Part of Speech Features for Sentiment Classification based on Latent Dirichlet Allocation

Eka Surya Usop[1], R. Rizal Isnanto[2], Retno Kusumaningrum[3]
1. Magister of Information System, Universitas Diponegoro, Semarang, Indonesia
2. Department of Computer Engineering, Universitas Diponegoro, Semarang, Indonesia
3. Department of Informatics, Universitas Diponegoro, Semarang, Indonesia
ekasurya@student.undip.ac.id[1], rizal_isnanto@undip.ac.id[2], retno@live.undip.ac.id[3]

***Abstract*— The input data used in the sentiment analysis process by using machine learning generally is Bag of Word (BoW). However, the input data using BoW is not enough to improve the machine learning in defining the polarity in a document. Therefore, need input in the form of more specific feature so that it is capable to give the more maximal result. Part of Speech (POS) is one of the techniques to create the more specific feature in a document. By using the POS-based feature in a document, then the occurrence of the word class like adjective or negation can be detected. The adjective and negation are the main sign of the sentiment or opinion in a document. This study is aimed to use POS technique to conduct feature selection. The result of the POS-based feature process will be the input for sentiment analysis process by using Latent Dirichlet Allocation (LDA) method. The result of this research showed that the document which has passed the POS-based feature process can give accuracy score higher with the difference about 7.8% than the document without feature selection process of POS.**

Keywords*— Sentiment Analysis, POS-based feature, Machine Learning, Part of Speech, Latent Dirichlet Allocation

I. INTRODUCTION

The sentiment analysis has been currently growing rapidly, mainly since the appearance of Web 2.0 and the technology to the social media for instance Blog, Facebook, and Twitter. Twitter is one of the microblogging sites that allow the users to write about many topics and discuss current issues. It causes twitter being the rich data sources to be analyzed [1].

Sentiment analysis is computation study in mining text field which learns about idea or opinion, sentiment, emotion even the attitude expressed in the text. The main task of sentiment analysis is conducting sentiment classification to find out the polarity of the document. That is contained sentiment value positively or negatively [2].

Mostly the techniques for sentiment classification use machine learning method. The input used in the machine learning is unigram token or can be said Bag of Word (BoW). However, BoW is not enough, instead of the more specific features which is used in sentiment classification process. The right features can give the best results in the classification process [3]. The feature selection process such as n-gram, TF, TF-IDF and POS Tagging can be used to improve the machine learning performance [4].

In this study, the sentiment classification process will use the input data the BoW and the data with the POS-based feature. The data with the POS-based feature is formed based on the POS occurrence and the frequency of the POS occurrence. The POS-based feature process in this study uses the data from Dictionary of Indonesian Language (Kamus Besar Bahasa Indonesia/ KBBI). Those three types of input data are used in sentiment classification by using Latent Dirichlet Allocation (LDA) method. The accuracy value of those three input data will be compared to obtain the best classification model.

II. RELATED WORK

This section explains about the previous studies about sentiment analysis and feature selection which is used as input data in classification.

One of the first researchers was about sentiment analysis using machine learning which was conducted by Pang and Lee [5] used Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM). In the present study, those three methods were used to find out the sentiment in the document of movies' review. The result was that SVM has the better result than other methods.

The similar research has been conducted by Bilal et al [6]. In this research, the sentiment analysis was conducted by three classification method those are NB, decision tree, and K-Nearest Neighbors (KNN). The feature used in classification process is BoW and TF-IDF. The analysis result showed that Naive Bayes method has better performance than Decision tree and KNN methods. It can be seen from the higher accuracy value, precision, and recall.

Another research of sentiment analysis was conducted by Wan and Gao [7], they conducted sentiment analysis to measure the services of Airline Service from twitter data, and this research used several algorithms of machine learning, for instance, Naïve Bayes, SVM, Bayesian Network c4.5 decision tree and random forest. The feature used was n-gram and POS. Research result showed that it can increase the accuracy result on sentiment classification with Twitter data.

III. METHODOLOGY

This section explains the steps conducted by the researcher in this study, start from the preprocessing, POS-based features and the last is sentiment classification process using LDA method.

A. Preprocessing

The process of preprocessing text which is aimed to eliminate the inconsistent data, data duplication, reduce the noise, and correct the wrong writing of the data. Some steps of preprocessing text processes such as tokenization, normalization, and stopword removal [8].

