TEST Engineering & Management

March-April 2020
ISSN: 0193-4120 Page No. 3953 - 3961

The Accuracy Comparison of Social Media Sentiment Analysis Using Lexicon Based and Support Vector Machine on Souvenir Recommendations

Wilis Kaswidjanti[1], Hidayatullah Himawan[1] and Parasian DP Silitonga[2]

[1]Informatics Engineering Department, Veterans National Development University, Indonesia
[2]Informatics Engineering Department, Santo Thomas Catholic University, Indonesia

Article Info
Volume 83
Page Number: 3953 - 3961
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 26 March 2020*

Abstract: Social media is one source of information that can be obtained by organizations / vendors or consumers. One of the information obtained from social media is the opinion of social media users on a matter. These opinions can be positive, negative or neutral. Sentiment analysis can be used to assess the sentiments of the opinions expressed by social media users. Sentiment analysis method used is a machine learning algorithm that can help the classification process. In this study the comparison of the accuracy of opinion sentiment analysis on the recommendation of favorite souvenirs in the Yogyakarta area using the Lexicon Based method and Support Vector Machine. The processed data is Twitter and Instagram social media sentiment data. The training data used were 1000 sentiment data, while the test data for the testing process were 50 sentiment data. The test results obtained the greatest accuracy is using a Lexicon Based 87.78%, then the greatest precision results are using a Lexicon Based of 94.23%, while for the greatest recall results are using a Support Vector Machine of 100%.

Keywords: *Souvenir Recommendations, Sentiment Analysis, Social Media, Twitter, Instagram, Lexicon Based, Support Vector Machine*

I. Introduction

The role of social media in influencing people can no longer be denied [1][2]. Opinions expressed on social media are personal opinions about a matter or event, for example regarding typical souvenirs in the form of food or trinkets from the city of Yogyakarta. This opinion influences people to shop [3][4], just like shopping for souvenirs which is their favorite.

In order to find out souvenir recommendations, sentiment analysis on Twitter and Instagram is needed, which involves opinions about souvenirs, especially in the Yogyakarta area. Sentiment analysis is a process of data collection and natural language processing aimed at processing and analyzing people's opinions, people's sentiments, opinions, attitudes, emotions, etc. from comments they post on social media [5][6][7]. The purpose of sentiment analysis is how to recognize positive and negative sentiments [8] [9] [10].

In the sentiment analysis process, a machine learning algorithm is implemented that can help the classification process. Several methods can be applied together for sentiment analysis, to achieve a more accurate sentiment analysis using various approaches and techniques [11]. The choice of method used is very influential on decisions made.

Lexicon Based is based on the assumption that contextual sentiment orientation is the sum of sentiment orientation of each word or phrase. Lexicon-based approach depends on the words in the opinion [5]. Support Vector Machine (SVM) is one of the widely used supervised machine learning algorithms for textual polarity detection. Support Vector Machine (SVM) is a machine learning algorithm for classification and regression prediction to maximize prediction accuracy [12]. Therefore, this research will be carried out, namely the comparison of the use of the Lexicon Based method and Support Vector Machine (SVM), and comparing the results of the accuracy of the performance of the system testing produced from each method so that it can be seen which performance is superior to the two that method.

In this study sentiment analysis on opinions in the form of sentiments obtained from Twitter and Instagram social media using SVM and Lexicon Based methods so that it can be applied to find out which souvenirs have the most positive reviews from the public or tourists who have ever bought the souvenirs.

II. MATERIALS AND METHOD

Sentiment analysis or so-called opinion mining is the process of finding and processing opinions, opinions, emotions, and attitudes from the text [13][14]. Data needed as material for research analysis is sentiment data derived from tweets and captions from Twitter and Instagram social media [15]. Sentiment data is obtained by crawling Twitter and Instagram data to get sentiment used as research material [16]. This data crawling process utilizes the Twitter streaming API and Instagram explore. The needs analysis that will be used in this study :

1. Input data: Twitter and Instagram Sentiment Data (Test Data), Stopword Data, Lexicon Dictionary Data, Basic Word Data (Dictionary), and Synonym Word Data. Stopwords are common words and are often used in a language but these common words do not have an influence on text classification, such as: “untuk”, “kapan”, “yang”, “di”, “iya”, “itu”, etc. Sentiment Dictionaries is a collection of words that contain positive keywords and negative keywords which are used as a lexicon dictionary in this study. Basic word data is obtained from an online Indonesian dictionary. The base word data includes the base word id, word id, stopword base word, sentiment base word, and word base status. Synonym Data are several words that have the same or almost the same meaning. In this study, synonym data contains a collection of words arranged according to words that have almost the same meaning. So that it can speed up the word prepocessing process. Synonym data is obtained from a dictionary on the web http://bsd.pendukasi.id.
2. Process: Text Preprocessing, Feature Extraction Process, Training Process and Data Test.
3. Data Output: Pre comment data from preprocessing results that are ready to be classified and processed for the determination of opinions or sentiments, Statistics contains the amount of data in each sentiment category in the form of graphs, and Testing opinions or sentiments consisting of the accuracy of opinions or sentiments on the system that has been built.

