Mining Customer Opinion for Topic Modeling Purpose: Case Study of Ride-Hailing Service Provider

Reggia Aldiana Wayasti, Isti Surjandari, Zulkarnain
Industrial Engineering Department,
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
regialdiana@gmail.com, isti@ie.ui.ac.id, zulkarnain@ie.ui.ac.id

***Abstract*—The popularity of ride-hailing services in the form of smartphone application as a transportation solution has become center of attention. The convenience offered has made many people use it in daily life and discuss it on social media. As a result, ride-hailing service providers utilize social media for capturing customers' opinions and marketing their services. If customers' statements about ride-hailing services are analyzed further, service providers can get insight for evaluating their services to meet customers' satisfaction. Text mining approach can be useful to analyze large number of posts and various writing styles to extract hidden information. Furthermore, by applying topic modeling, service providers can identify the important points that were spoken by customers without previously giving label or category to the text. Latent Dirichlet Allocation was used in this study to extract topics based on the posts from ride-hailing customers published on Twitter. This study used 40 parameter combinations for LDA to get the best one to obtain the topics. Based on the perplexity value, there were 9 topics discussed by customers in their posts including the top words in each topic. The output of this study can be used for the service providers to evaluate and improve the services.**

Keywords—text mining; topic modeling; latent Dirichlet allocation; social media analytics; ride-hailing service

I. INTRODUCTION

The high mobility of urban population causes the demand for transportation continues to increase, while the availability of public transportation is still inadequate in terms of quantity and quality. As a result, people tend to use private vehicle to accommodate their needs to travel. The ever-increasing number of private vehicles creates impacts such as congestion, pollution, reduced public space, and declining quality of life [1].

Various solutions have been developed to decrease the usage of private vehicles yet still fulfill people's needs for transportation at the same time. One of the innovative solutions that has become popular in the society is ride-hailing services in the form of smartphone applications. The advantages offered by those ride-hailing applications are the ease of ordering and getting the vehicle due to better availability, comfort, assured security, and more affordable price compared to conventional transportation [2].

Currently, there are many application-based ride-hailing service providers. They are able to expand their business to various countries and cities, and continue to innovate to bring other services that facilitate customers. The ongoing innovation and expansion has made the ride-hailing business more developed and become topic of conversation, including in social media. This social media phenomenon drives service providers to utilize it for capturing customers' opinions and complaints, as well as providing up-to-date information about their services and promotions, aiming to reach higher revenue and customer loyalty [3].

If those statements in social media are analyzed more thoroughly, ride-hailing service providers can define and evaluate the quality of services that has been provided based on the customer's point of view. However, the large number of posts and various writing styles make the analysis process more complicated and time consuming if done manually [4]. Therefore, text mining approach can be useful in extracting hidden information within the large number of social media posts which is in form of textual data [5]. One of the applications of text mining is topic modeling to identify the important points that were spoken by customers without previously giving label or category to the text [6].

Since both ride-hailing and social media are also trending in Indonesia, this study aimed to analyze customers' opinions and complaints to one of the service providers. The opinions and complaints were obtained from Twitter, which is one of the most active social media in Indonesia [7]. The approach used for the topic modeling was Latent Dirichlet Allocation (LDA) to identify aspect from the occurrence of the word or phrase [8]. The outcome of this study can help the ride-hailing service provider to provide and improve the services to meet customers' needs.

II. RELATED WORK

There have been several researches that use LDA for topic modeling and sentiment analysis purposes. Research by Duan,

ISBN: 978-1-5386-4571-0 (c) 2018 IEEE

Ai, and Li [9] uses LDA with user interest combination to make an effective recommendation for microblog users. The information in microblog is divided into topics, then the time interval is set and the weights of user interest are accumulated. Another research that uses LDA for topic modeling is done by Allahyari and Kochut [10], which combines topics and concepts for document tagging with categories in Wikipedia. However, in the topic model, the hierarchical structure of the categories is not taken into account directly. Tong and Zhang [11] use LDA for topic modeling in Wikipedia and Twitter. The aim is to search, explore, and recommend articles for the readers in Wikipedia, and analyze users' interest in Twitter.

