TELKOMNIKA, Vol.16, No.4, August 2018, pp. 1752~1759
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v16i4.8194

WCLOUDVIZ: Word Cloud Visualization of Indonesian News Articles Classification Based on Latent Dirichlet Allocation

Retno Kusumaningrum*, Satriyo Adhy, Suryono
Department of Informatics, Universitas Diponegoro, Jl. Prof. Soedarto, SH Tembalang, Semarang, Central Java, Indonesia
*Corresponding author, e-mail: retno@live.undip.ac.id

Abstract

Latent Dirichlet Allocation (LDA) is a widely implemented approach for extracting hidden topics in documents generated by soft clustering of a word based on document co-occurrence as a multinomial probability distribution over terms. Therefore, several visualizations have been developed, such as matrices design, text-based design, tree design, parallel coordinates, and force-directed graphs. Furthermore, based on a set of documents representing a class (category), we can implement classification task by comparing topic proportion for each class and topic proportion for the testing document by using Kullback-Leibler Divergence (KLD). Therefore, the purpose of this study is to develop a system for visualizing the output of LDA as a classification task. The visualization system consists of two parts: bar chart and dependent word cloud. The first visualization aims to show the trend of each category, while the second visualization aims to show the words that represent each selected category in a word cloud. This visualization is subsequently called WCloudViz. It provides clear, understandable and preferably shared the result.

Keywords*: Latent dirichlet allocation, Topic modeling, News articles classification, Data visualization, Word cloud*

1. Introduction

Latent Dirichlet Allocation, also known as LDA, is one of topic modeling algorithms. In LDA, each document is represented as a random mixture over a set of latent topics and each of the topic is represented as a distribution over a vocabulary [1]. Until this moment, LDA has been widely implemented in various domains, such as text document [2]–[8], image domain [9]–[14], music domain [15], [16], and etc.

LDA can be implemented for classification task as well as clustering task. As clustering task, the output of LDA is a set of topics in which each topic is represented by a set of words and its probability $(\varphi_{k,t})$. This LDA's output is a corpus-level output. In addition, LDA also results an output at document-level, i.e. topic proportion for each document (θ_d). Furthermore, when we have a set of documents representing a class (category), we can compute topic proportion for each class (θ_c). By implementing Kullback-Leibler Divergence, we can find the most similar distribution between topic proportion for each class and topic proportion for testing document based on the smallest value of KLD [12].

In accordance with clustering tasks, there are several visualizations that have been developed. The first design is matrices design in which rows correspond to terms and columns to topics. A developed matrices-based system is Termite [17]. The second design is text-based design [18]. This visualization exploits the use of text to illustrate topic as well as the use of text to explain the represented words for each topic, related documents for each topic, and related topics for each topic. Based on the selected document, we can also find the document's related topics, document's text, and document's related documents.

The third design is tree design. This design is a popular design for visualizing LDA's output such as relationship-enriched visualization [19], Diachronic visualization [20], LDAExplore [21], and etc. LDAExplore also implements parallel coordinates to show topic proportion for each document. The first vertical coordinate is coordinate for displaying

Received December 2, 2017; Revised January 24, 2018; Accepted July 3, 2018

document's ID, while second vertical coordinate and so on is for representing values of topic proportion (distribution) for each topic. A line is drawn to connect the document's ID and related topic proportions for each remaining vertical line. The last design is force-directed graphs, such as relationship-enriched visualization [19] and Diachronic visualization [20]. In addition, there are some visualizations that have been developed for broader topic (information retrieval) such as map view, tree view, and bubble view which are implemented in [22], graph-view [23], heat map, interactive stream graph, and context focus which are implemented in [24].

However, those visualizations are focused on visualizing word-topic distribution produced by LDA. Therefore, the aim of this study is to develop a system for visualizing output of LDA as classification task. We implement it against the classification of Indonesian news articles. This is shown to reduce the drawbacks of Indonesian online news presentations, i.e. repetition of news contents and news splitting into some headlines. It is known as yellow journalism which gives the negative impact for readers, that are, a shift of citizen characters arising from excessive news report such as the outspread of "Alay" culture, the diversion of political issues, common corruption cases, etc. This classification and visualization system is expected to be used to know the news trends and the words that represent them for each category. The system is subsequently called as WCloudViz.