The stages of sentiment analysis that will be used on this system are: Text Analysis Preprocessing, Cleansing, Tokenizing, Find Synonyms, Stopword Removel, and Stemming. SVM is trained to classify pairs of words that are synonymous and not synonymous [17]. The training phase features a feature extraction process in the comments by using Term Frequency and TF-DF (Term Frequency-Inverse Document Frequency). Steps in this step: Calculate Term Frequency and Calculate TF-IDF (Term Frequency-Inverse Document Frequency).

In the process of sentiment analysis, the first step is to enter hashtags as a keyword to retrieve sentiment data by the system from the Twitter API and Instagram API. Then, in the system, sentiment data will be searched in real time according to the inputted data. After sentiment is found, the results of the sentiment data collection which is the next text data will enter the preprocessing stage. At this stage the text data is the stage for cleaning up and uniformizing the words so that the words are ready for extraction to the next stage. The next stage is the TF-IDF feature weighting, this weighting is used before carrying out the classification stage. After weighting the next stage features the text will be classified using Lexicon Based and Support Vector Machine and will produce text data that has positive and negative sentiment classes.

Support Vector Machine (SVM) is a relatively new technique for making predictions, both in the case of classification and regression [18]. Support Vector Machine enters the supervised learning class, where in its implementation there needs to be a training phase using SVM sequential training and followed by a testing phase. The classification process using the SVM method is divided into 2 stages, namely the training data process and the test data process.

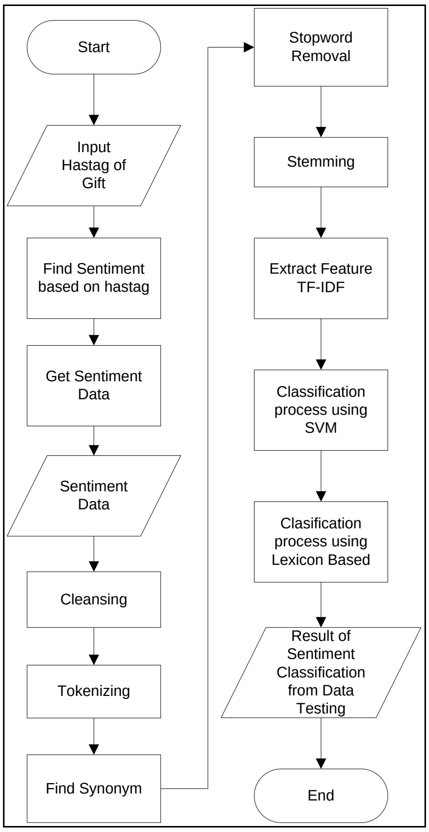
Start
Input Hastag of Gift
Find Sentiment based on hastag
Get Sentiment Data
Sentiment Data
Cleansing
Tokenizing
Find Synonym
Stopword Removal
Stemming
Extract Feature TF-IDF
Classification process using SVM
Clasification process using Lexicon Based
Result of Sentiment Classification from Data Testing
End

Fig.1. Sentiment Analysis Flowchart

Classification using the Support Vectore Machine is divided into two processes, namely the training process and testing. The training process is used to produce a sentiment analysis model which will later be used as a reference for classifying sentiments with new testing or raw data. The steps of sentiment classification algorithm using Support Vectore Machine:

1. Training. The training process will be carried out using the TF-IDF feature. The steps are as follows: (1) Calculate the weight and bias factor for the boundary plane; (2) The next step, the comment data is converted into the support vector format; (3) Kernelisasi uses the linear kernel function; (4) Look for the value of Lagrance Multiplier; (5) Calculations for data test.
2. Calculations for test data. The next process is the classification process in the support vector machine in class determination using the decision function f (x), which is the sign () function.

