ICIC Express Letters
Volume **11**, Number **10**, October **2017**

ICIC International ©2017 ISSN 1881-803X
pp. 1579–1584

A TOPIC MODELING METHOD BASED ON THE AFFECTIVE FILTER

Lin Wang[1], Hyunjin Jeon[2], Kangcheol Lee[2], Mungyeong Choe[2]
Jaehyun Park[2,*], Byung Cheol Lee[3], Jangwoon Park[3] and Heejin Jeong[4]

[1]Department of Library and Information Science
[2]Department of Industrial and Management Engineering
Incheon National University
119 Academy-ro, Yeonsu-gu, Incheon 22012, Korea
{ wanglin; onue95; rkdcjf0655; mgyeong }@inu.ac.kr; *Corresponding author: jaehpark@inu.ac.kr

[3]Department of Engineering
Texas A&M University – Corpus Christi
6300 Ocean Dr., Corpus Christi, TX 78412, USA
{ Bruce.Lee; Jangwoon.Park }@tamucc.edu

[4]Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave., Ann Arbor, MI 48109, USA
heejinj@umich.edu

Received April 2017; accepted July 2017

Abstract. *Documents that convey information are made up of many words. The topic modeling technique is used to identify the distribution of words and to recommend and derive a word that can be used as a main word. Typically, there are algorithms for the identification of word distribution in documents, such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis, and these algorithms are utilized in various fields. However, because the frequencies of words take an important role in documents, these algorithms possess the limitation of not being able to derive the main word if its frequency in a document is low. Thus, the objective of this paper is to present an affective filter, using the circumplex model, which is designed to overcome the abovementioned limitations. The proposed framework was verified through case studies that were about questions used in Korean university entrance examinations. Thus, there was no significant difference between the existing LDA and the developed recommendation method in terms of perceived reliability. Although we did not confirm statistical superiority, the proposed framework is expected to be used as a basic data to improve and develop the topic modeling technique. It is also expected to be applied in practical fields such as analyses of literature and social networking service comments.*
Keywords: Latent Dirichlet allocation, Topic model, Circumplex model, Perceived trust

1. **Introduction.** Quite recently, with the advent of the Fourth Industrial Revolution, information has become more important than ever before, rather than physical production facilities [1]. As everything is digitized in this situation, information is expanding quantitatively and finding important information is a challenge in any field.

Topic modeling is recognized for its utility and value in that it finds and recommends important information from a given data. Topic modeling is a technique that allows one to estimate the keywords of a document based on the distribution of words in the document. Typically, there are latent Dirichlet allocation (LDA), probabilistic latent semantic analysis (pLSA) and biterm topic model (BTM) for such estimation [2,3]. Such automated algorithms are applied not only to document analysis but also to various fields in which a large amount of data, such as image processing and biological information analysis, exist [4,5]. However, studies have focused more on the frequency of words in a document. Therefore, there are limitations to retrieve keywords via these algorithms because the frequency of the words has a significant influence on the outcome. Because of this disadvantage, topic modeling algorithms such as LDA and pLSA are not actively used in the field of information engineering. Instead, a lot of new compound models based on existing topic modeling algorithms have been widely developed [6].

The main purpose of this study is to develop a method that suggests keywords by considering not only the frequency of words but also affective characteristics. To this end, we propose a sentimental model based on LDA, which is the most popular method in topic modeling. The affective model derived from this study is expected to easily summarize user opinions and preferences with regard to service planning and research.

2. **Affective Model.** Affect can be defined as the perceived image/impression of affective quality with non-task related goals, while emotion can be defined as a neurophysiological state which is consciously accessible [7-9]. Research that derives emotion models has been actively conducted to analyze cognitive function [9,10]. In the case of the circumplex model by Russell and pleasure-arousal-dominance emotional state model by Mehrabian [11,12], it is remarkable that it is widely used in topic modeling despite some criticism. This model assumes that human emotions can be classified into three axes: pleasure, arousal, and dominance. Pleasure represents how delightful, arousal represents how excited and dominance indicates how dominant.

