Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 3, December 2019, pp. 1555~1561
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1555-1561

An automatic lexicon generation for Indonesian news sentiment analysis: a case on governor elections in Indonesia

Media Anugerah Ayu, Sony Surya Wijaya, Teddy Mantoro
Department of Computer Science, Faculty of Engineering and Technology, Sampoerna University, Indonesia

Article Info

Article history:

Received May 1, 2019
Revised Jul 1, 2019
Accepted Jul 28, 2019

Keywords:

Automatic lexicon generation
Indonesian lexicon
Lexicon generation
News sentiment analysis
Sentiment analysis

ABSTRACT

Sentiment analysis has been popularly used in analyzing data from the internet. One of the techniques used is lexicon based sentiment analysis. Generating lexicon is not an easy process, and lexicon in Bahasa Indonesia is rarely available. This paper proposes an automatic lexicon generation in Bahasa Indonesia for sentiment analysis purpose. Experiments were performed using the generated lexicon for doing sentiment analysis on Indonesian political news about the 2018 governor election in three provinces in Indonesia. The conducted experiments show promising results where it can predict the candidate's rank, the election winner, and the percentage of votes for each candidate with better accuracy than the previous work which used manually generated lexicon.

Corresponding Author:

Media Anugerah Ayu,
Department of Computer Science,
Sampoerna University, Jakarta-Indonesia.
Email: media.ayu@sampoernauniversity.ac.id

1. INTRODUCTION

The Rapid development of internet technology which is supported by the fast development of mobile devices has made online news websites become a favorite source of information that accessed by many people in the current digital era. In Indonesia, according to the Alexa Rank, two of the top five websites are news websites [1]. This mean that news websites can be a beneficial source to obtain a big picture on the current issue and they could reflect the happenings in the coun-try either its social happening, political happening, or economical happening. Dur-ing the election year, analyzing the news sentiment on the candidates involved would bring valuable insight that can map on what has happened during the campaign and what possibly can happen on the D day.

Sentiment analysis is defined as a process of detecting, extracting, and classi-fying user's opinions and attitudes toward certain topics [2]. Sentiment analysis determines whether a text has a positive, neutral, or negative sentiment. The rise of opinion dataset over the internet makes sentiment analysis become a potential field to study [3]. News articles and social media posts are among the datasets used in the sentiment analysis. Mada and Nurwidyantoro [4] analyzed the senti-ment of economic news in Indonesia. Whereas, [5] and [6] did sentiment analysis in Facebook. Then, Ozturk and Ayvaz [7] used twitter post as the dataset of sen-timent analysis about the Syrian refugee crisis.

The basic of sentiment analysis is sentiment categorization. One of the basic approaches for this sentiment categorization is lexicon based approach. In this approach the lexicon plays a major role in the process of sentiment classification. Lexicon based approach classifies the sentiment based on the dictionary provid-ed. Several researchers have studied the construction of English lexicon in various domains. It even has been provided in libraries of some programming languages like R and Phyton. However, for language other than English the lexicon diction-ary is limited. Therefore lexicon dictionary for sentiment in non-

English language needs to be developed. Lexicon generation can be performed manually or automatically. Turkish lexicon containing 5405 words was developed manually in-volving three experts by Ozturk and Ayvaz [7]. Lexicon for Indonesian politic was also built manually by Soroinda, Rachim, and Wonggo [8] involving a politi-cal expert and 300 seed words. This resulted on 12 sentiment words where 6 for positive sentiment and another 6 for negative sentiment. The disadvantage of generating lexicon manually is the need for experts in the domain context. As shown in a study by Fast et al. [9] using experts in constructing domain-specific lexicon is very difficult. Constructing lexicon manually is usually also time con-suming and labour intensive [10].

This paper presents a study on lexicon based sentiment analysis using Indonesian news articles as the datasets where the lexicon is automatically generated. The proposed sentiment analysis shows promising results when it was tested on the case of governor elections in 3 provinces in Indonesia.

Next section of this paper presents reviews on research works related to the focus of this study, i.e. sentiment analysis and lexicon generation. The following section discusses the methodology used in conducting this study. It is then fol-lowed by the result and discussion section. This section presents results from the experiments conducted and as well discussions on the meanings and implications of the results. This paper is then closed with a conclusion section [4].

2. RELATED WORK

This section discusses reviews on previous works related to sentiment classification and lexicon generation, especially on the techniques/approaches used.

