SVM can be explained simply as an attempt to find the best hyperplane that functions as a separator of two data sets from two different classes. The best hyperplane separator between the two classes can be found by measuring the margin of the hyperplane and finding its maximum point. Margin is the distance between the hyperplane and the closest data from each class. The closest data is referred to as a support vector. SVM can classify separate data linearly (linearly separable) and non-linear (nonlinearly separable).

Linearly separable data is data that can be separated linearly. The training data is declared by $D = \{(x_i, y_i)\}_{i=1}^{n}$ and $x_i = \{x_1, x_2, \dots, x_n\}$ are attribute (feature) sets for class-i training data. Whereas the class label from x_i data is denoted by $y_i \in \{-1, +1\}$. According to Zaki [19] the SVM linear classification hyperplane is defined as follows:

$$h(x) = w^T x_i + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b \qquad (7)$$

To get the best hyperplane is to find a hyperplane located in the middle between two areas of class boundaries and to get the best hyperplane, the same as maximizing the margin or distance between two sets of objects from different classes [14].

The search for the best separator field with the largest margin value can be formulated into a constraint optimization problem, namely:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \qquad (8)$$

$$s.t.\ y_i(w^T \cdot x_i + b) \geq 1, \forall x_i \in D \text{ [4].}$$

The condition for a function to be a kernel function is to fulfill Mercer's theorem which states that the resulting kernel matrix must be positive semi-definite. According to Prasetyo [12] the commonly used kernel functions are as follows:

1. Linear Kernel
   $K(x_i, x) = x_i^T x$
2. Polynomial Kernel
   $K(x_i, x) = \left(x_i^T x + r\right)^p$
3. Gaussian Radial Basic Function (RBF) Kernel
   $K(x_i, x) = exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right)$
4. Sigmoid Kernel
   $K(x_i, x) = \tanh\left(x_i^T x + r\right)$

x_i and x are training data pairs. The parameters $\sigma, r, p > 0$ are constants.

There are two options for implementing an SVM multiclass by combining several binary SVMs or combining all data consisting of several classes into a form of optimization problem [14]. However, in the second approach optimization problems that must be solved are far more complicated. One method commonly used to implement SVM multiclass is the "one against all" approach.

2.10 Model Evaluation

To determine the accuracy of the classificator model in predicting new data that is not included in the training data, an evaluation of the model will be carried out. K-Fold Cross Validation is used to calculate the accuracy of the classifier function model for new data.

To present the results of the K-Fold Cross Validation the following confusion matrix is used:

Table 1. Confusion Matrix

		Prediction		
		-1	0	1
Actual Value	-1	a_{11}	a_{12}	a_{13}
	0	a_{21}	a_{22}	a_{23}
	1	a_{31}	a_{32}	a_{33}

From the confusion matrix, the performance accuracy units of the model can be seen by:

$$Akurasi = \frac{(a_{11} + a_{22} + a_{33})}{(a_{11} + a_{12} + a_{13} + a_{21} + a_{22} + a_{23} + a_{31} + a_{32} + a_{33})}$$

In other words, to find accuracy, it is enough to do the diagonal sum of the confusion matrix, then divide the amount of data used [13].

3. Research Methodology

3.1. *Data collection*

The data to be used is secondary data obtained directly from Twitter social media, in the form of tweets or posts about the comments and opinions of the Indonesian people to the services provided by Go-jek Indonesia. The data taken is tweet or posting in Indonesian on social media twitter.

3.2. *Data processing*

Data processing is carried out through the following process:

1. Pre-processing is carried out namely case folding, data cleaning, language normalization, stopword removal, stemming, and tokenization of all data so that feature extraction can be performed.
2. Feature extraction is done to get the features (term) of each tweet to be used in the classification model. The feature extraction process is weighting with TF-IDF, scaling features with Min-Max Scaler, and feature selection using Univariate Chi Square.
3. Classification.The classification process carried out in this study is to apply the Multiclass Support Vector Machine (SVM) algorithm to group data into three sentiments, namely data with negative, positive or neutral sentiment towards the services provided by Gojek Indonesia.
4. After the classification results are obtained, an evaluation of the model will be carried out to determine the level of accuracy and errors of the system in predicting the testing data sentiment class.