The training data process starts with inputting training data that has been manually labeled sentiments in the database which will then be carried out in the preprocessing stage. Furthermore, feature extraction will be performed to calculate the appearance of each word or token, then change the word into a support vector format and label the class on each word term. After that, calculating the weight (w) and bias factor (b), if the value has been obtained, the kernel is functioned using a linear kernel to be used for the calculation of the classification of test data. If the matrix value has been obtained then the last step is to calculate the Lagrange Multiplier α value which is useful for calculations on the next test data. Then the process of the test data stages begins with inputing the test data taken from the Instagram and Twitter API data which has sentiments which then enter the preprocessing text stage then compare the word or token with the training data token stored in the data in the database. If the word matches the word in the database, it will be given a weight to each term. Next do the classification calculation using the kernel linear function by multiplying the new word matrix test data with the term matrix which is a document in the training data. After the matrix results are obtained then the determination of the sentiment class using the decision function f (x) is the sign () function by entering the values of α and b obtained from the results of the training data. The sign () function is a normalization function, if the value of x on the sign function (x)> 0 then the function gives a value of 1 or positive, but if the value of x in the function is <0 then the function will give a value (-1) or negative, then the test data process is complete.

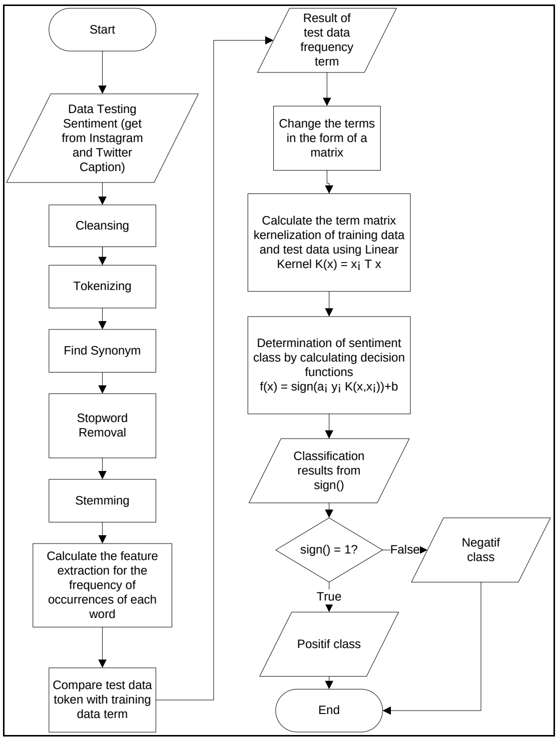
Start
Data Testing Sentiment (get from Instagram and Twitter Caption)
Cleansing
Tokenizing
Find Synonym
Stopword Removal
Stemming
Calculate the feature extraction for the frequency of occurrences of each word
Compare test data token with training data term
Result of test data frequency term
Change the terms in the form of a matrix
Calculate the term matrix kernelization of training data and test data using Linear Kernel K(x) = xᵢ T x
Determination of sentiment class by calculating decision functions f(x) = sign(aᵢ yᵢ K(x,xᵢ))+b
Classification results from sign()
sign() = 1?
False
Negatif class
True
Positif class
End

Fig. 2. SVM Classification Flowchart

Stages in determining the sentiment category on social media Twitter and Instagram using Lexicon Based :

1. Build a sentiment dictionary that is divided into positive and negative sentiments.
2. Preprocessing Text. The preprocessing stage in the lexicon analysis has differences in the analysis using the support vector machine. The difference is that there is no step in word weighting or feature extraction. Overall the steps taken are: Cleansing, Tokenizing, Stopping (Stopword Removel), and Stemming.

In this lexicon based classification process, it starts from entering test data in the form of sentiments taken from Instagram and Twitter. Then do the text preprocessing stage. After that, compare the word stemming results with the lexicon sentiment dictionary data already contained in the database. If in a comment dominant is more positive then the sentiment that will be generated is positive, if not the sentiment result is negative.

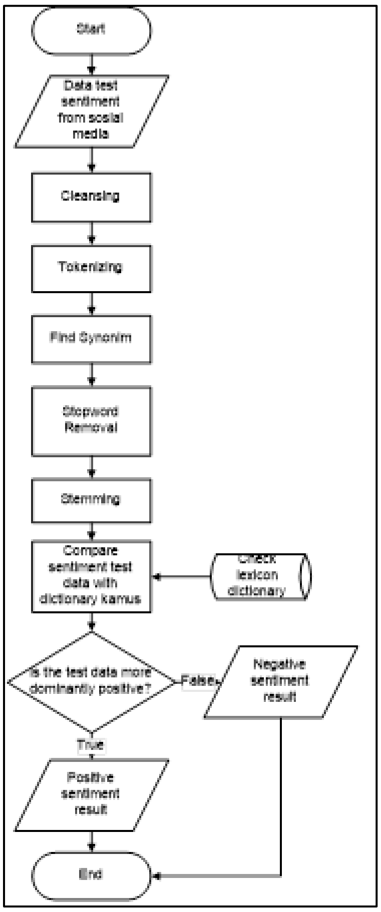
Start
Data test sentiment from sosial media
Cleansing
Tokenizing
Find Synonim
Stopword Removal
Stemming
Compare sentiment test data with dictionary kamus
Check lexicon dictionary
Is the test data more dominantly positive?
False
Negative sentiment result
True
Positive sentiment result
End