Generally there are three main techniques used in sentiment classification, i.e. rule-based, lexicon-based (lexical knowledge), and machine learning. Rule-based is a classification by defining several rules to the text. According to the research by Devika, ruled-based classification performs better in sentence level rather than word level. Moreover, the performance of rule-based classification is determined by the rules used. If more rules used, it will be more complicated [11].

Lexical knowledge classifies sentiment of a term based on the dictionary provided, which is usually called as lexicon [12]. The process with this technique done by counting and weighting the sentiment words that have been selected through evaluation process [13]. The third approach is machine learning. While rule-based is classifying using defined rules, machine learning offers a classification algorithm which learns from data provided or processed. Machine learning is a process of approximating the upcoming output by using training data or only the input itself. Machine learning methods commonly divided into two categories: supervised learning and unsupervised learning [14].

Supervised learning is a machine learning that uses a sample data as the training data. Based on the training data, supervised machine learning predicts the output of the input data. Algorithms used in supervised learning includes artificial neural network (ANN), multi-layer perceptron, and decision tree. There are some researchers which uses supervised learning as their approaches for sentiment analysis. Sharma and Dey [15] used Back Propagation Artificial Neural Network (BPANN) to evaluate the sentiment of movie reviews. Lexica from previous re-search are used as training data. There are three lexica used, which are Hatzivass-iloglou & McKeown, General Inquirer, and the Opinion Lexicon. The movie re-view from IMDB.com is used for test data. Besides performing sentiment analysis for the movie review, Sharma and Dey also analyzed the performance of BPANN. They evaluate the performance of the approach by comparing the precision and recall of the example and result data. Sharma and Dey showed that BPANN was performing well in sentiment analysis and reducing the dimensionali-ty.

Support vector machine (SVM) and naïve bayes (NB) were used in sentiment analysis conducted by Zhang et al. [16] for internet restaurant reviews written in Cantonese. Whereas Pang et al [17] used NB, SWM, and maximum entropy as the methods for their sentiment classification. Another work, by Li et al [18], ap-plied deep learning through an approach called a sentiment-feature-enhanced deep neural network (SDNN) for text classification of its sentiment analysis. Deep learning was also adopted by Empath a sentiment lexicon constructed by Fast et al. [9].

Contrary to supervised learning, there is no training data required in unsuper-vised learning. It only needs input data. Unsupervised machine learning works by finding the regularities in the input data. Unsupervised learning algorithms include different type of clustering techniques. One of the clustering techniques is K-Means clustering. K-means clustering works by assigning data points to k cluster centroid. Each data point is grouped to the closest centroid iteratively until the centroid does not change [14].

The application of k-means clustering in sentiment analysis is presented in Fei Liu research [19]. In their study, documents of movie review clustered into two clusters: positive and negative. Fei Liu was also implementing Term Frequency-Inverse Document Frequency (TF-IDF) and the voting mechanism to improve the accuracy of the clustering approach. TF-IDF is an algorithm for calculating the weight of a term.

A term is important when it has a high frequency in a single document, but the low frequency in the collection of all other documents. TF-IDF used to improve the efficiency of raw data, while the voting mechanism is used to extract more stable clustering result. According to their research result, it shows that sentiment analysis using only k-means clustering is giving a poor outcome, to improve the accuracy, TF-IDF and voting mechanism technique should be implemented.

3. RESEARCH METHODS

This section presents the research process conducted in this study and also the research design which is based on experimental design research. Two main pro-cesses as part of this research are lexicon generation and sentiment analysis. Da-ta collection is the first step for the two processes. Data collection was done for the input for lexicon generation and another data collection needs to be done for preparing datasets for the sentiment analysis process. The flow of data collection process is depicted in Figure 1. Whereas the flow of process for lexicon generation and sentiment analysis can be seen from the diagram presented in Figure 2.

