Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 11, No. 2, Desember 2021

ISSN 0216 – 0544
e-ISSN 2301– 6914

ASPECT EXTRACTION IN E-COMMERCE USING LATENT DIRICHLET ALLOCATION (LDA) WITH TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

[a] Satyawan Agung Nugroho, [b] Fitra A. Bachtiar, [c] Randy Cahya Wihandika

[a, b, c] Department of Informatics Engineering Universitas Brawijaya, Malang, 65145. Indonesia
E-mail: satyawanan@student.ub.ac.id, fitra.bachtiar@ub.ac.id, rendicahya@ub.ac.id

Abstract

Social media is a common thing that people use. Posts or comments found on social media describe someone's feelings and opinions so there have to be important topics that can be extracted from social media. In the e-commerce field, topic is an interesting thing to know because it can describes people's opinion towards a product. However, the large number of social media users is currently making the process of finding topics from social media difficult, so computer assistance is needed. One method that can be used is Latent Dirichlet Allocation (LDA). LDA is a good method for extracting topics, but the drawback is that sometimes the topics are incomprehensible. To cover up the drawback, TF-IDF feature selection method is used so that less important words can be skipped so LDA can generate a better topic. The best hyperparameter values obtained were 10 iterations, 10 topics, α and β values consecutively 0,1 and 0,01. The best feature selection percentile value is 90. This value is used to find the threshold that can be used as the lower limit of the TF-IDF value of each word so that the word with greater TF-IDF value can be used as feature.

Key words: Aspect Extraction, Latent Dirichlet Allocation, Perplexity, Term Frequency – Inverse Document Frequency, Topic Modelling.

INTRODUCTION

These days, e-commerce is commonly used by people around the world to buy things they want with ease. In 2019, 90% of Indonesian citizens who are 19 to 64 years old made a transaction on an e-commerce website [1]. On most e-commerce websites, users can leave comments on the section provided. These comments usually contain text, video, or image data.

Each comment certainly has a topic that is the core of information that represents user's thoughts on the product that they comment on. These comments are proven to be able to describe the user's emotion or opinion accurately [2]. However, as the number of e-commerce users increases, it becomes difficult to get a topic or generalization from a collection of user expressions. Also, most social media do not show the summary of the information. Therefore, there is a need to create a summarization on a given topic. The topic extraction may take a long time and the resulting topic can be subjective [3]. Given these limitations, computer assistance is needed to facilitate the process of extracting topics from these many sources.

One of methods could be used in extracting aspects of textual information is known as topic modeling. Topic modeling is an unsupervised approach that assumes each document contains several topics and these topics are a probability distribution of a collection of words. The output of this approach is several groups of words that are considered to represent a topic in a document [4]. There are two methods commonly used in the topic modeling approach, namely Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) [4]. PLSA is a method that determines the words forming the topic by looking for the probability of a topic in the document then looking for the probability of a word in the topic [5]. The two methods, PLSA and LDA, can be used to model the topics contained in the document. However, in the LDA method was proven to solve topic modeling problems by obtaining a lower perplexity value compared to the PLSA method.

In several studies, LDA has also proved to be effective in extracting aspects of information. In [2], LDA was used to find review topics on TripAdvisor, especially in the city of Phuket. In this study, LDA is said to be able to find topics with little bias. In [6], LDA is also used to find review topics on TripAdvisor but focuses on hotel reviews. The result is that LDA can identify 30 topics and the topics were named by two researchers. The topics are named based on the logical relationship formed between words that often appear on a topic. Whereas in the [3], LDA was used to find the topic of complaints given to the Customer Financial Protection Bureau. The results obtained are that LDA can find 40 meaningful topics. Although LDA is a method that can work effectively to extract topics, the resulting topics are sometimes not clearly understood by humans. This is because the words contained in a topic are sometimes less related to one another so that the relationships between words are less understandable and make it difficult to know the topics that are formed in the word collection [7].

In this study, LDA and Term Frequency-Inverse Document Frequency (TF-IDF) is used to detect topics in e-Commerce documents. Feature reduction is used to reduce the number of features in the document. To reduce the appearance of unclear words, the feature selection method using TF-IDF is used to remove words that are deemed less meaningful. Compared to N-Gram, the results of feature selection carried out with TF-IDF produce better features seen from the results of its application to machine learning algorithms [8]. The topic distribution (α), distribution of a word in a topic (β), and number of topics will determine the topic extraction results [9]. Thus, LDA hyperparameters will be determined by experiment to maximize the results. Thus, this research expects to help determine the best hyperparameters for LDA as well as many words that need to be removed to improve the quality of the resulting topic.