Moghaddam and Ester [12] grab text data from review website for aspect-based opinion mining. They use and evaluate six LDA-based models (LDA, S-LDA, D-LDA, PLDA, S-PLDA, and D-PLDA) for topic modeling to identify the impact of each model to the result. Afterwards, the opinion phrases are extracted based on grammatical relations. On the other hand, Putri and Kusumaningrum [13] use LDA for identifying the tendency of the reviews and classifying visitors' sentiment in TripAdvisor, in which different set of parameters are used. Lim and Buntine [14] apply LDA-based opinion model method that utilize hashtags, mentions, emoticons, and sentiment words. The interaction between target and opinion can be modeled directly so that the opinion prediction can be improved.

III. Research Methodology

Text mining is a part of data mining which is a process of analyzing unstructured text data to extract hidden information [5]. In recent years, the research related to text mining has received much attention and is actively conducted along with the increasing number of text data obtained from various sources, including social media. The use of text mining is also popular for business needs because it produces deep and innovative analysis [15].

Generally, there are three steps in text mining to finally get informations from the data [16]. The text data is first collected from certain source, then pre-processing step is done to transform text data to be more structured so that it will be easier to discover information in the text data.

Text mining can be done with both supervised and unsupervised methods. Compared to unsupervised ones, the result produced by supervised method tend not to be flexible because it is limited to the feature and parameter established previously [17]. Therefore, this study used topic modeling approach to extract the important points based on customers' opinion.

This study used Latent Dirichlet Allocation (LDA), which is a method for probabilistic topic modeling. Probabilistic topic modeling itself is a method to easily organize and discover information and topic within the text data. The main task of LDA is finding the topic which is distributed in the corpus with the high probability. LDA uses bag of words assumption that not counts on the order of words [6].

In general, this study consists of three main steps namely data collection, text pre-processing, and topic modeling using LDA.

A. Data Collection

In this study, the text data used was obtained from the messages posted on Twitter that are directed to one of the ride-hailing service provider in Indonesia. Those messages, also called tweets, were downloaded using a program that can obtain tweets through Twitter API. The tweets used for this study were those that is posted during January 2018. The Twitter API can only grab tweets for a week long, so the process of downloading the tweets could only done once a week and four times in total.

Since the scope of the service provider is only in Indonesia, the tweets collected are in Indonesian. In addition, this study only focuses on tweets from customers about the services including their opinion, experience, compliments, and complaints perceived while using the services. Before moving to the next step, the tweets went through screening process to remove unnecessary and irrelevant ones. The number of tweets that were proceed to pre-processing step was 3,139 tweets.

B. Text Pre-processing

The tweets collected were still unstructured due to various writing styles such as the usage of slangs, abbreviations, and non-alphanumeric characters. Therefore, pre-processing steps were necessary to be done to get the data ready to be processed in the next step.

There were five steps of pre-processing done in this study. Tokenization was done to split up the sentences into terms. The next step was case folding that converts all the letters into lowercase. Next, the words were changed into standard language in Indonesian by spelling normalization. After that, filtering was done to remove mentions, retweets, hashtags, URLs, and non-alphanumeric characters. In addition, stop words and other words that have no particular and related meaning to the context were also removed. Stemming was also done to change each word to the root form.

After the steps were completely done, all of the tokens were converted to Bag of Words (BoW). BoW captures the appearance of words in each sentence or document with reference to the tokens [13]. The generated BoW was then proceed as the input for topic modeling step.

C. Topic Modeling

This study used Latent Dirichlet Allocation (LDA), which is a method for probabilistic topic modeling. Probabilistic topic modeling itself is a method to easily organize and discover information and topic within the text data. The main task of LDA is finding the topic which is distributed in the corpus with the high probability. LDA uses Bag of Words assumption that not counts on the order of words [6].

There are several parameters in LDA, just like in the model in Fig. 1 [6]. In the model, D is the number of

documents, while N is the number of words in the document. The parameters that have to be defined are number of topics (K), number of iterations, hyperparameter of topic distributions in document (α), and word distribution in topic (β). The topic assigned for the word in the document is defined as $z_{d,n}$, and $w_{d,n}$ the words observed in the document.