Tokenization is a process of breaking the text document to be the units called token. The token can be the word, number, and punctuation. To obtain the optimal token, then in the tokenization process was also conducted the filtering process used to eliminate the illegal characters of the document. The characters eliminated were URL, username, emoticons, hashtags, etc. Moreover, it's also conducted the casefolding which was aimed to change all of the tokens being the small font.

Normalization is used to spelling improvement. Twitter data have the high frequency of misspelling; therefore the spelling repair is needed in the word or token. The spelling improvement process is important because the token with misspelling will not readable in the next process. Normalization process in the current study was used the Jaro-Winkler Distance algorithm. Each word will be measured the distant by using KBBI.

Stopword removal was aimed to eliminate the words that often appear but do not a have contributed to data analysis process. By eliminating the stop-word then the data dimension can be reduced and computation time increase. For example like the word "dan","yang","di","ke", (and, that, in, to) those words are often found in all document.

B. POS-based Features

Part of Speech (POS) is a categorization way of word classes, such as noun, verb, adjective, etc. By using POS, word class annotation process for each word in the document can be conducted automatically. POS received the input of text in Indonesian language and will give an output of word line enclosed with related word class [9].

The current research used feature selection with POS to create two types of feature that will be used in data classification process. Document forming process was being the POS feature used in this research, based on the KBBI. On the first POS feature (POS-1), the words in the training document and testing document were firstly matched with KBBI and changed to be the word class form in accordance with the tagging as shown in Table I if the words were in KBBI and if the words were not in KBBI then they will be deleted from the document. The second POS feature (POS-2) was obtained by counting the number of the word class occurrence of POS-1. The examples of feature selection process as shown in Table II.

TABLE I. TAG PART OF SPEECH

Tag	Deskripsi
VB	Verb
NN	Noun
ADJ	Adjective
ADV	Adverb
NNP	Proper noun.
NEG	Negation

TABLE II. POS-BASED FEATURES PROCESS

Document	dukung ahok pemimpin penuh inovasi
POS-based KBBI	dukung/verb ahok/xxx pemimpin/noun penuh/adjective inovasi/noun
POS-1	verb noun adjective noun
POS-2	VB-1 NN-2 ADJ-1 ADV-0 NNP-0 NEG-0

C. Latent Dirichlet Allocation (LDA)

LDA is a general model of the probability of a corpus. The main task of LDA was finding the latent variable in a text document that is the topic. The topic can be described from the words distribution in the corpus. The word with high probability, have high possibility to be a topic [10]. By classifying the data based on the certain topics on the data in large number, then it will be easier in information digging from that data [11].

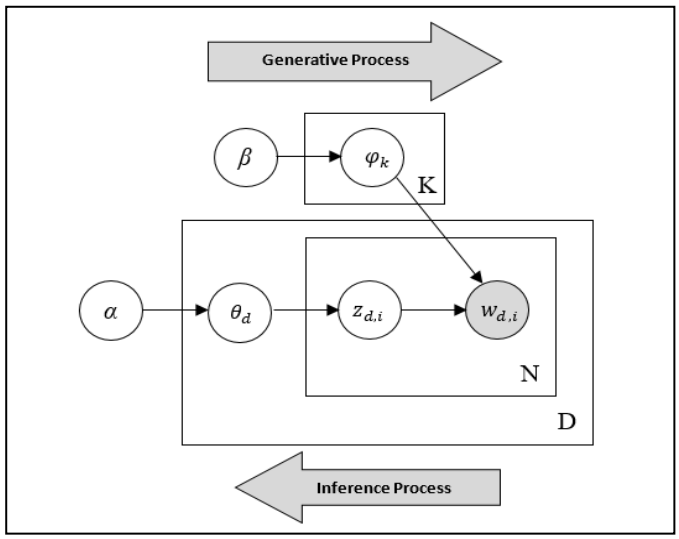

Fig. 1. Graphical model of Latent Dirichlet Allocation

LDA has two models; those are the generative process and inference process. In the case of sentiment analysis on this study, the LDA model used is inference process. The LDA model of inference process was aimed to identify the latent variables (φ_k, θ_d) and obtain the word w distribution for each topics k by adding the hyperparameter α and β. General process on LDA is shown in Fig 1.

LDA inference has two main processes those are training and testing. The training process was aimed to obtain the w word distribution for each k topics [12]. To obtain that word distribution, LDA inference uses Gibbs sampling algorithm [13]. This method was chosen because easy to apply and imply [12]. The main concept of Gibbs Sampling method is conditional distribution samples generation from each variable that the score has been known. The score of each variable was always upgraded in accordance with the probability score of other variables [14].