METHODOLOGY

Topic Extraction Overview

The overview of the Topic Extraction Model is shown in Figure 1. The first step of the topic extraction is data pre-processing. Basic text pre-processing is performed such as tokenizing etc. The second step is TF-IDF feature selection to reduce the number of

features. The third step is extracting topic using LDA.

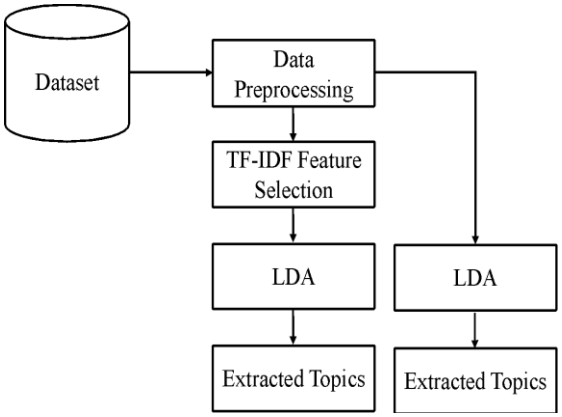

Fig 1. Topic extraction model

There are two topic extraction models which use feature selection and not using feature selection. The details of each steps is described in the following:

Data Collection

The data used in the study were collected from database repository. The data used are obtained from the Kaggle website "Grammar and Online Product Review" which can be accessed at https://www.kaggle.com/datafiniti/grammar-and-online-product-reviews. The dataset used is a collection of product reviews taken from several online shopping sites. The dataset contains 71046 data with various features such as review sentences, review titles, categories, review sources, and many others. For aspect extraction, the only required feature is the review text so that other features can be ignored. To keep the data variated, the order of the data is randomized to reduce the likelihood of data from the same category appearing close together. The total data used are 1000 data by taking the top 500 data as training data while the bottom 500 data as test data.

Data Pre-Processing

At this stage, pre-processing is performed to prepare the data so that feature selection and topic extraction can be done. First, tokenizing is done to separate each word in the documents, remove punctuation marks, and also change all letters to lowercase. Then, filtering is done to eliminate less important words. After that, lemmatization is performed to change the affixed words into its basic form. In this research, stemming was not used because the data used was in English and if stemming was used, many words in English would be cut off.

TF-IDF Feature Selection

TF-IDF is a weighting method commonly used in text mining. Weighting is done to convert text data into the numeric form so that it can be processed by the computer. TF is the value of the occurrence of a word in a document. The more often a word appears in a document, the greater the TF value. IDF is the value of the occurrence of a word in all existing documents. The more often a word appears in several documents, the more common the word is so that it should not be used [10].

The steps needed to calculate TF-IDF value according to [11] are:

1. Calculate the TF value of each word in each document using Equation (1).

$$wTF(t) = \{0, TF = 0 \;\; 1 + log\, log\, (TF(t)),\; TF > 0 \quad (1)$$

2. Calculate the IDF value of each word using Equation (2).

$$IDF(t) = log(\frac{|D|}{DF(t)}) \quad (2)$$

3. Calculate the TF-IDF value using Equation (3).

$$TF\, IDF(t) = wTF_{t,d} * IDF \quad (3)$$

Where *wTF (t)* is the weight of the *TF* of a word, *TF* is the occurrence of a word in a document, *IDF (t)* is the IDF weight of a word in a document, $|D|$ is the number of existing documents, *DF (t)* is the number of documents that contain the word *t* in it, *TF-IDF (t)* is the TF-IDF value of a word in a document.

Once every word in the document has its TF-IDF value, the next step is to omit words that have low TF-IDF values. The process is as follows:

1. Sort the words based on the TF-IDF value in descending order.
2. Choose a percentile value and calculate it. The calculated value is then selected to be the threshold.
3. Words that have a TF-IDF value greater than the threshold are the words that are going to be used, whereas the words that have a TF-IDF value less than the threshold are not going to be used.

The percentile value is chosen through the experiment. The values that are going to be

experimented on are 10, 20, 30, 40, 50, 60, 70, 80, and 90.