Fig.3. Lexicon Based Classification Flowchart

III. RESULT

In SVM to calculate accuracy using k-fold cross validation. Sentiment data used in this training were 1450 sentiment data. The testing process uses the approach k = 10, the 10-fold cross validation data will be divided into 10 subsets of the same size and different data. For each iteration, one part is used for testing data, the rest is used for training data. For example in the first iteration the first subset is used as test data, the second subset to ten is used as training data. In the second iteration and the second subset as test data, the first, third and ten subsets are as training data and so on until the tenth iteration. Whereas Lexicon is based to test its accuracy using the confusion matrix method. Evaluation is carried out to determine the performance of the method so that it can be used for classification.

Table 1. 10-Fold Cross Validation Test Results

Iteration	Data Pelatihan		
	Accuracy	Precision	Recall
1	91.37%	91.37%	100%
2	86.33%	86.33%	100%
3	89.93%	89.86%	100%
4	87.05%	87.05%	100%
5	81.29%	81.16%	100%
6	84.17%	84.17%	100%
7	88.49%	88.49%	100%
8	80.58%	80.58%	100%
9	86.33%	86.23%	100%
10	91.37%	91.37%	100%

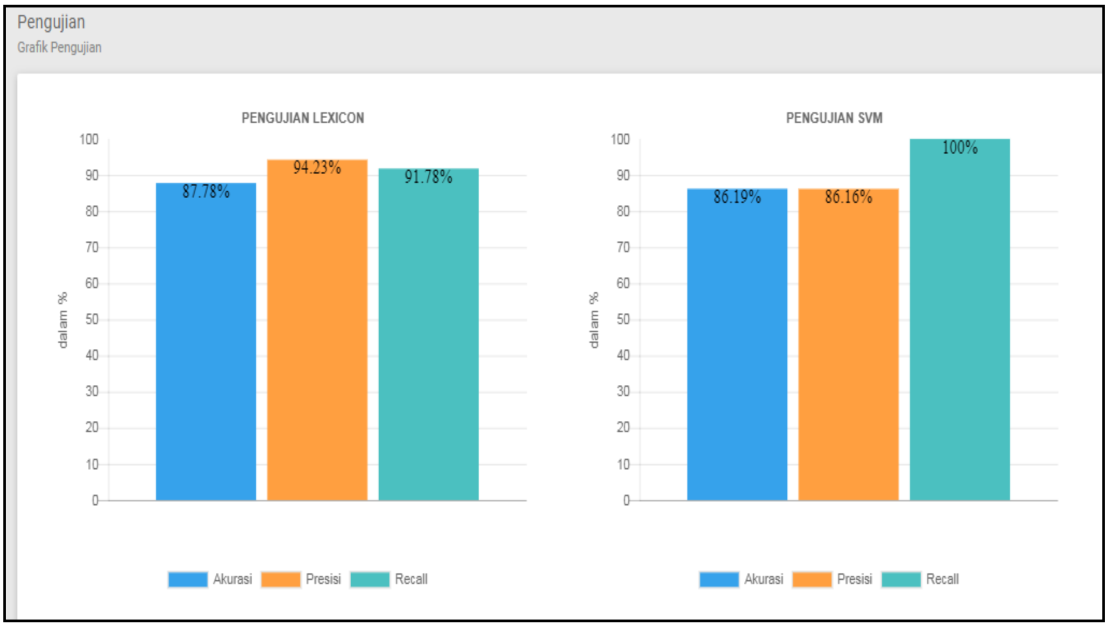

Fig. 4. Testing Graph on The System

From the analysis table of 50 test data, a confusion matrix table can be formed, can be seen in table 2, table 3 and table 4.