Various researchers have analyzed vocabularies based on the premise that the circumplex model corresponds to the actual cognitive model. Bradley and Lang [13] analyzed the distribution of pleasure, arousal, and dominance for over 1,000 English words. In this study, an affective filter (AF) was constructed using the database. The following Table 1 shows some of the databases of vocabularies presented in Bradley and Lang [13].

Table 1. An example of valence, arousal, and dominance features of words

Description	Word No.	Valence Mean (SD)	Arousal Mean (SD)	Dominance Mean (SD)	Word Frequency
Advantage	629	6.95 (1.85)	4.76 (2.18)	6.36 (2.23)	73
Beautiful	654	7.60 (1.64)	6.17 (2.34)	6.29 (1.81)	127
Blue	544	6.76 (1.78)	4.31 (2.20)	5.63 (1.64)	143
Concentrate	78	5.20 (1.28)	4.65 (2.13)	4.97 (1.75)	11
Delayed	721	3.07 (1.74)	5.62 (2.39)	3.64 (1.94)	25

3. **Method.** As an analyzed data, an English paragraph of sample questions from college entrance examination in Korea between 2012 and 2017 is selected. Each paragraph consists of 150 words on average and each paragraph contains a subject in the form of a sentence. The following remarks are noteworthy: 1) it can be said that the topic sentence has been verified as it is used in college entrance examination, 2) it was adopted as an analysis subject in that various topics were treated in a balance. Finally, 46 paragraphs were selected except when the derived words are not included in the Bradley and Lang database after pre-application of the LDA.

As a method of recommending keywords, we set up three existing LDA, AF and hybrid LDA and AF methods. First, in the case of LDA, a parameter that is considered to summarize the paragraph is applied through the preliminary experiment. The setting values are as follows: 2 topics, 5 words per topic and 5000 repetitions, using the Gensim package that works with Python. Second, AF was developed by constructing an affective model. To do this, three people who had no problem in the topic summary and cognitive ability participated in the survey to select the top three words that indicate 'The word that best describes the subject'. Then, the characteristics of pleasure, arousal, and dominance

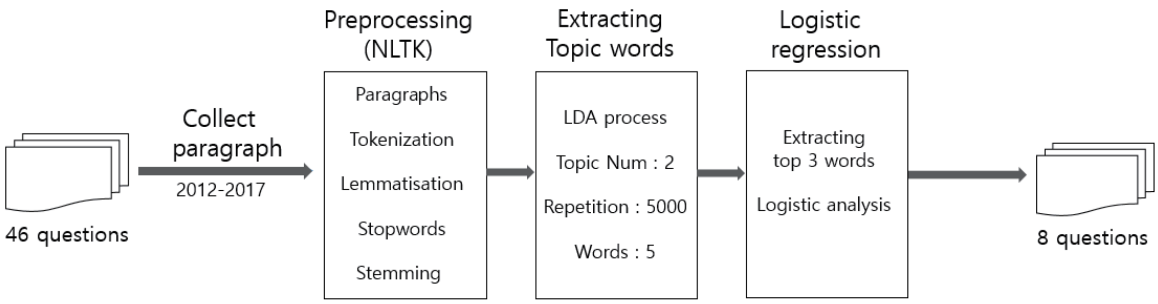

Figure 1. Document extraction process

of the words extracted from each set of paragraphs were constructed using a logistic regression model. We then applied the regression model to eight paragraphs that were not included in the training set and suggested the vocabulary as an AF vocabulary. Third, in case of hybrid method, it shows all words recommended in LDA and AF by highlighting them. For this purpose, the LDA recommended word was highlighted in the ordinary style and the AF recommended word was highlighted in the boldface style.

A total of eight paragraphs and lists of recommendations for each method were used to verify the three recommendations. Twenty-four participants (22 men and 2 women with the mean age of 24.4 and standard deviation of 1.2 years) participated in the topic summarization and cognitive abilities to confirm the reliability of the recommendations.