LDA Topic Modelling

There are several steps in the LDA Topic Modelling. LDA considers existing documents to be a collection of several hidden topics and each topic is a collection of several words [6]. The steps needed to extract aspects with LDA are:

First, initialize the α, β, K, i value. α is a parameter that determines the topic mixture in a document. The higher the α value, the more topics will be contained in a document. β is a parameter that determines the word mixture in each topic. The higher the β value, then each topic will have more words [12]. K is the number of topics, and i is the number of iterations.

In this study, the α value will be searched between 0.1 to 1, the β value will be searched between 0.01 to 1, the K value will be searched between 2 to 40, and the i value will be searched between 10 to 100.

The α and β values are both extremely important to the topic that will be generated by LDA, therefore to make sure these 2 values are assigned with the best possible value, there will be an experiment conducted on these 2 values. The experiment is conducted by running LDA with the K and *i value* set to the best value, and the α and β will be changed through each runs. There will be 2 scenarios regarding this experiment. The first one is where both the α and β values are increased after each runs and the second one is where the α value is increased and β value is decreased. The purpose of these 2 scenarios is to find out the best combination of α and β values.

Second, assign a random topic for each word in each document. To simplify the calculation process, the arrangement as in Table 1 and Table 2 can be used.

Table 1. Words Topics Table Example

	1st Topic to kth Topic
1st Word to nth Word	Number of nth word in kth topic

Table 2. Topics Documents Table Example

	1st Topic to kth Topic
1st Document to mth Document	Number of kth topic in mth document

Third, for each word (n = 1 to N), subtract 1 from the n^{th} word in Table 1 and the m^{th} document in Table 2. The word is considered missing because the topic probability calculation will be done for that word.

Fourth, Calculate the probability of the n^{th} word for each topic (k = 1 to K) using Equation (4).

$$w_n t_k = \frac{count(w_n|t_k)+\beta}{count(t_k)+W*\beta} * \frac{count(w\ in\ D\ |t_k)+\alpha}{count(w\ in\ D)+K*\alpha} \quad (4)$$

Where W is the number of unique words, K is the number of topics, D is the document where w_n is located.

Fifth, calculate the highest topic probability of a word using Equation (5). The topic with the highest probability will become the new topic of w_n. Then the values in Table 1 and Table 2 are updated after assigning new the topic to w_n.

$$p(t_k) = \frac{t_k}{\sum_{k=1}^{K} t_k} \quad (5)$$

Sixth, repeat the 3rd to 6th steps according to the specified number of iterations.

Seventh, calculate the ϕ value for each word on each topic using Equation (6).

$$\Phi(w_n t_k) = \frac{count(t_k)+\beta}{(\sum_{m=1}^{N} count(t_k))+N*\beta} \quad (6)$$

Where ϕ is the word distribution of the k^{th} topic, w is word, and t is topic.

Eight, calculate the Θ value for each topic in the document using Equation (7).

$$\Theta(d_n t_k) = \frac{count(t_k)+\alpha}{count(w\ in\ d_n)+K*\alpha} \quad (7)$$

Where Θ is the distribution of topics in the nth document, w is word, d is the document, and t is topic.

The flow chart for the overall model, pre-processing, the TF-IDF process, the LDA process, and the perplexity value calculation was made. Figure 2 provides an overview of the model created. The dataset is divided into two parts, training data and testing data. Each data is then pre-processed with tokenization, case folding, filtering, and stemming. Then, if feature selection is done, the model will select the feature to keep using the TF-IDF method. The next step is to extract the topic from the dataset using LDA and evaluate it using the perplexity score.

Evaluation

The results of the method implementation are evaluated. To measure the performance of LDA, the perplexity value is used. Perplexity is a value that describes how uncertain the model is, so that the lower the perplexity value, the better the resulting model will be [9]. To get the perplexity value from the model created, a calculation can be done using Equation (8).

$$Perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} loglog\,(p(w_d))}{\sum_{d=1}^{M} N_d}\right\} \quad (8)$$

The LDA hyperparameter testing conducted includes the best number of iterations testing, the best number of topics testing, and the best α and β values testing. And also to test the feature selection method, the best percentile value as the lower limit in feature selection is tested.

RESULT AND DISCUSSION

In this research, four experiments in finding the optimum parameter are done. Each testing will try to find the best hyperparameter value for LDA and the TF-IDF feature selection as well. Initially, the number of topics used was 7, the α value was 0.1, the β value was 0.01 [13].