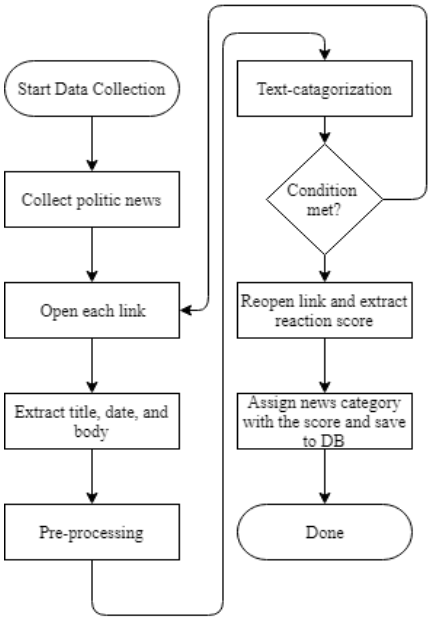

Figure 1. Data collection process

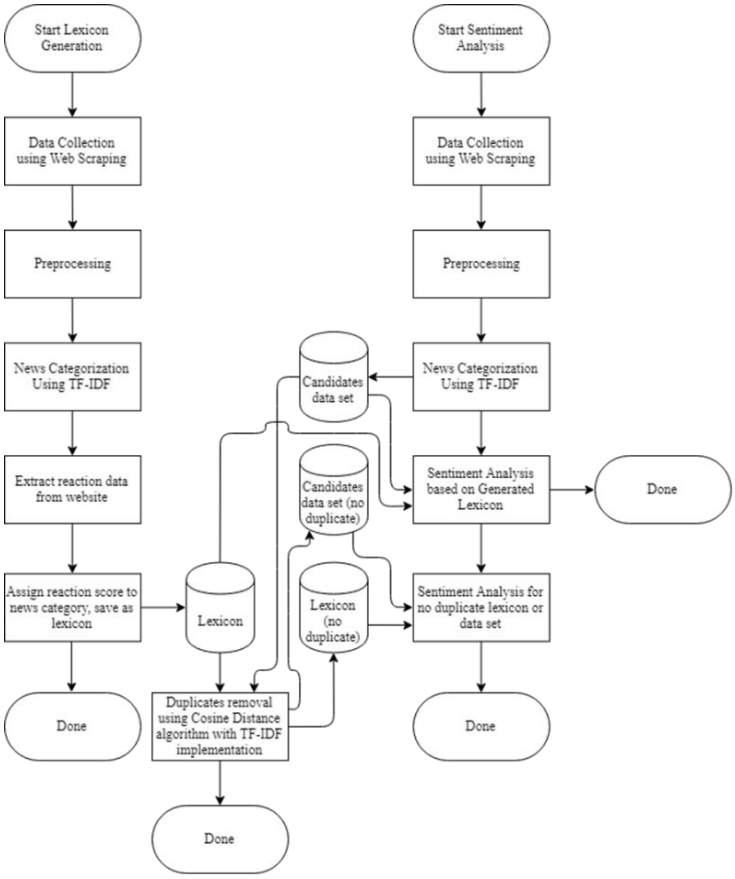

Figure 2. Lexicon generation and sentiment analysis process

Another step done in the two processes is text categorization. Text categoriza-tion is a process which is performed in lexicon generation and also in sentiment analysis. In this study the text categorization process was conducted based on weighting scheme. TF-IDF algorithm was used to weight the most important term in a document which can define the category of the document. The TF-IDF was calculated based on the following formula:

$$w_{i,j} = tf_{i,j} \times (\frac{N}{df_i}) \quad (1)$$

where $w_{i,j}$ is the weight for term i in document j, $tf_{i,j}$ is the term frequency of i in document j, N is the number of the document in the collection, and df_i is the document frequency of term i in the collection.

3.1. Lexicon Generation

Lexicon is constructed to determine the text sentiment score. In this study the lexicon was generated automatically, utilizing detik.com as the source of data. This site was chosen due to its popularity as the most visited news websites in In-donesia [1]. Detik.com offers reaction button to the news which consists of eight (8) reactions as depicted in Figure 3. In this study, 2 reactions, i.e. entertained and shocked were removed to avoid the ambiguity of the reaction and possible sar-casm reaction. Also, as Ekman's theory suggested of six basic human's emotional categories [20]. Happy, inspired, and proud reactions are considered as positive sentiments. Whereas sad, afraid, and angry reaction are considered as negative sentiments.

Figure 3. Reaction buttons feature on Detik Website

The percentage for each reaction represents readers' feedbacks on the news. These percentages were collected automatically and the total sentiment was then calculated using the following formula:

Total sentiment= $\sum$ Positive Sentiment- $\sum$ Negative Sentiment

3.2. Sentiment Analysis

In this study the sentiment analysis was done based on the generated lexicon discussed earlier. The dataset for the lexicon based sentiment analysis performed in this study is the news about the candidates of 2018 governor election in Jawa Barat, Jawa Tengah, and Jawa Timur provinces. Data was collected from news articles published within the registration of candidates to the end of campaign pe-riod, i.e. 8 January 2018 - 23 June 2018. As depicted through a diagram present-ed in Figure 2, after the data is collected from three news websites, i.e. detik.com, tribunnews.com, and liputan6.com, then the data will go through a preprocessing phase. It is then followed by text categorization process using TF-IDF algorithm. Following that, the sentiment score of the category will be checked based on the generated lexicon. It is expected that the results from this sentiment analysis will be able to be used to predict the outcome of the elections.