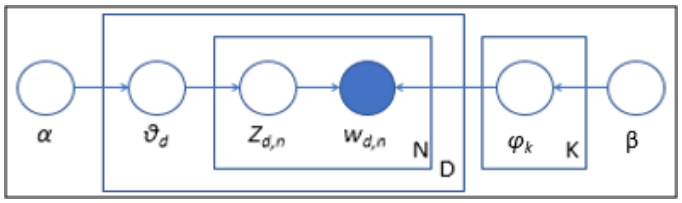

Fig. 1. Graphical model of LDA

LDA has two different processes namely generative and inference process. As generative process, LDA is applied to create document in which the parameters such as word distribution in the topic (φ_k) and the proportion of topic for each document ($\theta_{d,n}$) are known previously. On the other hand, the inference process is applied to identify those parameters and word distribution ($w_{d,n}$) among number of topics based on the available document [13]. This study used the inference process of LDA, with Gibbs sampling algorithm, in which the idea is to update each of the variable based on the conditional probability of other variables [18]. Therefore, the main task is to obtain the value of φ_k and $\theta_{d,n}$ with the calculation using following formulas,

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

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

where $n_{t,k}$ is number of words that are assigned in a topic, $n_{d,k}$ is the number of words in the document that are assigned in a topic, and V is the amount of different words in the document [18].

Due to the usage of inference process, the parameters had to be set first. The parameters were the number of topics, number of iterations, and the alpha (α) hyperparameter. The number of topics were varied from 5, 6, 7, 8, and 9. The number of iterations were 2000, 3000, 4000, and 5000. The alpha (α) hyperparameters were set to 0.1 and 0.01.

For the evaluation process, this study used perplexity value. Calculating perplexity value is useful to define the ability of probability model to predict a sample. The more the perplexity value is, the less the probability model can predict a sample The combination used was the one that has least perplexity value. Afterwards, the topics were extracted based on the combination. The equation for calculating perplexity is as follows [11]:

$$Perplexity(w) = exp\left\{\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\} \quad (3)$$

where $n^{(jd)}$ is the frequency of occurrence of j^{th} word in d^{th} document.

IV. Result and Analysis

There were 40 parameter combinations for data processing with LDA. Fig. 2 and Fig. 3 show the results of the perplexity of all parameter combinations. Based on both graph, the smallest perplexity value is 90.41 which comes from the combination of which the number of topic 9, number of iteration 4000, and the alpha hyperparameter 0.1. Therefore, the topics were extracted based on this result.

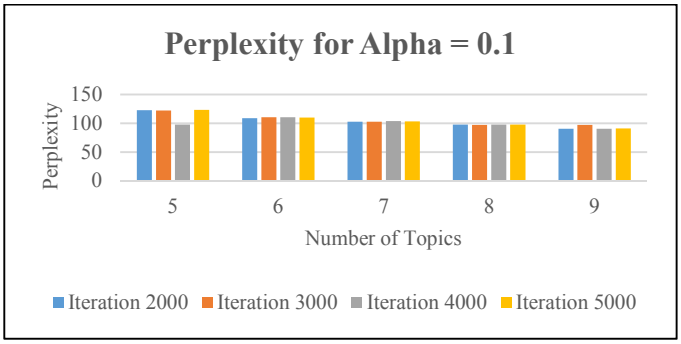

Fig. 2. Perplexity for the combinations with alpha = 0.1

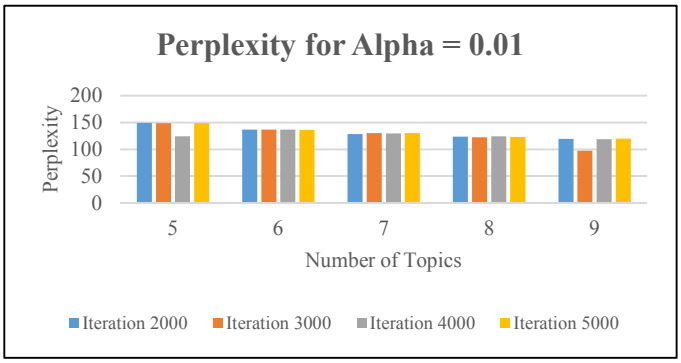

Fig. 3. Perplexity for the combinations with alpha = 0.01

Based on the result above, it can be concluded that number of topics can affect the perplexity. This is due to when the number of topics is small, there may be possibility of ambiguity in the topic. The number of words in a topic may be too big that irrelevant words can be appeared and makes the interpretation of the topic more difficult.