Table 2. Confusion Matrix Table

Analysis Method	Confusion Matrix			
	TP	FN	FP	TN
SVM	41	4	3	2
LB	40	5	1	4

Table 3. Test results using Support Vector Machine

Test Sentiment	Accuracy	Precision	Recall
50 test data	86%	93.20%	91.11%

Table 4. Test results using Lexicon Based

Test Sentiment	Accuracy	Precision	Recall
50 test data	88%	97.56%	88.89%

After testing, it can be seen that the lexicon based method provides better accuracy and precision than the support vector machine method with an accuracy of 87.78% and a precision of 94.23%. As for the recall results, the support vector machine method is better than the lexicon based method with a recall of 100%.

IV. CONCLUSION

1. The results of the analysis on the system that was built showed that sentiment analysis conducted using sentiment data obtained from social media Twitter and Instagram can produce a recommendation for a favorite, especially in Yogyakarta with a total of 30 souvenirs that have been sorted by souvenirs. has the highest weight.
2. The test results on the built system show that the lexicon based method provides better accuracy and precision than the support vector machine method with an accuracy of 87.78% and a precision of 94.23%. As for the recall results, the support vector machine method is better than the lexicon based method with a recall of 100%.
3. The lexicon based method is highly dependent on the amount of sentiment dictionary data used, the more the amount of dictionary data used the greater the accuracy value, and conversely the less the amount of dictionary data used the smaller the accuracy value produced.

REFERENCES

[1] R. Rezapour, L. Wang, and J. Diesner, “Identifying the Overlap between Election Result and Candidates ’ Ranking based on Hashtag-Enhanced , Lexicon-Based Sentiment Analysis,” 2017.

[2] R. V Karthik, S. Ganapathy, and A. Kannan, “A Recommendation System for Online Purchase Using Feature and Product Ranking,” *2018 Elev. Int. Conf. Contemp. Comput.*, pp. 1–6, 2018.

[3] S. V. S. Ananth, C. Pmss, and D. Ph, “Live Twitter Knowledge as a Corpus for Sentiment Analysis and Opinion Mining,” vol. 13, no. 7, pp. 3679–3686, 2017.

[4] K. Y. Rao, “Product Recommendation System from Users Reviews using Sentiment Analysis,” no. July, 2017.

[5] A. Kabiri, “Translation Is Not Enough : Comparing Lexicon- based Methods for Sentiment Analysis in Persian,” no. Ml, pp. 36–41, 2017.

[6] J. Serrano-guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-viedma, “Sentiment analysis : A review and comparative analysis of web services,” *Inf. Sci. (Ny).*, vol. 311, pp. 18–38, 2015.

[7] K. Ravi and V. Ravi, *A survey on opinion mining and sentiment analysis : tasks , approaches and applications*, no. June. Elsevier B.V., 2015.

[8] S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, “Feature Selection and Ensemble Construction: A Two-step Method for Aspect Based Sentiment Analysis Md,” *Knowledge-Based Syst.*, 2017.

[9] M. Singson, K. Nagarattiname, and T. Gogoi, “The Sellout: Readers Sentiment Analysis of 2016 Man Booker Prize Winner,” 2017.

[10] F. I. Tanesab, I. Sembiring, and H. D. Purnomo, “Sentiment Analysis Model Based On Youtube Comment Using Support Vector Machine,” vol. 6, no. 8, pp. 180–185, 2017.

[11] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” *Knowledge-Based Syst.*, vol. 89, pp. 14–46, 2015.

[12] M. Ahmad and I. Ali, “Sentiment Analysis of Tweets using SVM,” vol. 177, no. 5, pp. 25–29, 2017.

[13] A. I. Journal, S. Liu, and I. Lee, “Email Sentiment Analysis Through k-Means Labeling and Support Vector Machine Classification,” vol. 9722, no. May, 2018.

[14] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis : A survey,” no. October 2017, pp. 1–25, 2018.

[15] H. Saif, Y. He, M. Fernandez, and H. Alani, “Contextual semantics for sentiment analysis of Twitter,” *Inf. Process. Manag.*, vol. 52, no. 1, pp. 5–19, 2016.

[16] R. Liang and J. qiang Wang, “A Linguistic Intuitionistic Cloud Decision Support Model with Sentiment Analysis for Product Selection in E-commerce,” *Int. J. Fuzzy Syst.*, vol. 21, no. 3, pp. 963–977, 2019.

[17] K. Chinniyan, S. Gangadharan, and K. Sabanaikam, “Semantic Similarity based Web Document Classification Using Support Vector Machine,” vol. 14, no. 3, 2017.

[18] Y. Liu, J. Bi, and Z. Fan, “A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm Yang,” *Inf. Sci. (Ny).*, 2017.