Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization

Mohammad F. A. Bashri, Retno Kusumaningrum
Department of Informatics
Universitas Diponegoro
Semarang, Indonesia
fajarainul@student.undip.ac.id, retno_ilkom@undip.ac.id

***Abstract*— Sentiment analysis is a field of study that analyzes sentiment. One method for doing sentiment analysis is Latent Dirichlet Allocation (LDA) that extracts the topic of documents where the topic is represented as the appearance of the words with different topic probability. Therefore, we need data representation in visual form that is easier to understand than text and tables. One form of data visualization is wordcloud that provides a visual representation of words frequency. This research will perform sentiment analysis from the students' comments toward a university, in this case the Universitas Diponegoro, using LDA and topic polarity wordcloud visualization. The purpose of this study is to generate the topic polarity wordcloud of the students' comments by using the best combination of parameters. The best combination is the parameter with the value of alpha 0.1, value of beta 0.1, number of topics 9, threshold 10^{-7}, and perplexity values 8.07. Such parameter combination produces 3 topics as positive sentiment and 6 topics as negative sentiment. In addition, we also compare the proposed method to several algorithms such as Naïve Bayes and Logistic Regression. The final result shows that the proposed method outperforms the Naïve Bayes and Logistic Regression in terms of F-Measure by 61%, 54%, and 56%, respectively.**

Keywords—Latent Dirichlet Allocation, topic polarity, wordcloud, sentiment analysis, data visualization

I. INTRODUCTION

Currently, opinion has a major role to every human action. When someone wants to make a decision, he needs to know what the other people's opinions are about what he will do. For example, when a person wants to buy a product from an online store, he'll look at the reviews from other users to decide whether to buy the product or not. For institutions that engaged in the field of goods and services, they need consumer's opinion regarding the goods or services they produce.

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1]. Sentiment analysis has several benefits such as product monitoring that is to monitor the level of satisfaction with the education system applied. Several methods have been implemented to analyze sentiment on Indonesian documents, such as Naive Bayes classifier [2]-[5], SVM [2][4], and Maximum Entropy [2].

One method that can be used to perform sentiment analysis is Topic Models. Topic Models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents [6]. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible for human annotation. At this time, the LDA model is one of the most popular probability topic models, and it has more comprehensive assumptions of text generation than others [7].

Representing data in visual form makes it easier to understand the content of the data rather than representing the data in the form of text or tables. Visualization helps people to understand the data that is hard to understand. In spite of the amount of the data used is very large, the data pattern can be understood quickly and easily. Data visualization conveys information in a way that is universal and simpler. One form of data visualization is wordcloud. Wordcloud is a visual representation of the occurrence frequency of words in text books or websites. Font size determines the occurrence frequency of a word: the larger the font size, the greater the word frequency is and in opposite, the smaller the font size, the smaller the word frequency is [8].

Currently, wordcloud is generated by the frequency of words occurrence in the document, but cannot accommodate the occurrence of the word related with topics. In addition, wordcloud cannot indicate the polarity of the contents of the document. Therefore, we need an alternate method to produce wordcloud by generating words that represent topics and document's polarity. This research will conduct the visualization of sentiment analysis using topic polarity worcloud visualization.

II. LDA FOR TEXT CLUSTERING

A. LDA in General

Latent Dirichlet Allocation (LDA), a probabilistic topic modelling, presents each document as a random mixture over a set of latent topics, and each topic is represented as a distribution over a vocabulary [6]. LDA is divided into two processes, namely LDA as a generative process and LDA as an inference process. LDA as the generative process is used to generate the document in which we already know the values of word-topic probability (φ_k) and topic proportion for each document (θ_d). On the other hand, the LDA as the inference

ISBN: 978-1-5090-4911-0 (c) 2017 IEEE

process is used to determine the value of latent variables, i.e. the probability of word-topic and topic proportion of known documents [9].