3.3. Experimets

Experiments were conducted to evaluate the performance of the proposed technique. Three variables involved in the experiments, i.e. lexicon data and can-didate data as independent variables, and sentiment analysis accuracy as the de-pendent variable. The experiments conducted were based on the design of exper-iments presented in Table 1. Data resulted from the experiments conducted is evaluated based on its accu-racy performance. Evaluating the accuracy of the experiments from the senti-ment analysis results can also show how accurate the lexicon generated in this study is.

Table 1. Design of Experiment with Two Independent Variables

	Candidate with Duplicates (Cd)	Candidate without Duplicates (C)
Lexicon with Duplicates (Ld)	Experiment LdCd	Experiment LdC
Lexicon without Duplicates (L)	Experiment LCd	Experiment LC

The accuracy of sentiment analysis processes is evaluated based on three in-dexes, i.e. precision, recall rate, and accuracy, calculated using the following equations:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where TP is the number of true positive instances, FN is the number of false negative instances, FP is the number of false positive instances, and TN refers to the number of true negative instances as presented in Table 2 as a confusion matrix [21].

Table 2. Confusion Matrix Table

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

In this study the actual positive and actual negative data is taken from the re-sults of 2018 governor election from Komisi Pemilihan Umum (KPU), the official commission of the 2018 governor election. Two predictions will be evaluated us-ing the confusion matrix which are the prediction of the 2018 governor election rank and winner. The mean absolute percentage error (MAPE) is then calculated to measure the error of vote percentage predictions. The following formula is used for the error calculation.

$$Percent\ Error = \frac{Predicted\ Value - Actual\ Value}{Actual\ Value} x\ 100 \quad (5)$$

$$MAPE = \frac{\sum Percent\ Error}{n} x\ 100 \quad (6)$$

4. RESULTS AND ANALYSIS

This section presents results from the experiments conducted which based on the design of experiments described in sub-section 3.3. Three types of accuracy were evaluated, which are: 1. the accuracy of the sentiment analysis results in predicting correct order of the candidate rank, 2. the accuracy of the sentiment analysis results in predicting the election winner, and 3. the accuracy in predicting total vote percentage of each candidate. Here election winner defined as the candidate with the highest score, while vote percentage de-fined as total vote gained by a candidate from all the vote submitted. All the prediction will be compared to the result from KPU, the official commission for the election. The results of the accuracy performance evaluation from the experiments conducted are presented in Table 3, 4 and 5.

Table 3. The accuracy performance of the lexicon based sentiment analysis in predicting the candidate ranks

Experiment	Province	Precision (%)	Recall (%)	Accuracy (%)	Average Accuracy (%)
LdCd	Jawa Barat	50	50	75	91.67
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LC	Jawa Barat	25	25	62.5	87,5
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LCd	Jawa Barat	25	25	62.5	87,5
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LdC	Jawa Barat	25	25	62.5	87,5
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	

The experimental results presented in Table 3 shows that LdCd sentiment analysis result has better accuracy in predicting correct order of the governor elec-tion rank rather than the other three conditions. Election candidate ranks in Jawa Tengah and Jawa Timur province has been predicted with 100% accuracy in all experiments. Data in Table 4 shows experimental results for election winner prediction. The data shows that the sentiment analysis yielded 100% accuracy in predicting the winner of 2018 governor election in Jawa Barat, Jawa Tengah, and Jawa Timur.

Data presented in Table 5 show that basically the MAPE are not exceeding 40% with the highest resulted from the experiments are 39.14 percent. On aver-age the LCd experiment has the highest MAPE of 26.20 percent, and the lowest MAPE of 18.25 percent has been resulted from LdC.

Table 4. The accuracy performance of the lexicon based sentiment analysis in predicting the candidate winner

Experiment	Province	Precision (%)	Recall (%)	Accuracy (%)	Average Accuracy (%)
LdCd	Jawa Barat	100	100	100	100
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LC	Jawa Barat	100	100	100	100
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LCd	Jawa Barat	100	100	100	100
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	
LdC	Jawa Barat	100	100	100	100
	Jawa Tengah	100	100	100	
	Jawa Timur	100	100	100	

Results presented in Table 5 also suggest that the performance of this auto-matic lexicon generation based sentiment analysis is better than the previous one conducted by Soroinda, Rachim, and Wonggo [8] where the lexicon was manual-ly generated. The previous work showed more than 48 percentage of errors.