In section 2, we provide the research method related to the design consideration including task analysis, data and backend, as well as prototype design. The explanation about result and discussion will be presented in section 4. Finally, in section 5, we conclude this study with what we perceive to be the future work.

2. Research Method

As explained in previous section, WCloudViz is aimed to visualize the output of LDA as classification task. Therefore, there are three tasks, which form WCloudViz. The first task is identifying news trend for each class (category), and visualizing it as bar chart is the second task. The last task is word cloud generation of word representing for each selected category. The outline of WCloudViz formation is depicted in Figure 1. The following sub-sub sections explain each task in detail.

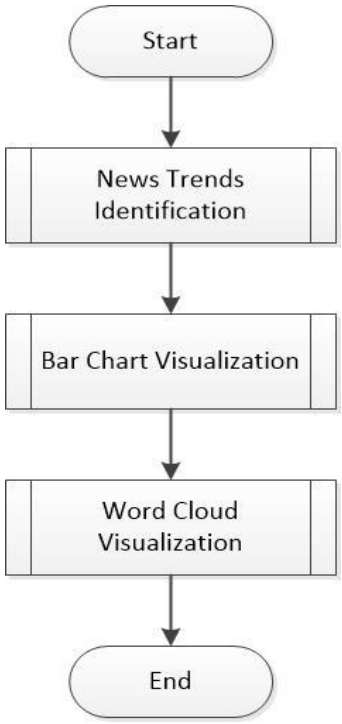

Figure 1. The outline of WCloudViz formation

2.1. News Trend Identification

Since we aim to visualize the output of LDA as classification task, the process of news trend identification will be conducted based on LDA classification model. The detail about how to generate the LDA-based classification model for news articles classification can be read in Ref. [25]. Moreover, the procedure of news trend identification described in Figure 2.

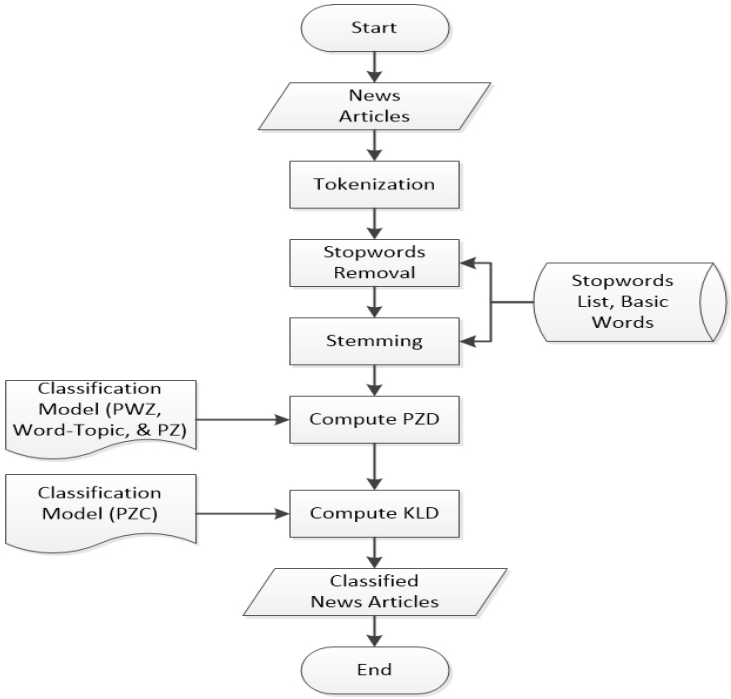

Figure 2. Procedure of news trend identification based on LDA classification model

Tokenization is a process to eliminate characters that are likely to be insignificant for information extraction, such as white space and punctuation marks. Afterwards, the second process is stopword removal, implemented for eliminating words that are likely to be insignificant for information extraction. In this research, there are 758 words in Indonesian that are considered as stopwords. Table 1 shows example of Indonesian stopwords and its English translation for each word.