4. Results and Discussion

4.1 *Data Acquisition*

Data is obtained using the Python application with the Twitterscraper library. The amount of data successfully acquired was 1070 tweets which were then divided by 300 training data for each negative, neutral and positive class, and 170 testing data. One example of training data in this study can be seen in Table 2.

Table 2. Examples of Training Data

No	*Tweet*
1	Walaupun @gojekindonesia lagi agak eror di Jogja, (dapet driver jauh jauh terus) mitra gojek tetap memberikan layanan dengan sigap dan mantap. @gojekindonesia
2	Sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah.

4.2. Pre-Processing Data

In the process of text pre-processing which includes Case Folding can be seen in Table 3, Data Cleaning in Table 4, Language Normalization in Table 5, Stopword Removal in Table 6, Stemming in Table 7, and Tokenisation in Table 8.

Table 3. Case Folding

Before *Case Folding*	After *Case Folding*
Walaupun @gojekindonesia lagi agak eror di Jogja, (dapet driver jauh jauh terus) mitra gojek tetap memberikan layanan dengan sigap dan mantap. @gojekindonesia	walaupun @gojekindonesia lagi agak eror di jogja, (dapet driver jauh jauh terus) mitra gojek tetap memberikan layanan dengan sigap dan mantap
Sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah.	sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah.

Table 4. Data Cleaning

Before Data Cleaning	After Data Cleaning
Walaupun @gojekindonesia lagi agak eror di Jogja, (dapet driver jauh jauh terus) mitra gojek tetap memberikan layanan dengan sigap dan mantap. @gojekindonesia	walaupun gojekindonesia lagi agak eror di jogja dapet driver jauh jauh terus mitra gojek tetap memberikan layanan dengan sigap dan mantap
Sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah.	sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah

Table 5. Normalisation of Language

Before Language Normalization	After being normalized
walaupun @gojekindonesia lagi agak eror di Jogja, (dapet driver jauh jauh terus) mitra gojek tetap memberikan layanan dengan sigap dan mantap. @gojekindonesia	walaupun gojekindonesia lagi agak eror di jogja dapat driver jauh jauh terus mitra gojek tetap memberikan layanan dengan sigap dan mantap

sejujurnya ga berani komplain ke gojek soalnya sering gojek ga jaga privasi kita and nanti driver dtg kerumah nyerang kan ga lucu kan yah.	sejujurnya tidak berani komplain ke gojek soalnya sering gojek tidak jaga privasi kita dan nanti driver datang kerumah nyerang kan tidak lucu kan ya

Table 6. Stopword Removal

Before Stopword is deleted	After Stopword is deleted
walaupun @gojekindonesia lagi agak eror di jogja dapat driver jauh jauh terus mitra gojek tetap memberikan layanan dengan sigap dan mantap	gojekindonesia eror jogja driver mitra gojek layanan sigap mantap
sejujurnya tidak berani komplain ke gojek soalnya sering gojek tidak jaga privasi kita dan nanti driver datang kerumah nyerang kan tidak lucu kan ya	sejujurnya berani komplain gojek gojek jaga privasi driver rumah nyerang lucu ya

Table 7. Stemming Process

Before *Stemming*	After *Stemming*
gojekindonesia eror jogja driver mitra gojek layanan sigap mantap	gojekindonesia eror jogja driver mitra gojek layan sigap mantap
sejujurnya berani komplain gojek gojek jaga privasi driver rumah nyerang lucu ya	jujur berani komplain gojek gojek jaga privasi driver rumah serang lucu ya

Table 8. Results of Tokenisation

Before Tokenisation	After Tokenisation
gojekindonesia eror jogja driver mitra gojek layan sigap mantap	gojekindonesia eror jogja driver mitra gojek layan sigap mantap
jujur berani komplain gojek jaga privasi driver rumah serang lucu ya	Jujur berani komplain gojek gojek jaga privasi driver rumah serang

lucu
ya

4.3. Feature Extraction

Following is the TF-IDF weighting process with several documents in Table 9 to be weighted.