The experiment was carried out in the following way. First, participants identified one paragraph and a sentence topic. Second, they checked the recommended list of words by the LDA, AF and hybrid methods and rated them in terms of reliability. Eight paragraphs per participant were presented in different order using the Latin square balancing method because once the credibility of the recommendation drops, it can affect the paragraph evaluation later. After evaluating all the recommended words in each of the eight paragraphs, we asked an additional rating of 'how reliable they are' in debriefing.

Reliability is used as an index of evaluation. In this paper, it was presumed that automatic recommendation suggestion is a kind of artificial intelligence. In addition, an additional analysis of the trust problem on artificial intelligence is included in the evaluation index. All the ratings ranged of 0 to 10.

4. **Result.** A total of eight paragraphs were derived for each of the three methods (i.e., LDA, AF and hybrid) as shown in Table 2. Note that recommendations are not complete words because of a stemming process in a topic modeling. On the other hand, the regression model that is derived for the recommended word using AF is shown in Table 3. In particular, the arousal variable was found to be highly significant. A variable affective relation, which was subjectively rated by 3 people, indicates whether the paragraph is considered to be emotional or not.

The accuracy of classification based on the results of the above mentioned three evaluators was as high as 76.4% and the result of the Hosmer and Lemeshow test which was conducted to test goodness of fit was 0.615 ($p > .05$) as highly fit.

The evaluation results for the recommended language are shown in Figure 2. The reliability scores for LDA, AF and hybrid method are 5.2, 2.5, and 5.2, respectively. The results of the ANOVA showed that there was a significant difference in the reliability scores according to the three methods ($\alpha = 0.05$). As shown in Figure 2, there was no significant difference between LDA and hybrid method in the Student-Newman-Keuls test.

Table 2. An example of paragraph

Script	Many disciplines are better learned by entering into the doing than by mere abstract study. This is often the case with the most abstract as well as the seemingly more practical disciplines. For example, within the philosophical disciplines, logic must be learned through the use of examples and actual problem solving. Only after some time and struggle does the student begin to develop the insights and intuitions that enable him to see the centrality and relevance of this mode of thinking. This learning by doing is essential in many of the sciences. For instance, only after a good deal of observation do the sparks in the bubble chamber become recognizable as the specific movements of identifiable particles.
Subject	Importance of learning by doing.
AF	**Learn**
LDA	Relev, central, essenti, think, mode, instanc Disciplin, learn, mani, abstract, example, think
Hybrid	Relev, central, essenti, think, mode, instanc Disciplin, **learn**, mani, abstract, example, think

Table 3. Derived affective model

Factors	B	*p-value*
Pleasure	21.66	0.998
Arousal	−0.77	0.042*
Dominance	0.20	0.540
Affective relation	1.22	0.046*
Intercept	−24.75	0.997

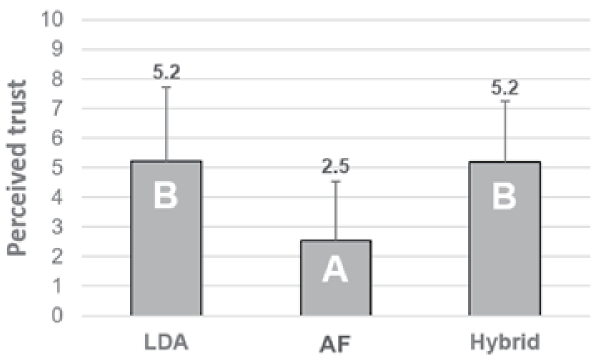

Figure 2. Perceived trust rate

5. **Discussion.** As a result of the evaluation of the reliability of the recommended words, there was no significant difference between LDA and hybrid method. When only the AF was applied, there was a difference between the other methods, which means that it was difficult to construct keywords in this case. In this study, we proposed a hybrid method using AF to improve the existing LDA method, but there was no statistical difference compared to the existing method. There are some reasons for this: first, the sample size of participants who were mobilized to develop the regression model was small and the three opinions may differ from those of the general public. Second, there is a possibility that the circumplex model may not be suitable for the keyword term recommendation because the pleasure, arousal and dominance values inherent in this model may not be appropriate since the emotional characteristics of each vocabulary are separated from the

subject composition. Finally, the paragraph itself to be analyzed may not be suitable for use with AFs.