The equation for calculating the probability of word-topic and topic proportion is as follows [10]:

$$\varphi_k = p(w=t|z=k) = \frac{n_{t,k} + \beta_t}{\sum_{t=1}^{V} n_{t,k} + \beta_t} \quad (1)$$

$$\theta_d = p(z=k|d) = \frac{n_{d,k} + \alpha_k}{\sum_{k=1}^{K} n_{d,k} + \alpha_k} \quad (2)$$

This research will use LDA as the inference process using Gibbs sampling algorithm. The basic idea of Gibbs Sampling algorithm is to update each of the variables respectively based on the conditional probability of all other variables, although the joint probability distribution is unknown [10].

B. Evaluation

This research will use perplexity as evaluation process. Each iteration in the Gibbs Sampling will calculate the value of perplexity. The purpose of calculating the value of perplexity is to provide the stop conditions in the process of Gibbs Sampling. Perplexity equation used is as follows [11].

$$Perplexity(\vec{w}|m) = exp\{-\frac{\sum_{d=1}^{M} log\, p(\vec{w}|m)}{\sum_{d=1}^{M} N_d}\} \quad (3)$$

In addition, we perform a comparison of various methods based on the F-measure as an evaluation metric with the following formula.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

III. METHODOLOGY

This section will discuss the research methodology conducted. Fig. 1 describes the methodology that is used in this study. The first step taken is collecting data by spreading online form. Data comments are collected from students toward their university, then preprocessing is performed on the collected data, including tokenization, stopwords removal, stemming, and forming bag of words. Tokenization is letters alternation to lowercase, symbols removal, and row of words separation into array of word. Next, we perform stopwords removal. In this study, adjective (e.g. good, bad, new, old) and negative words (e.g. not, do not) are not removed because they will be used in the process of establishing topic polarity wordcloud. Next, we perform stemming process that changes words to their root word. In this study, we use Sastrawi as stemmer library that implements Nazief-Adriani algorithm, improved with Confix Stripping (CS) algorithm [12], improved with Enhanced Confix Stripping (ECS) [13], and improved again with Modified ECS [14]. Last stage in preprocessing is forming bag of words. The purpose of preprocessing is to convert unstructured text to structured text. In addition, preprocessing is aimed to keep the words or phrases that are considered important in text mining.

Once bag of words is formed, the next process is calculating the LDA Collapsed Gibbs Sampling. At each iteration of LDA Collapsed, Gibbs Sampling will calculate the value of perplexity. Iteration will stop if it meets two conditions; the iteration reaches the maximum number of iterations and the difference value of perplexity iteration *i* with iteration *(i-1)* is less than a predetermined threshold value.

The next stage is forming topic polarity wordcloud. A probability value word-topic (PWZ) is then performed to sequence largest to smallest and top *n* word with the largest PWZ value for each topic (in this study *n* is 15). These *n* words will be used to visualize topic polarity wordcloud. Polarity forming is performed by checking each word into sentiment lexicon at Table I.

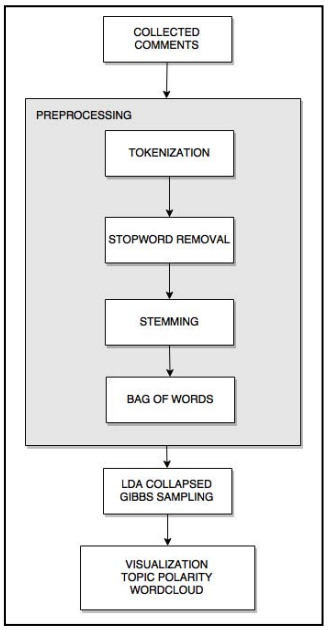

Fig. 1. Research methodology

TABLE I. SENTIMENT LEXICON

Positive	**Negative**
large, new, clean, beautiful, comfortable, cool, fine, neat, safe, airy, spacious, beautiful, strong, elegant, modern, bright, aromatic, full, fragrant, mild, quiet, nice, clear, honest, friendly, fair, firm, polite, correct, creative, clever, industrious, enterprising, young, sociable, cheap, low, fast, easy, smooth, satisfied	small, ancient, dirty, bad, hot, ugly, danger, sultry, narrow, fragile, dilapidated, rundown, full, cracked, broken, spooky, dark, foul, distant, empty, shabby, long-time, random, noisy, torn, rowdy, chaotic, obstreperous, worn, nasty, sketchy, lying, arrogant, rude, wrong, stupid, lazy, angry, confused, skipping, tired, old, expensive, high, slow, difficult, jammed, hassle, no, not, do not, as yet, less