Table 9. Token from training data (X_1, X_2, X_3) and testing (X_4)

Document	Token
X_1	gojekindonesia eror jogja driver mitra gojek layan sigap mantap
X_2	jujur berani komplain gojek gojek jaga privasi driver rumah serang lucu ya
X_3	hai gojekindonesia layanan goclean hilang aplikasi golife ya area tangerang pake goclean sangat saying
X_4	driver menit hapus aplikasi gojek hp

In the word "application", the total number of documents (N) = 4, and the frequency of occurrence of the word "application" in all documents (df) = 2.
Calculate IDF:

$$idf_{aplikasi} = \log\left(\frac{4}{2}\right) = \log(2) = 0{,}301$$

Calculate the weight (w) in document X1:

$$w_{X3,aplikasi} = 1 * 0{,}301 = 0{,}301$$

The overall weight of the features (terms) can be seen in Table 10.

Table 10. Weighting of TF-IDF

word	Tf				Df	N/df	Idf	weight (TF-IDF)			
	X_1	X_2	X_3	X_4				X_1	X_2	X_3	X_4
Mitra	1	0	0	0	1	4	0,602	0,602	0	0	0
Gojek	1	2	0	1	3	1,333	0,125	0,125	0,25	0	0,125
Gojekindonesia	1	0	1	0	2	2	0,301	0,301	0	0,301	0
Sigap	1	0	0	0	1	4	0,602	0,602	0	0	0
Mantap	1	0	0	0	1	4	0,602	0,602	0	0	0
Jujur	0	1	0	0	1	4	0,602	0	0,602	0	0
Goclean	0	0	2	0	1	4	0,602	0	0	1,204	0
Pivasi	0	1	0	0	1	4	0,602	0	0,602	0	0
Hilang	0	0	1	0	1	4	0,602	0	0	0,602	0
Layanan	1	0	1	0	2	2	0,301	0,301	0	0,301	0
Driver	1	1	0	1	3	1,333	0,125	0,125	0,125	0	0,125
Complain	0	1	0	0	1	4	0,602	0	0,602	0	0

Serang	0	1	0	0	1	4	0,602	0	0,602	0	0
Pakai	0	0	1	0	1	4	0,602	0	0	0,602	0
Saying	0	0	1	0	1	4	0,602	0	0	0,602	0
Rumah	0	1	0	0	1	4	0,602	0	0,602	0	0
Lucu	0	1	0	0	1	4	0,602	0	0,602	0	0
Berani	0	1	0	0	1	4	0,602	0	0,602	0	0
Jaga	0	1	0	0	1	4	0,602	0	0,602	0	0
Aplikasi	0	0	1	1	2	2	0,301	0	0	0,301	0,301
Ya	0	1	1	0	2	2	0,301	0	0,301	0,301	0
Tanggerang	0	0	1	0	1	4	0,602	0	0	0,602	0
Menit	0	0	0	1	1	4	0,602	0	0	0	0,602
Hapus	0	0	0	1	1	4	0,602	0	0	0	0,602
HP	0	0	0	1	1	4	0,602	0	0	0	0,602

Then a feature weight scaling process is performed to maintain the range of each weight of each term at 0-1, with the aim of avoiding numerical difficulties during the calculation process.

After the scaling stage weights are completed, feature selection is used. The feature selection method that can be used is Univariate Chi Square. In the Univariate method, we will look for the number of features with the highest Chi Square value.

4.4. Classification in Python.

The classification process with the entire training and testing data is done using the help of Python 3.7 programming software and jupyter notebook as a text editor for the program created. In this study a non-linear SVM algorithm using the Gaussian Radial Base kernel. The parameters in SVM are determined by trial and error.

The results of classifications from community tweets to Go-jek Indonesia accounts using the Multiclass One vs Rest Support Vector Machine method with several C parameter values with γ and accuracy generated using 100% features and 70% features (Table 13).

Table 11. Classification Accuracy with 100% Features

C	0,1	0,2	0,3	0,4	0,5
1	81	80	82	77	74
1,5	83	85	85	81,2	78
2	84	86	84,7	81,2	77,6
2,5	84	87	84,7	81,2	77,6
3	84	87	84,7	81,2	77,6
3,5	84	86	84,7	81,2	77,6
4	84	86	84,7	81,2	77,6
4,5	84	86	84,7	81,2	77,6
5	86	87	84,7	81,2	77,6
5,2	86	87	84,7	81,2	77,6
5,5	86	87	84,7	81,2	77,6
6	86	87	84,7	81,2	77,6

In the classification done using 100% of the features obtained the highest accuracy is 0.87% with a parameter value of C of 2.5.