Table 1. Applications in Each Class

Part of Speech	Stopwords in Indonesia	English Translation
Pronoun	saya, kamu, mereka, kita, ini, itu	I, you, they, we, this, that
Adverb	sekarang, agak	currently, some
Preposition	setelah, sebelum, kepada, dalam, atas, bawah	after, before, to, in, on, under
Conjunction	dan, karena	and, because

Stemming, the last text processing, is a process to derive each word from their root form. This research implements Sastrawi-stemmer library for stemming process. Sastrawi-stemmer is incrementally built. As a base, it implements Nazief Nazief-Adriani algorithm. Subsequently, it is improved by Confix Stripping (CS) algorithm and Enhanced Confix Stripping (ECS) algorithm. The final improvement is a modification of Modified ECS. The selection of Sastrawi-stemmer library for stemming process is based on several issues, namely it overcomes overstemming problem by using basic words dictionary, it overcomes the understemming problem by adding rules, as well as it is able to stem compound words correctly.

On the other hand, we already have the classification models including PWZ (word-topic distribution) values and PZ (topic distribution) values. Both values are in corpus level. In addition, we also have a set of words that represent each topic. Based on the model, we compute the PZD (topic proportion for each document) for each unseen news article by using the following formula.

$$\theta_{d,k} = p(z=k) * \sum_{i=1}^{N} \varphi_{k,t}^{(i)} \quad (1)$$

Where $\theta_{d,k}$ is equal to PZD, the probability of k-th topic in d-th document, $p(z=k)$ is the probabilityof topic z is equal to k, and $\sum_{i=1}^{N} \varphi_{k,t}^{(i)}$ is the sum of i-th term in the corresponding document equal to term t and assigned as topic k, while N is the number of term in the corresponding document.

Afterwards, we can compute the similarity distribution between pre-defined PZC (topic proportion for each class) and pre-computed PZD (resulted based on Equation 1) based on Kullback-Leibler Divergence (KLD). Let $P = \{p_i\}$ be the topic proportion of the unseen document and $Q = \{q_i\}$ be the topic proportion of each class; then the KLD can be computed by using the following formula.

$$KLD = \frac{DPQ + DQP}{2} \quad (2)$$

$$DPQ = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \quad (3)$$

$$DQP = \sum_i q_i \log\left(\frac{q_i}{p_i}\right) \quad (4)$$

Finally, the most similar distribution between topic proportion for each class and topic proportion for unseen document is obtained based on the smallest value of KLD.

2.2. Bar Chart Visualization

Bar chart is visualized based on the number of documents for each class (category) over the whole submitted new articles. Figure 3 shows the detailed procedure for generating bar chart based on output of the previous task.

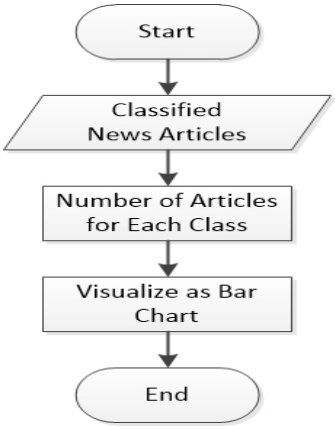

Figure 3. Procedure of bar chart visualization

According to the Figure 3, the input of this process is a collection of news articles that have been classified from previous step into six categories (class) including economic, tourism, criminal, sport, and politic. The next process is computing a number of articles for each category. Subsequently, plot the bar chart with x-axis represents six news categories and y-axis represents number of articles for each category.

2.3. Word Cloud Visualization

Based on the selected class (category), we can generate word cloud of represented words based on the following procedure as shown in Figure 4. We use the initial value of the biggest font size that is 150, and the number of words that are 20. Both values can be changed in accordance with the requirements.