Paragraphs are composed of various topics, which are mainly types of information transmission. If the emotion plays an important role, as in a literary work, the AF may work properly. Studies with similar premises have been conducted [14].

Nevertheless, the significance of this study can be seen from various perspectives. First, the proposed framework can be applied to other fields. It can be applied to important documents such as literary works, or in the analysis of the emotional characteristics and distribution of short articles in various social networking service. Second, it is meaningful to extend the application field of the circumplex model. In this study, the circumplex model was extended to the topic analysis. As the result of this study, if the affective filter is applied to the existing numerical vocabulary by the mathematical method, the vocabulary can be analyzed in more various ways.

6. Conclusion.

In this paper, we proposed a methodology to analyze the circumplex model based on the promising LDA algorithm to overcome the limitations of existing topic modeling techniques. For this purpose, we derived actual recommended words using paragraphs in a college entrance examination and evaluated the performance of recommendation qualitatively. The results showed that there was no statistically significant difference between LDA and AF with LDA.

As a result of this study, it is expected that the proposed AF-based topic modeling analysis method can be used as a basic technique to complement existing topic modeling techniques. Furthermore, it is expected to be used in analyzing and comparing the features of many documents that are circulated on the Internet and digitized.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2015R1C1A1A010541-48).

REFERENCES

[1] K. Schwab, *The Fourth Industrial Revolution*, Penguin UK, 2017.
[2] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.
[3] X. Cheng, X. Yan, Y. Lan and J. Guo, BTM: Topic modeling over short texts, *IEEE Trans. Knowledge and Data Engineering*, vol.26, no.12, pp.2928-2941, 2014.
[4] A. Bosch, A. Zisserman and X. Muñoz, Scene classification via pLSA, *Computer Vision – ECCV*, pp.517-530, 2006.
[5] Y. Guo, S. J. Barnes and Q. Jia, Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation, *Tourism Management*, vol.59, pp.467-483, 2017.
[6] Y. Wang, J. S. Lee and I. C. Choi, Indexing by latent Dirichlet allocation and an ensemble model, *Journal of the Association for Information Science and Technology*, vol.67, no.7, pp.1736-1750, 2016.
[7] H. K. Kim, S. H. Han, J. Park and J. Park, Identifying affect elements based on a conceptual model of affect: A case study on a smartphone, *International Journal of Industrial Ergonomics*, vol.53, pp.193-204, 2016.
[8] J. Park, S. H. Han, H. K. Kim, Y. Cho and W. Park, Developing elements of user experience for mobile phones and services: Survey, interview, and observation approaches, *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol.23, no.4, pp.279-293, 2013.
[9] J. A. Russell, Core affect and the psychological construction of emotion, *Psychological Review*, vol.110, no.1, pp.145-172, 2003.
[10] E. M. Daly, W. J. Lancee and J. Polivy, A conical model for the taxonomy of emotional experience, *Journal of Personality and Social Psychology*, vol.45, no.2, pp.443-457, 1983.
[11] J. A. Russell, A circumplex model of affect, *Journal of Personality and Social Psychology*, vol.39, no.6, pp.1161-1178, 1980.
[12] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament, *Current Psychology*, vol.14, no.4, pp.261-292, 1996.

[13] M. M. Bradley and P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, *The Center for Research in Psychophysiology*, University of Florida, Technical Report C-1, pp.1-45, 1999.
[14] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang and D. M. Blei, Reading tea leaves: How humans interpret topic models, *Proc. of the 22nd International Conference on Neural Information Processing Systems*, pp.288-296, 2009.