4.5. Confusion Matrix

Confusion Matrix table for classification results with RBF kernel and parameters C = 5.2 and γ = 0.1 with 1977 features (term) can then be made to find out the classification results and errors in the classification of each class.

Table 12. *Confusion Matrix*

Class	-1	0	1
-1	53	2	3
0	3	41	3
1	1	2	62

From Table 12, it can be seen that there are 53 negative tweets (true negative), 41 neutral tweets (true neutral), and 62 positive tweets (true positive) that have predictions according to the label given. There were 4 testing tweets that were labeled negative and neutral which failed to be classified correctly, while for positive testing data there were 6 tweets that failed to be classified correctly.

5. Conclusion

The Support Vector Machine algorithm can be used to classify text sentiments for 3 classes using the Multiclass One Vs Rest method. The highest accuracy generated for Multiclass tweet classification with Support Vector Machine is 91.8% with 1977 training data features. Document classification error is caused because in a class there are words that are the same as other classes and the weight of words in the other categories is greater than the class they should be, so tweets classified are more likely to approach other classes.

6. References

[1] Athoillah, M., Irawan, M.I., dan Imah, E.M., (2015): Support Vector Machine untuk Image Retrieval, Prosiding Seminar Nasional Matematika dan Pendidikan Matematika,.

[2] Damayanti, S. A. S., (2017): Transportasi Berbasis Aplikasi Online: Go-Jek sebagai Sarana Tranportasi Masyarakat Kota Surabayal, Jurnal Sosiologi, .

[3] Feldman, R., dan Sanger, J., (2006): The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, Cambridge.

[4] Gunn, S. R., (1998): Support Vector Machines for Classification and Regression, University of Southampton, Southampton.

[5] Han, J., Kamber, M., dan Pei, J., (2012): Data Mining: Concepts and Techniques, 3, Elsevier, USA.

[6] Hsu, C. W., dan Lin, C. J., (2002): A Comparison of Methods for Multi-class Support Vector Machines, IEEE Transaction on Neural Network, 13(2), 415– 425.

[7] Liu, B., (2012): : Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.

[8] Luqyana, W. A., Cholissodin, I., dan Perdana, R. S., (2018): Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(11), 4704–4713.

[9] Manning, C. D., Raghavan, P., dan Schutze, H., (2008): Introduction to Information Retrieval, Cambridge University Press, Cambridge.

[10] MIT, T., (2013): Language Data Reveals Twitters Global Reach, MIT Technology Review, diakses melalui http://www.technologyreview.com.

[11] Nugroho, D. G., Chrisnanto, Y. H., dan Wahana, A., (2016): Analisis Sentimen pada Jasa Ojek Online Menggunakan Metode Naive Bayes, Prosiding SNST, FT, Universitas Wahid Hasyim Semarang.

[12] Prasetyo, E., (2014): Data Mining: Mengolah Data menjadi Informasi Menggu-nakan Matlab, Penerbit ANDI, Yogyakarta.

[13] Pratama, M. L., dan Murfi, H., (2014): Studi Komparasi Metode Multiclass Support Vector Machine untuk Masalah Analisis Sentimen Pada Twitter, Matematika UI, .

[14] Santosa, B., (2007): Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Graha Ilmu, Yogyakarta.

[15] Sipayung, E. M., Maharani, H., dan Zefanya, I., (2016): Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier, Jurnal Sistem Informasi, 8(1).

[16] Sun, C., Wang, X., dan Xu, J., (2009): Study on Feature Selection in Finance Text Categorization, Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, .

[17] Vapnik, V., dan Cortes, C., (1995): Support Vector Networks, Machine Learning, 20, 273–297.

[18] Vijayarani, S., dan Dhayanand, S., (2015): Liver Disease Prediction using SVM and Nave Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR), 4(4).

[19] Zaki, M. J., dan Meira, W. J., (2013): Data Mining and Analysis: Fundamental Concepts and algorithms, Cambridge University Press, Cambridge.