The alpha (α) hyperparameter can also affect the perplexity in the result. The value of alpha determines the distribution of topics per document. Smaller value of alpha resulting in larger perplexity, which means the topics distributed in the document are fewer. Meanwhile, if the alpha value is larger, the distribution tend to be more stable, resulting in better perplexity [13].

The best combination with smallest perplexity value is further analyzed to extract the words that appear in each topic. From those words, the main point discussed in each topic can be defined. Table I shows the top words that appear in the topics.

The words in Table I represent customers' experience of using the services from the ride-hailing service providers. In the application there are several services provided, and the most discussed other than ride-hailing is food delivery in

Topic 2, loyalty points in Topic 4, electronic money in Topic 7, and instant courier in Topic 8. In addition, customers also discuss about how the firm handle and solve problems and complaints in Topic 1, application performance in Topic 3, experience and

TABLE I. TOP WORDS APPEARED IN THE TOPICS

Topic	Words	Translation
1	"pesan", "bantu", "email", "keluh"	"order", "help", "email", "complaint"
2	"makan", "pesan", "aplikasi", "harga", "fitur", "chat", "beda"	"eat", "order", "application", "price", "feature", "chat", "different"
3	"pesan", "aplikasi", "akun", "pakai", "login", "layan", "langgan"	"order", "application", "account", "use", "login", "service", "customer"
4	"voucher", "pakai", "poin", "bayar", "ribu"	"voucher", "use", "point", "pay", "thousand"
5	"tipu", "telepon", "akun", "tolong", "orang", "kode"	"fraud", "call", "account", "help", "person", "code"
6	"kemudi", "pesan", "jalan", "batal", "tolong", "jemput"	"drive", "order", "way", "cancel", "help", "pick"
7	"isi", "saldo", "masuk", "tarik", "akun"	"top-up", "balance", "increase", "withdraw", "account"
8	"kemudi", "barang", "pesan", "ambil", "kirim", "jam"	"drive", "item", "order", "pick", "send", "hour"
9	"naik", "tarif", "kasih", "terima", "mahal", "kemudi", "ribu"	"increase", "tariff", "give", "accept", "expensive", "drive", "thousand"

risk of fraudulence in Topic 5, drivers' behavior and trip experience in Topic 6, and the fare for ride-hailing in Topic 9. The given category names for the topics are shown in Table II, and the explanations are as follows.

A. Responsiveness and the time taken to solve customers' complaints and issues

The tweets related to this topic are mostly from customers who already report their problem or bad experience to the customer service. The time taken for solving the issues can be varied. Some customers got their issues finished in just few minutes, yet some others are still waiting for the issue to be solved or finished. Those whose problems solved faster give compliments, while those who are waiting complains more about the slow response of the customer service.

B. Experience and issues in food delivery service

One of the popular service in the application is food delivery service because it helps customers who cannot go outside to get food when they are hungry. Therefore, customers often give compliment for this service and the convenience offered. Meanwhile, there are some of them who faced problems when ordering foods such as errors in the application, different cost details before and after ordering, and difficulty in contacting drivers because there is no chat feature. Besides, the restaurant database is not updated very often, and the price in the application is sometimes different than the real price in the merchant.

C. The stability of the application to be used daily

In this topic, many of the customers complains that their application cannot be used to make orders of some or all services. When it occurs, they have to contact the customer service to solve the issue, yet it happens frequently even for days. Another issue related to this topic is the possibility of being logged out automatically from the application and cannot login again to make orders. Those issues are dangerous because customers can easily switch to other ride-hailing application. However, in spite of the complaints about the stability, there are still some customers that give compliment about the better application interface in the updated version

D. Loyalty points and rewards

There is a loyalty point program in the application where customers can get token after using each of the service. This token can be swiped to finally get points, and the points can be collected to be redeemed with vouchers or merchandise. This program is beneficial for loyal customers, but the total points to be redeemed keep increasing and the value of the vouchers changes from time to time. Some customers feel objected to this issue.