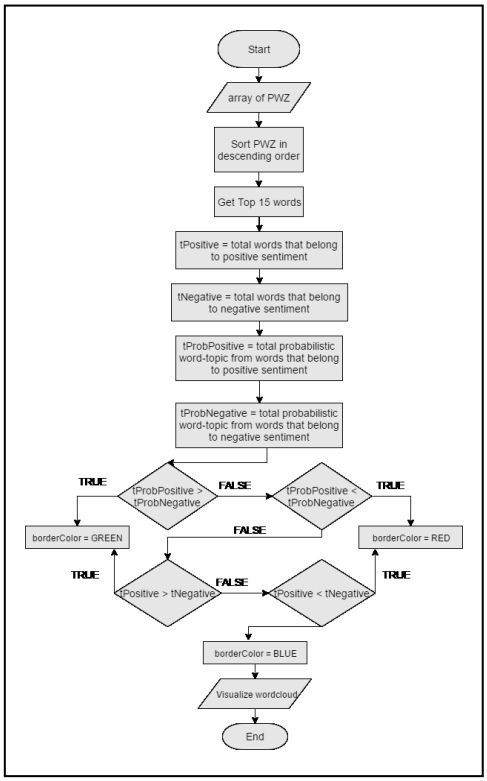

Fig. 2. Procedure for visualize topic polarity wordcloud

Fig. 2 describes the workflow of topic polarity wordcloud process. If total value of word-topic probability of words that marked as positive sentiment (tProbPositive) is greater than total value of word-topic probability of words that marked as negative sentiment (tProbNegative), wordcloud will be printed as positive sentiment (printed with green border), and in contrast, wordcloud will be printed as negative sentiment (printed with red border). Another condition is if tProbPositive is equal to tProbNegative, it will count the total number of words that marked as positive sentiment (tPositive) and words that marked as negative sentiment (tNegative). If tPositive is greater than tNegative, wordcloud will be printed as positive sentiment (printed with green border). On the other hand, if tNegative is greater than tPositive, wordcloud will be printed as negative sentiment (printed with red border). Last condition is if tPositive is equal to tNegative, wordcloud will be printed as neutral sentiment (printed with blue border).

IV. Experiments and Results

This section will describe the experimental setup, experimental scenarios, and the results and analysis.

A. Experimental Setup

In this study, we use 690 data comments collected from online form. Each comment performs preprocessing, i.e. tokenization, stopwords removal, stemming process, and forming bag of words. Next step is calculating LDA Collapsed Gibbs Sampling. Last step is forming topic polarity wordcloud.

This experiment is implemented using PHP programming under Windows 10 Pro 64 bit and the following hardware spesifications :

- Intel(R) Core(TM) i3-2350M CPU @2.30GHz
- 4 GB Memory (RAM)
- 500 GB HDD

B. Experimental Scenarios

The scenario of the research is to determine the best combination of parameters for sentiment analysis using LDA, that is determining the best parameter combination by calculating the average value of perplexity for each combination of parameters. Combination of parameters that is used in this study is the combination with the value of alpha 0.1, 0.01, 0.001, value of beta 0.1, 0.01, 0.001, number of topic 2, 3, 4, 6, 9, and value of threshold 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}. Therefore, the total is 225 combinations of parameters. Once the best parameter combination is determined, topic polarity wordcloud visualization is performed. In this experiment, we implement several values of alpha and beta since both values influence the resulted values of φ_k and θ_d. The smaller value of alpha causes fewer numbers of topics per document, whereas the smaller value of beta causes fewer numbers of words per topic. In other words it results sparse vector of φ_k and θ_d, which influences LDA's performance.