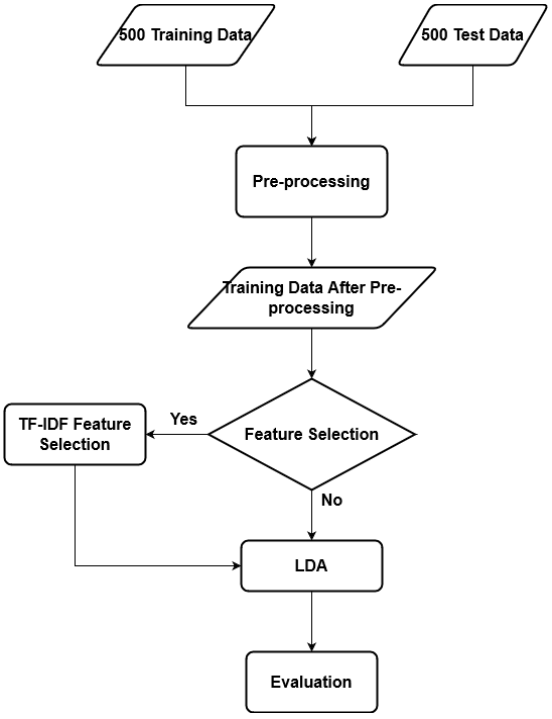

Fig 2. Model framework

Number of Iterations

The number of iterations testing is done to determine the number of iterations of the LDA process required to obtain the best quality of the topic. In this testing, the number of topics used was 7, the α value was 0.1, the β value was 0.01, and feature selection is not done. There are 10 number of iterations tested, namely 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The results of the

number of iterations test can be seen in Table 3 and Figure 3.

As can be seen in Table 3, the resulting average perplexity value is not much different from the smallest perplexity value of 60.015 and the largest value of 62.147. At Figure 2, the perplexity value for each number of iterations fluctuates and also uncertain, so it can be said that a large number of iterations does not affect the quality of the resulting topic. Thus, in this experiment, the number of iteration used is 10 which yield the lowest average perplexity value.

Table 3. Number of Iterations Testing

Number of Iterations	Average of Perplexity
10	60,01569026
20	60,74777191
30	60,14812664
40	62,14717162
50	60,72480628
60	60,4960433
70	60,20455816
80	61,01785
90	60,35923249
100	60,47171711

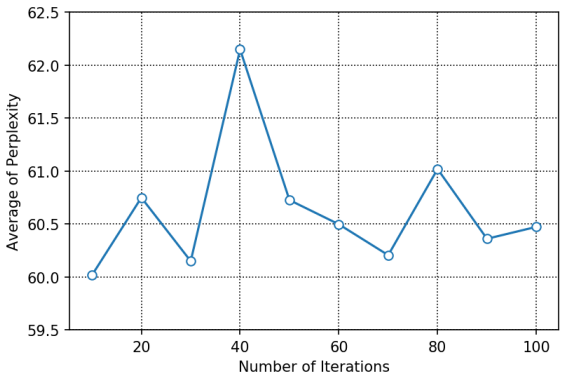

Fig 3. Number of iterations testing graph

Number of Topics

The number of topics testing is done to find out the best number of topics for aspect extraction using LDA. The parameter values

used remain the same in the previous testing, namely the α value of 0.1 and the β value of 0.01. However, because the number of iterations does not affect the quality of the topic, the number of iterations used is 10 so that the computation time can be reduced. The number of topics tested are 2, 5, 10, 15, 20, 25, 30, 35, and 40. The test results for the number of topics can be seen in Table 4.

As can be seen in Table 4, the perplexity value decreases as the number of topics increases. The best number of topics determined using the elbow method [2] as can be seen in Figure 3. As can be seen in Figure 3, the elbows of the graph are on the number of topics 5, 10, and 15. To get the best number of topics, LDA was executed three times with the number of topics of 5, 10, and 15 respectively. The result is that 10 topics gives the smallest percentage of ambiguous topics compared to 5 and 15.