The first step of word cloud visualization is counting frequency of word occurrence in all documents of selected category. Get maximum frequency and get top-20 most occurred words. Subsequently, compute the font size of all obtained words based on the following formula. We use default value of *MaxFont* is 150 pt.

$$FontSize = \left\lfloor \frac{Frequency}{MaxFrequency} \times MaxFont \right\rfloor \quad (5)$$

As an exception and to provide the convenience of the reader, then the default value of the smallest FontSize is set at 14 pt despite FontSize calculation results based on the formula (5) gives the results of less than 14 pt.

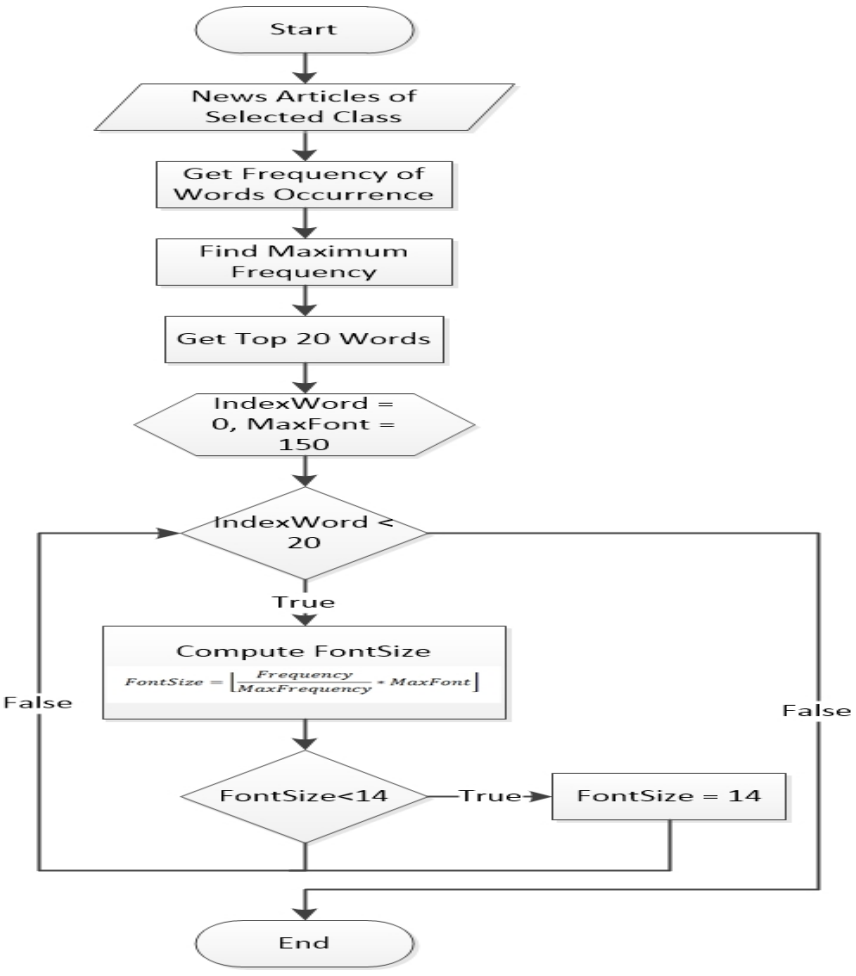

Figure 4. Procedure of word cloud visualization

3. Results and Analysis

In this section, it is explained the results of research and its comprehensive discussion about interaction and visualization. For the demonstrator of WCloudViz, we employ the news articles data in Indonesian. These articles are obtained from local newspaper, i.e. Jawa Pos - Radar Semarang. Articles are saved in .txt format for each headline.

3.1. Prototype Design

As explained in previous section, there are three tasks which form WCloudViz, consisting of news trend identification, bar chart visualization, and word cloud visualization. Two first tasks are represented in one interface as shown in Figure 5. The third task is represented in another interface, which is loaded after the user clicks one of the bars in the chart. The interface can be seen in Figure 6.