D. Sentiment Classification using LDA

As described previously, the classification process has two parts that are interrelated those are training process and testing process. The training process is the process to generate the classification model. The process of generating a classification model in this study was used LDA as inference process by using a Gibbs sampling algorithm. The classification models produced from the training process were hereafter used in the testing process. The classification model is the value of word distribution for each topic (φ_k), topic proportion for each document (θ_d), topic proportion for each class, and topic distribution [15].

The testing process was aimed to know the extent of that model is capable to conduct the document classification and to know the best classification model from all the data processed. In training process, we applied Kullback Leibler Divergence (KLD) to measure distribution similarity between topic proportion for each class from model classification and topic proportion from the testing process [16]. the smallest value of KLD indicated as the sentiment class of the document [12].

The classification process uses 10 fold cross validation technique to divide the dataset into training and testing data. The performance from classification model is measured from an average of accuracy value over 10-folds. The framework of sentiment classification process is shown in Fig 2.

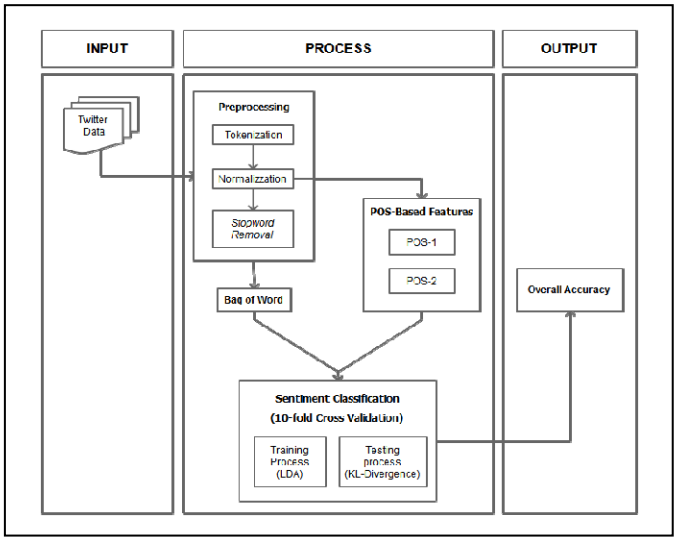

Fig. 2. Sentiment classification process

IV. Experiment and Result

This section describes the experimental stages, beginning with experimental setup, experimental scenarios used and last is the discussion of experimental results.

A. Experimental Setup

The data used in the current research was collected by using advanced search facility provided by twitter. This facility is easier the users to get the data with such criteria and able to take the data without any limited time. The data used for this research were 500 documents of tweet including 250 data with the positive label and 250 data with the negative label. Those data then were used to sentiment classification process using LDA method. There were training process and testing process in sentiment classification process. Therefore, the data were divided into train data and test data with 10-fold cross validation technique. The dataset in the total of 500 document was divided into 10 parts, each part contains 50 documents including 25 documents with the positive label and 25 document with the negative label. By using 10-fold cross validation technique then the training process and testing process would be conducted in ten times of experiments. The result from each experiment was the average value of accuracy or can be said overall accuracy (OA).

This experiment was implemented with PHP and MySQL on Windows 10 - 64 bit. The hardware used was the computer with the specification of Intel Processor Core i3, memory 4 GB, and hard disk 500 GB.

B. Experimental Scenarios

This study was conducted three scenarios of the experiment. Each experiment has different input data those are POS-1, POS-2, and BoW. Each input data will be analyzed by parameter combination of Gibbs sampling that was alpha (0.1, 0.01, 0.001) and beta (0.1, 0.01, 0.001) and three topics. Therefore, in each experiment, there were 27 parameter combinations. The goals of this experiment were obtaining the highest OA from three scenarios.