Table 4. Number of Topics Testing

Number of Topics	Perplexity
2	157,879
5	77,88667
10	46,70364
15	34,81418
20	28,8117
25	23,18521
30	20,25588
35	18,13561
40	16,24489

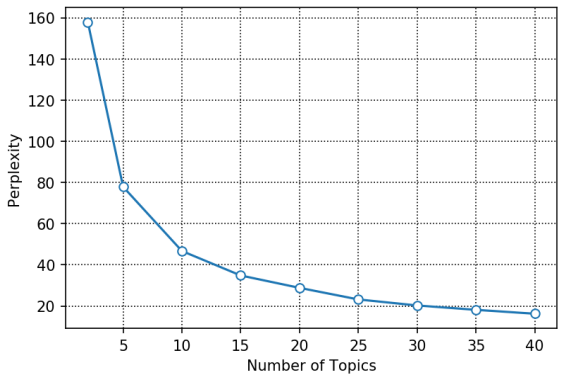

Fig 4. Number of topics testing graph

α and β Values Testing

α and β values testing is done to determine the best α and β values for aspect extraction using LDA. Two test scenarios are done. The first one where the α and β values both increasing. The second one where the α value increasing and the β value decreasing. The α values used are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. While the β values used are 0.01, 0.12, 0.23, 0.34, 0.45, 0.56, 0.67, 0.78, 0.89, 1. The number of iterations used is 10 and the number of topics used is 10. The test results of α and β values can be seen in Table 5 and Table 6.

In each test scenario, the best combination of α and β value is 0.1 and 0.01 (scenario 1) with the perplexity value of 46.98 and 1 and 0.01 (scenario 2) with the perplexity value of 35.01. When compared, the better combination of α and β value is 0.1 and 0.01 compared to 1 and 0.01 because less clear topics are found. The graph of the each of the scenario testing can be seen in Figure 4.

Table 5. 1st Scenario of α and β Value Testing

α	β	Perplexity
0,1	0,01	46,98358914
0,2	0,12	65,16709189
0,3	0,23	80,78008111
0,4	0,34	93,27983594
0,5	0,45	100,9357222
0,6	0,56	119,757602
0,7	0,67	131,5270641
0,8	0,78	134,0401075
0,9	0,89	141,6319856
1	1	143,5700096

Table 6. 2nd Scenario of α and β Value Testing

α	β	Perplexity
0,1	1	190,5237431
0,2	0,89	174,6433409
0,3	0,78	162,0143474
0,4	0,67	148,0302188
0,5	0,56	128,2162396
0,6	0,45	88,00661544
0,7	0,34	70,43666933
0,8	0,23	52,63991443
0,9	0,12	43,95964157
1	0,01	35,01748026

Percentile Value Testing

Percentile value testing is done to determine the percentile value that will be used in feature selection to increase LDA's performance in aspect extraction. The percentile values used are 10, 20, 30, 40, 50, 60, 70, 80, 90 with α 0.1, β 0.01, the number of iterations 10, and the number of topics 10. The results of percentile value testing can be seen in Table 7.

As can be seen in Table 7 that the lowest perplexity value is obtained with the percentile value of 90, that is 5.9, so it can be said that for feature selection, the best value that can be used is 90. However, the percentile value from 10 to 50 does not change much. This is caused by the numerous 0.005 values on the TF-IDF average, causing the threshold value used in the percentile value of 10 to 50 to be the same.

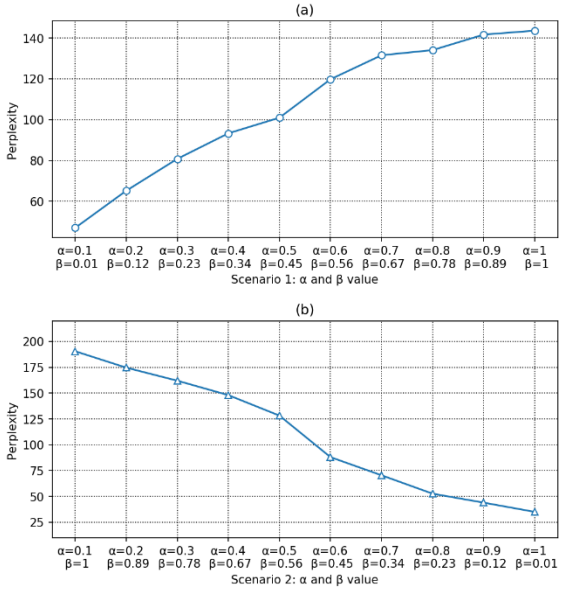

Fig 5. (a) Perplexity of scenario 1, (b) Perplexity of scenario 2

Global Testing

After getting the best hyperparameter value, global testing is conducted to show the results of the tests that have been done. In this global testing, 10 topics are used, the number of iterations is 10, the α and β values are 0.1 and 0.01. Feature selection is also done using 90 as the percentile value. The perplexity value obtained is 6.12 and the resulting topics can be seen in Table 8. The results obtained can be said to be quite good with less clear topics, only 3 out of 10 topics. This confirms that the hyperparameter value chosen is correct and the TF-IDF feature selection method is sufficient to help LDA to produce good topics. The topics spread and words in each topic can be seen in Figure 6, Figure 7, and Figure 8.