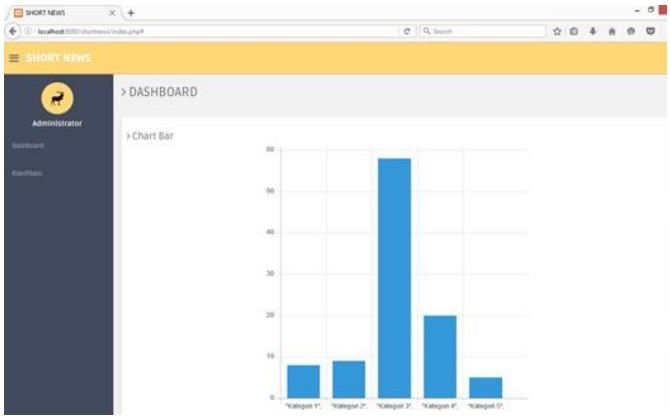

Figure 5. Bar chart visualization with embedded news trend identification

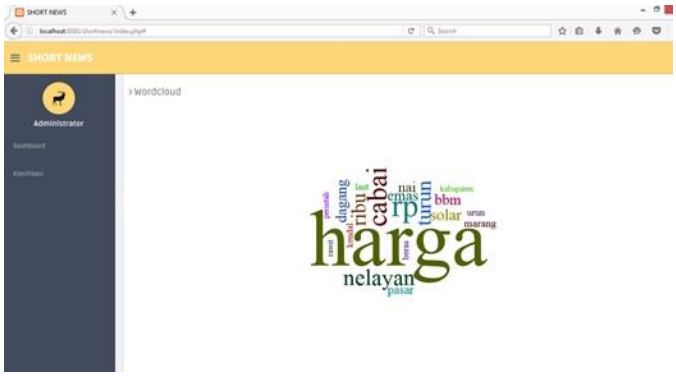

Figure 6. Word cloud visualization of selected category

3.2. Interaction and Visualization

Users can gain insight into the news categories that have become the reporting focus through the bar chart visualization, which means that the news category with most report has the highest bar chart and vice-versa. In addition, users can also obtain information on the topic of each news category. Each topic is represented as a word cloud, i.e. word-occurrence in a set of documents. Word cloud will be displayed when the user has selected one of news categories. Hereafter, it can be referred as the dependent word cloud.

For example, Figure 6 shows the economic category and it is represented by 10 words having high frequency, such as “harga” (price), “cabai” (chili), “Rp” (Indonesian currency), “turun” (decrease), “ribu” (thousand), “BBM” (fuel), “solar” (diesel), “nelayan” (fisherman), “dagang” (trade), and “emas” (gold). Furthermore, users can also draw conclusion based on the

word cloud, i.e. the current economic news headlines are widely reporting about the news related to the increasing chili prices up to hundreds of thousands rupiahs, the declining rupiahs, and the price of fuel, diesel fuel, as well as gold.

In accordance with the previous explanations, there are some points that can be accomplished by word cloud. First, word cloud provides clarity in identifying trends and pattern that will otherwise be unclear or difficult to see in a tabular format. Second, word cloud is easier to understand since it is more visually appealing than the textual format of data. The bigger the words represent, the more widely-reported the headlines are. The last point, word cloud is more likely to be shared since it presents pictorial representation of the data. Unfortunately, the resulting word cloud has not been depicted in accordance with the related topic or theme, thus it is less impactful.

4. Conclusion

WCloudViz is a visualization system based on LDA. It consists of two parts: bar chart and dependent word cloud. The first visualization aims to show the trend of each category, while the second visualization aims to show the words that represent each selected category in word cloud. This visualization provides clear, understandable, and preferably shared result. The strategic future work shall form each word cloud in accordance with related topic or theme, thus it can be more impactful and insightful.