C. Results and Analysis

Fig. 3 to Fig. 5 show a graph of the results of the average value of perplexity for the whole combination of parameters. Based on the graph, we can see the smallest perplexity value that is obtained by a combination of parameters K-5 with value of alpha 0.1, value of beta 0.1, the number of topics 9, value of threshold 10^{-7}, and the perplexity values is 8.07. Therefore, the best combination of parameters is used as an input in the form of topic polarity wordcloud visualization. Results of topic polarity wordcloud visualization can be seen in Fig. 6.

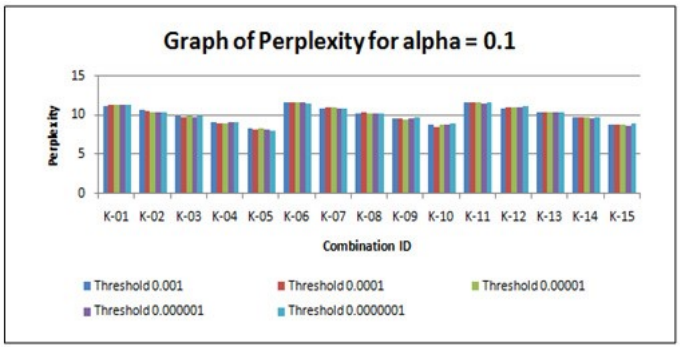

Fig. 3. Graph of perplexity for alpha = 0.1

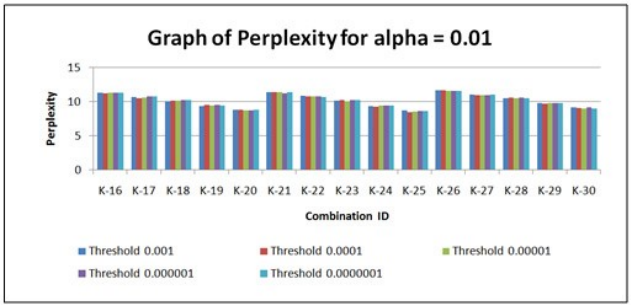

Fig. 4. Graph of perplexity for alpha = 0.01

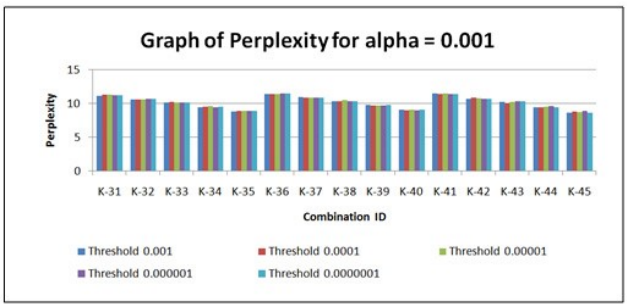

Fig. 5. Graph of perplexity for alpha = 0.001

Based on Fig. 6, we can see the comments leave mostly negative comments as from the total 9 topics, 6 topics are negative sentiments. Furthermore, from each wordcloud result, we can see what things that are discussed in this topic. For example, in topic 4, we know the facilities in the university (Universitas Diponegoro) are not satisfying, and so topic 4 is a negative sentiment.

In addition, we also compare the proposed method to several algorithms such as Naïve Bayes and Logistic Regression. The final result shows that the proposed method outperforms the Naïve Bayes and Logistic Regression in terms of F-Measure by 61%, 54%, and 56%, respectively. The proposed method gives the best result since this method optimally classifies negative sentiments. It can be seen based on the variety representation of negative sentiments, i.e. 6 topics. On contrary, the positive sentiments are only represented by 3 topics. This condition occurs since people more expressive in giving negative sentiments, thus a single comment consists of tens or even hundreds of words.