For example, as can be seen in Table 8, the first topic consists of like, would, one, first, good, little, cream, time, think, need words. These words are too different from each other and don't give any clear meaning. As for the Skin Treatment topic, the words are skin, use, moisturizer, feel, age, try, olay, even, great, anti. These words are clearly creating the meaning of beauty product or skin treatment.

Table 7. Percentile Value Testing

Percentile Value	Perplexity
10	25,03702307
20	24,61226083
30	24,38624955
40	24,80392676
50	24,04142863
60	16,39125937
70	16,39565084
80	10,5326157
90	5,900677961

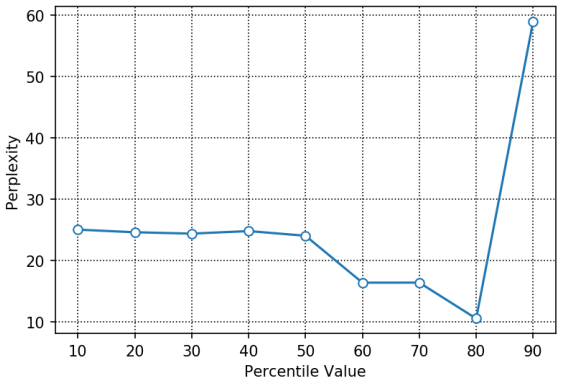

Fig 6. Percentile value testing graph

CONCLUSION

Based on the research that has been done about aspect extraction using LDA, it can be concluded that the appropriate hyperparameter values for the dataset used are the number of iterations 10, the number of topics 10, the α and β values, namely 0.1 and 0.01. Also, the best percentile value for feature selection used is 90 for the dataset used, or in other words, only 10% of words in the dataset are used. By using this parameters the words that surrounds the topics shows the relevances.

This study is still in initial stage. Further investigation as well as to improve the research that has been done regarding aspect extraction with LDA are as follows: First, the resulting topics sometimes display words that are less important so that additional pre-processing is needed to eliminate less important words. Also, the feature selection method used is new so that for further research, other feature selection methods can be used or modification of the methods used in this study. Second, the amount of data in this study is limited to 500 due to constrained computing device. Third, in terms of α dan β value there is a

need to find those value more objective. For example using optimization method. In this study both values are evaluated in turns.

Table 8. Global Testing Topics Result

Topics	Words
Unclear	like would one first good little cream time think need
Skin Treatment	skin use moisturizer feel age try olay even great anti
Movie	movie great good love kid watch enjoy family funny story
Housing	use clean wipe clorox love easy bathroom house everything make
Hair Treatment	hair conditioner use shampoo smell feel get really soft like
Unclear	look really color soft love find go lip buy make
Unclear	use mop work change get bottle spray stop buy start
Movie	movie buy great original godzilla like one well good love
Laundry	tide pod many laundry use detergent back know always bring
Fragrance	love scent well work recommend great fresh smell try best