Acknowledgement

The authors would like to acknowledge the research funding supported by Universitas Diponegoro under the grant of Research for International Scientific Publication (number 1052-36/UN7.5.1/PG/2016, year 2016)

References

[1] DM Blei, AY Ng, MI Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 2003; 3: 993–1022.
[2] DM Blei. Probabilistic Models of Text and Images. University of Carolina, Berkeley, 2004.
[3] V Ravi. Legal Documents Clustering using Latent Dirichlet Allocation. *International Journal of Applied Information Systems*. 2012; 2(6): 34–37.
[4] I Bíró, J Szabó, AA Benczúr. *Latent dirichlet allocation in web spam filtering.* Proc. 4th Int. Work. Advers. Inf. Retr. web, 2008: 29–32.
[5] W Sriurai. Improving Text Categorization by using a Topic Model. *Adv. Comput.*, 2011; 2(6): 21–27.
[6] I Bíró. Document Classification with Latent Dirichlet Allocation. Unpubl. Dr. Diss. Eotvos Lorand …, 2009.
[7] L Bolelli, Ş Ertekin, CL Giles. Topic and trend detection in text collections using latent dirichlet allocation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 2009; 5478 LNCS: 776–780.
[8] XH Phan, LM Nguyen, S Horiguchi. *Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections*. Proceeding 17th Int. Conf. World Wide Web - WWW '08, no. January 2008: 91–100.
[9] K Barnard, P Duygulu, D Forsyth, N de Freitas, DM Blei, MI Jordan. Matching Words and Pictures. *J. Mach. Learn. Res.*, 2003; 3: 1107–1135.
[10] C Wang, DM Blei, L Fei-fei. *Simultaneous image classification and annotation*. IEEE Conference on Computer Vision and Pattern Recognition. 2009:1903–1910.
[11] D Bratasanu, I Nedelcu, M Datcu. Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications. *IEEE J. Sel. Top. Earth Obs. Remote Sens.* 2011; 4(1): 193-120.
[12] R Kusumaningrum, H Wei, R Manurung, A Murni. Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image. *J. Appl. Remote Sens.* Jan. 2014; 8(1): 83690-1-18.
[13] R Kusumaningrum, HM Manurung, AM Arymurthy. CIELab Color Moments: Alternative Descriptors for LANDSAT Images Classification System. *INKOM*. 2014; 8(2): 109–114.
[14] M Liénou, H Maître, M Datcu. Using Latent Dirichlet Allocation. *IEEE Geosci. Remote Sens. Lett.* 2010; 7(1): 28–32.
[15] D Hu. Latent Dirichlet Allocation, Text, Images, & Music."
[16] R Cai, C Zhang, C Wang, L Zhang, W Ma. MusicSense: Contextual Music Recommendation using Emotional Allocation Modeling. *Emotion*. 2007: 553–556.

[17] J Chuang, CD Manning, J Heer. *Termite: Visualization Techniques for Assessing Textual Topic Models.* in Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI), 2012: 74–77.
[18] AJB Chaney, DM Blei. *Visualizing Topic Models*. in Proceedings of the Sixth International Conference on Weblogs and Social Media. 2003.
[19] A Smith, J Chuang, Y Hu, J Boyd-graber, L Findlater. *Concurrent Visualization of Relationships between Words and Topics in Topic Models.* in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014: 79–82.
[20] J Lamirel, N Dugu, P Cuxac. *Performing and Visualizing Temporal Analysis of Large Text Data Issued for Open Sources: Past and Future*. in Proceedings of the 12th International Conference on Beyond Databases, Architecture and Structures. 2016: 1340846.
[21] A Ganesan, K Brantley, S Pan J Chen. LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation. CoRR, 2015.
[22] MH Mamoon, HM El-Bakry, AA Salama. Visualization for Information Retrieval based on Fast Search Technology. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI).* December 2013; 1(4): 140-156.
[23] F Nadirman, A Ridha, Annisa. Searching and Visualization of References in Research Documents. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. June 2014; 12(2): 447-454.
[24] MNP Ma'ady, CK Yang, RP Kusumawardani, H Suryotrisongko. Temporal Exploration in 2D Visualization of Emotions on Twitter Stream. *TELKOMNIKA (Telecommunication Computing Electronics and Control).* February 2018; 16(1): 376-384.
[25] R Kusumaningrum, S Adhy, MIA Wiedjayanto, Suryono. *Classification of Indonesian News Articles based on Latent Dirichlet Allocation.* in Proceedings of the 3rd International Conference on Software and Data Engineering, 2016.