V. Conclusion

From the research that has been done, it can be interfered that the best combination of parameters for sentiment analysis is the combination of parameters with the value of alpha 0.1, value of beta 0.1, number of topics 9, and value of threshold 10^{-7}. Such parameter combination produces 3 topics as positive sentiment and 6 topics as negative sentiment with the topics being talked are facilities, bureaucratic, parking lots, and tuition. In addition, we also compare the proposed method to several algorithms such as Naïve Bayes and Logistic Regression, in which the proposed method outperforms the Naïve Bayes and Logistic Regression in terms of F-Measure by 61%, 54%, and 56%, respectively.

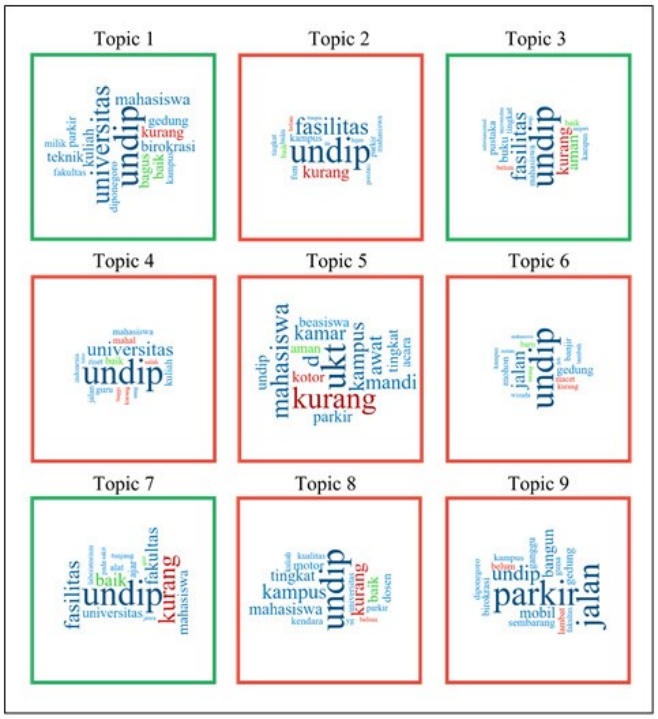

Fig. 6. Topic polarity wordcloud results

Acknowledgement

The authors would like to acknowledge the research funding supported by Universitas Diponegoro under the grant of Research for International Scientific Publication – Year 2017 (number 276-36/UN7.5.1/PG/2017).

References

[1] B. Liu, Sentiment Analysis and Opinion Mining.: Morgan & Claypool Publishers, 2012.

[2] Franky and R. Manurung, "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews", in Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications (ICACIA), Depok, Indonesia, 2008.

[3] G. A Buntoro, T. B Adji, and A. E. Purnamasari, "Sentiment Analysis Candidates of Indonesian Presiden 2014 with Five Class Attribute", International Journal of Computer Applications (0975 – 8887), vol. 136 No 2, February 2016.

[4] P. Aliandu, "Sentiment Analysis on Indonesian Tweet", in The Proceedings of The 7th ICTS, Bali, 2013, pp. 203 - 208.

[5] Y. E. Soelistio and M. R. S. Surendra, "Simple Text Mining for Sentiment Analysis of Political Figure using Naive Bayes Classifier Method", in The Proceedings of The 7th ICTS, Bali, 2013, pp. 99-104.

[6] D. M. Blei, "Probabilistic Topic Models", Communications of the ACM, vol. 55, pp. 77-84, 2012.

[7] Z. Liu, "High Performance Latent Dirichlet Allocation for Text Mining", Brunel University London, 2013.

[8] F. Miley and A. Read, "Using word clouds to develop proactive learners", Journal of the Scholarship of Teaching and Learning, vol. 11 No.2, pp. 91-110, 2011.

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

[10] 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.

[11] S. Chen and Y. Wang, "Latent Dirichlet Allocation", 2007

[12] J. Asian, "Effective Techniques for Indonesian Text Retrieval", 2007.

[13] A. Z. Arifin, I P. A. K. Mahendra, and H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language" , 2009.

[14] S. D. Tahitoe and D. Purwitasari, "Implementasi Modifikasi Enhanced Confix Stripping Stemmer untuk Bahasa Indonesia dengan Metode Corpus Based Stemming", 2010.