Table 5. Percentage of Error of each Sentiment Analysis Result

Experiment	MAPE Score (%)			Average (%)
	Jawa Barat	Jawa Tengah	Jawa Timur	
LcDc	33.93	31.35	1.79	22.36
LC	36.14	30.62	2.09	22.95
LCd	39.14	38.63	0.82	26.20
LdC	28.22	23.46	3.06	18.25

5. CONCLUSION

This study shows that an automatic lexicon generation for the Indonesian news sentiment analysis has been successfully developed. The lexicon has been tested for sentiment analysis of the 2018 governor election in three provinces in Indonesia. The process utilized TF-IDF algorithm for its text categorization phase. Data from the experiments conducted show a good performance of the sentiment analysis in predicting the election results. The experimental results show that the highest error of the analysis is 26.2 percent in predicting the vote percent-age of each candidate. This is a promising result compared to a previous work which used manually generated lexicon.

REFERENCES

[1] Alexa, "Top Sites in Indonesia," 2018. Retrieved from https://www.alexa.com/topsites/countries/ID
[2] A. Montoyo, et al., "Subjectivity and sentiment anal-ysis: an overview of the current state of the area and envisaged developments," *Decision Support Systems*, vol. 53, pp. 675–679, 2012.
[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, 2008.
[4] U. G. Mada, et al., "Sentiment Analysis of Eco-nomic News in Bahasa Indonesia Using Majority Vote Classifier," 2016.
[5] C. Troussas, et al., "Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning," *4th International Conference on Information, Intelligence, Systems and Applications (IISA 2013)*, pp. 198-205, 2013.
[6] A. Ortigosa, et al., "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, pp. 527-541.
[7] N. Öztürk and S. Ayvaz, "Sentiment Analysis on Twitter : A Text Mining Ap-proach to the Syrian Refugee Crisis," *Telematics and Informatics*, 2017.
[8] A. A. R. Soroinda, et al., "A Corpus-Based Lexicon Building in Indonesian Political Context Through Indonesian Online News," pp. 347-352, 2016.

[9] E. Fast, et al., "Empath: Understanding Topic Signals in Large-Scale Text," *Proceedings of the 2016 CHI Conference on Human Factors in Com-puting Systems (CHI 2016)*, pp. 4647-4657, 2016.
[10] C. S. G. Khoo and S. B. Johnkhan, "Lexicon-based sentiment analysis: Compara-tive evaluation of six sentiment lexicons," *Journal of Information Science*, vol. 44, pp. 491-511, 2018.
[11] M. D. Devika, et al., "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016.
[12] P. Melville, et al., "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification," *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, pp. 1275-1284, 2009.
[13] Z. Hailong, et al., "Machine Learning and Lexicon Based Me-thods for Sentiments Classification: A Survey," *Proceedings of 11th Web Information System and Application Conference (WISA 2014)*, pp. 262-265, 2014.
[14] E. Alpaydin, "Introduction to Machine Learning (2nd ed.)," The MIT Press, 2010.
[15] A. Sharma and S. Dey, "An Artificial Neural Network Based Approach for Sen-timent Analysis of Opinionated Text," *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, New York, NY, USA: ACM, pp. 37-42, 2012..
[16] Z. Zhang, et al., "Sentiment classification of Internet res-taurant reviews written in Cantonese," *Expert Systems with Applications*, vol. 38, pp. 7674-7682, 2011.
[17] B. Pang, et al., "Thumbsup?:sentiment classification us-ing machine learning techniques," *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP'02), Association for Computational Linguistics, Stroudsburg, Pa, USA*, vol. 10, pp.79-86, 2002.
[18] W. Li, et al., "An Improved Approach for Text Sentiment Classification Based on a Deep Neural Network via a Sentiment Attention Mechanism," *Future Internet*, vol. 11, 2019.
[19] G. Li and F. Liu, "Application of a clustering method on sentiment analysis," *Journal of Information Science*, vol. 38, pp. 127-139, 2012.
[20] P. Ekman, "An Argument for Basic Emotions," *Cognition e-Emotion*, vol. 6, pp. 169-200, 1992.
[21] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and ROC Curves," *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA: ACM, pp. 233-240, 2006.