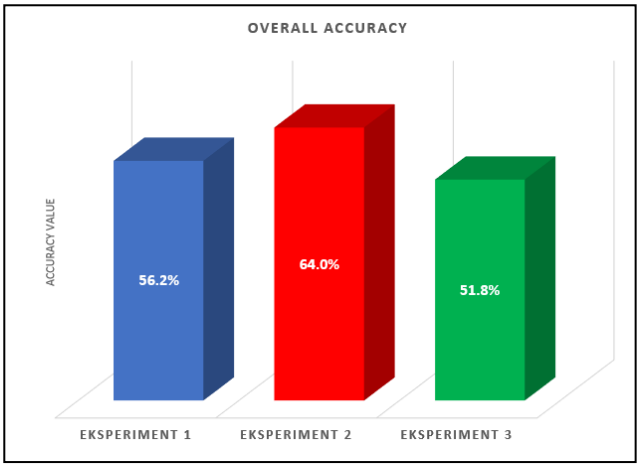

Fig. 3. Comparison of experiment result

C. Result and Analysis

Fig. 3 describes the comparison of the highest OA in each experiment. The first experiment revealed the highest overall accuracy value was 56.2 % with α=0.001 β=0.01 and topic=3. The second experiment revealed the highest overall accuracy value was 64 % with α=0.001 β=0.001 and topic=10. The third experiment revealed the highest overall accuracy value was 51.8 % with α=0.001 β=0.001 and topic=5.

The data in the first and second experiment was the document that has been processed through the feature selection process, while the data in the third experiment only passed the text preprocessing process without through the feature selection process. The result stated that the twitter data that have been processed using feature selection (experiment 1 and experiment 2) were able to give higher accuracy average value as much 7.8 % that the data that not through the feature selection process (experiment 3). The difference of the result was affected by the occurrences of word class like adjective, or negation in a document which was processed through the feature selection of POS.

The data in the first experiment and the second experiment have been processed using feature selection of POS, but both of the experiments have the different result. The result of the second experiment was able to give higher accuracy average value as much 7.8 % from the accuracy average value of the first experiment. That difference appeared because in the first feature of POS-1 was formed by showing the word class appearance of a document, meanwhile the feature of POS-2 which is the derivative of the feature of POS-1, was formed by showing the frequency of the word class appearance of a document.

V. Conclusion

According to the research result of sentiment analysis with feature selection process of Part of Speech (POS) and using the Latent Dirichlet Allocation (LDA) method, obtained the following conclusions.

- The result of this research showed that the document which has passed the POS-based feature process can give accuracy score higher with the difference about 7.8% than the document without feature selection process of POS.
- Sentiment classification using the feature of POS-2 was able to give highest accuracy average value of 64% with the highest accuracy on validity test with 10 fold cross validation value was 70 % in the fold 7.
- POS-based features was able to give impact toward the LDA method in increasing the better accuracy value in the classification process.
- The LDA method has good ability to conduct the sentiment classification in train document and test document used in the current research.

Acknowledgment

The first author would like to acknowledge the research funding supported by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia, in part of BPPDN scholarships 2015.

Reference

[1] E. Kontopoulos, C. Berberidis, T. Dergiades and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," *Expert Systems with Applications,* vol. 40, pp. 4065-4074, 2013.

[2] V. Ravi and K. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems,* vol. 89, pp. 14-46, 2015.

[3] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies,* 2012.

[4] R. Xia, C. Zong and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences,* vol. 181, pp. 1138-1152, 2011.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval,* vol. 2, pp. 1-135, 2008.

[6] M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naıve Bayesian, Decision Tree and KNN classification techniques," *Journal of King Saud University - Computer and Information Sciences,* vol. 28, pp. 330-334, 2016.

[7] Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in *2015 IEEE, 15th International Conference on Data Mining Workshops*, Atlantic City, NJ, USA, 2015.

[8] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal,* vol. 5, no. 4, pp. 1093-1113, 2014.

[9] A. Dinakaramani, F. Rashel, A. Luthfi and R. Manurung, "Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus," in *International Conference on Asian Language Processing (IALP)*, Kuching, 2014.

[10] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1-22, 2003.

[11] D. M. Blei, "Introduction to Probabilistic Topic Modeling," *Communications of the ACM,* vol. 55, no. 4, pp. 77-84, 2012.

[12] R. Kusumaningrum, H. Wei, R. Manurung and A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image," *Journal of Applied Remote Sensing,* vol. 8, 2014.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceeding of the National Academy of Sciences,* vol. 101, p. 5228–5235, 2004.

[14] G. Heinrich, "Parameter estimation for text analysis," 2009.

[15] R. Kusumaningrum, S. Adhy, M. I. A. Wiedjayanto and Suryono, "Classification of Indonesian News Articles based on Latent Dirichlet Allocation," in *International Conference on Data and Software Engineering (ICoDSE)*, Denpasar, 2016.

[16] I. Putri and R. Kusumaningrum, "Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia," in *International Conference on Computing and Applied Informatics*, Jakarta, 2016.