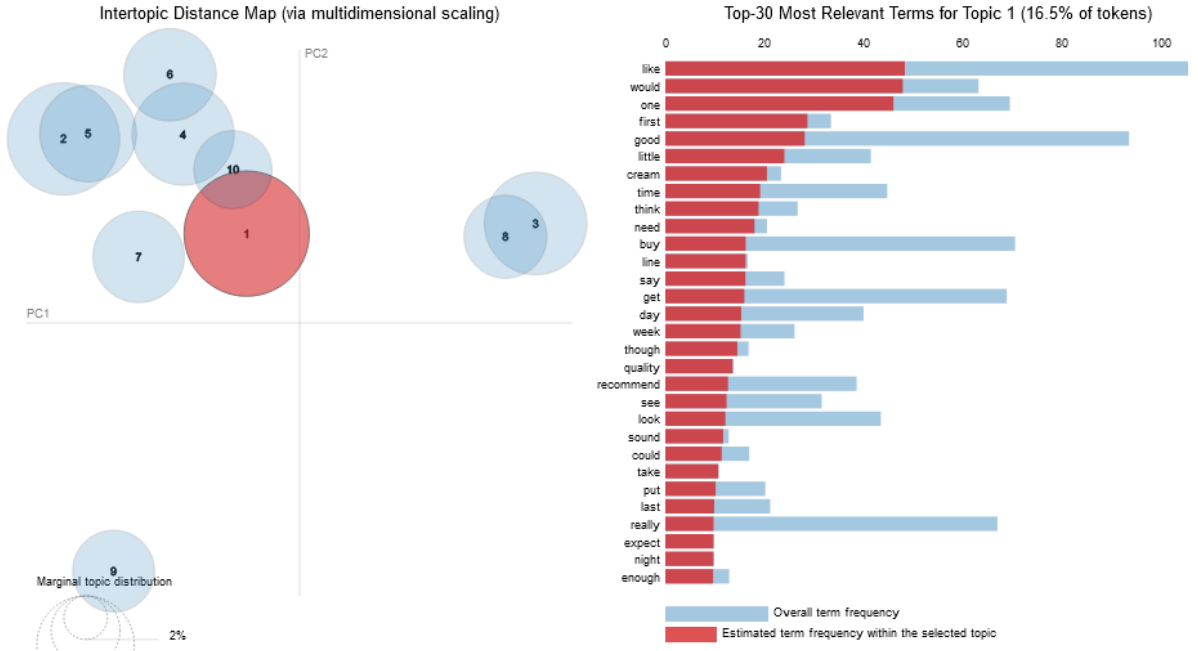

Fig 7. First topic(unclear) spreads and word relevance

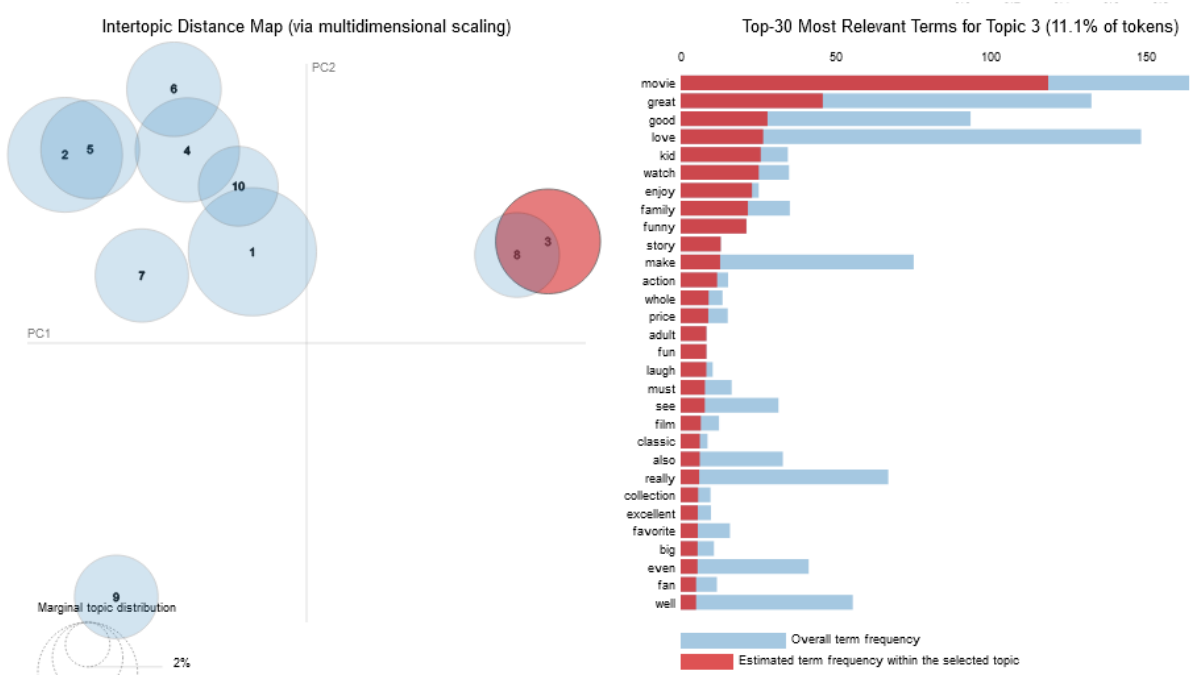

Fig 8. Third topic(movie) spreads and word relevance

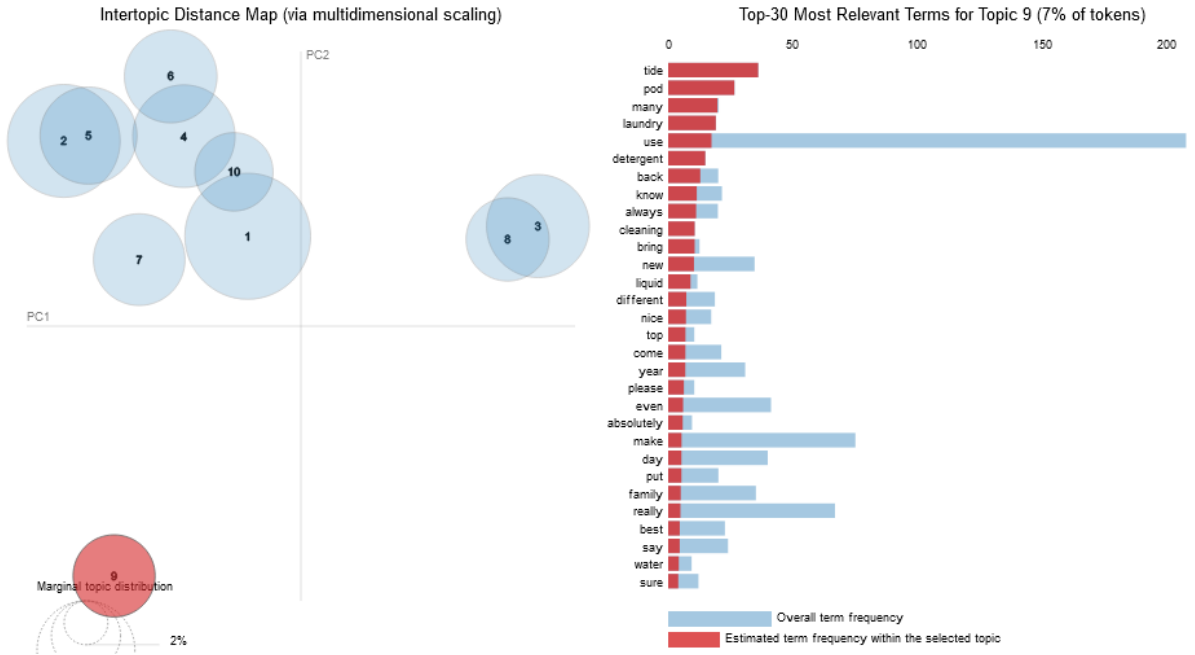

Fig 9. Ninth topic (laundry) spreads and word relevance.

REFERENCES

[1] S. Kemp and S. Moey, "Digital 2019 Spotlight: Ecommerce in Indonesia," 2019. [Online]. Available: https://datareportal.com/reports/digital-2019-ecommerce-in-indonesia.

[2] V. Taecharungroj and B. Mathayomchan, "Analysing TripAdvisor reviews of tourist attractions in Phuket , Thailand," *Tour. Manag.*, vol. 75, pp. 550–568, 2019.

[3] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer

complaints," *Expert Syst. Appl.*, vol. 127, pp. 256–271, 2019.

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

[5] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Mach. Learn.*, vol. 42, pp. 177–196, 2001.

[6] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews : Tourist satisfaction analysis using latent dirichlet allocation," *Tour. Manag.*, vol. 59, pp. 467–483, 2017.

[7] D. Mimno, H. M. Wallach, E. Talley, and M. Leenders, "Optimizing Semantic Coherence in Topic Models," *Proc. 2011 Conf. Empir. Methods Nat. Lang. Process.*, no. 2, pp. 262–272, 2011.

[8] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Comput. Sci.*, vol. 152, pp. 341–348, 2019.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[10] N. C. Wirawan, Indriati, and P. P. Adikara, "Analisis Sentimen Dengan Query Expansion Pada Review Aplikasi M- Banking Menggunakan Metode Fuzzy K-Nearest Neighbor (Fuzzy k-NN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 362–368, 2018.

[11] W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1750–1757, 2017.

[12] A. Agustina, "Analisis Dan Visualisasi Suara Pelanggan Pada Pusat Layanan Pelanggan Dengan Pemodelan Topik Menggunakan Latent Dirichlet Allocation (LDA) Studi Kasus: PT. Petrokimia Gresik," Institut Teknologi Sepuluh November, 2017.

[13] H. Hao, K. Zhang, W. Wang, and G. Gao, "A Tale of Two Countries : International Comparison of Online Doctor Reviews Between China and the United States," *Int. J. Med. Inform.*, vol. 99, pp. 37–44, 2017.

.