E. Fraudulence occurrence and possibility

Another problem complained by customers is the security of the application. Customers often get phone call from unknown number that claim to be the customer service who give information about prize for them. However, they end up getting their account hacked or losing their balance in the electronic wallet. Although customers have been told to be careful and many of them have been aware of this, the possibility of fraudulence is still exist and there has to be an act of ending the occurrence.

F. Drivers' behavior and customers' trip experience

Customers often share their experience when using the service. It can be related to the driver from ride-hailing service. According to the top words, the most told story is about the motorcycle or car drivers who do not pick up the customers, or customers have to wait for a long time for the driver to arrive at pick up location. Meanwhile, there are still some others who got good experience on the trip such as getting friendly driver, and got their things returned after previously left in the car.

G. Electronic money system

This feature of the application enable customers to make payment for the service directly. The balance can be purchased by topping up via drivers or banking services. Besides, the balance can be transferred and withdrawn to bank account. The thing that customers discuss the most is problem when topping up the balance because sometimes it is not automatically increase after the process. The other issue is the trouble in withdrawal and account verification for the process.

H. Instant courier service

The instant courier service is preferred by customers because they can send goods to others and get it delivered on the same day. This service is also partnered with famous online marketplace to make the delivery process faster. However, the drivers usually take long time to pick up the goods, even to deliver it to customers. In some cases, the goods is not delivered on the same day, or the orders get canceled in the middle of delivery process.

I. The fare of the services

One of the reason of using ride-hailing application is its cheaper fare than conventional transportation. However, recently customers complain about the fare that become more expensive without prior notice. In addition, the fare difference when using electronic money and cash become slighter. This issue made customers confused and some of them choose not to use the service until the fare become normal again.

TABLE II. CATEGORY NAMES FOR EACH TOPIC

Topic	Category
1	Responsiveness and the time taken to solve customers' complaints and issues
2	Experience and issues in food delivery service
3	The stability of the application to be used daily
4	Loyalty points and rewards
5	Fraudulence occurrence and possibility
6	Drivers' behavior and customers' trip experience
7	Eectronic money system
8	Instant courier service
9	The fare of the services

It can be found that the top words in each topic mostly contain negative sentiments just like the explanation above. This can happen because customers get trouble more often while using the services, and directly complain to the customer service through Twitter. The result of this study can be a consideration for the service provider to improve their service and make their system more safe so that their loyal customers will stay to use the services, and their new customers will be more loyal after getting better experience in using the services.

V. CONCLUSION

The result of this study gave insight to one of ride-hailing service providers to define the points discussed by customers about their opinion, experience, and complaints in social media. After processing the data from Twitter using LDA with 40 parameter combinations, the smallest perplexity value which is 90.41 was generated by a combination of 9 topics, 4000 iterations, and alpha value of 0.1. From the words appeared in each topic, the most used services and the most discussed points were defined so that it would be easier to plan the improvement for their services to meet customers' needs, as well as solving the problems and complaints.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Universitas Indonesia for funding this study through Thesis Research Grants for Indexed International Publication, No 2456/UN2.R3.1/HKP.05.00/2018.

REFERENCES

[1] K. H. Leung, "Indonesia's Summary Transport Assessment," Asian Development Bank, Manila, 2016.

[2] International Transport Forum, "App-Based Ride and Taxi Services," OECD, Paris, 2016.

[3] W. He, S. Zha and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management,* vol. 33, pp. 464-472, 2013.

[4] W. Claster, P. Pardo, M. Cooper and K. Tajeddini, "Tourism, travel and tweets: algorithmic text analysis methodologies in tourism," *Middle East Journal of Management,* vol. 1, no. 1, pp. 81-99, 2013.

[5] I. Surjandari, M. S. Naffisah and M. I. Prawiradinata, "Text Mining of Twitter Data for Public Sentiment Analysis of Staple Foods Price Changes," *Journal of Industrial and Intelligent Information,* vol. 3, no. 3, pp. 253-257, 2015.

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

[7] S. Kemp, "Digital in 2017: Southeast Asia," We Are Social, New York, 2017.

[8] F. Colace, L. Casaburi, M. De Santo and L. Greco, "Sentiment detection in social networks and in collaborative learning environments," *Computers in Human Behavior,* vol. 51, p. 1061–1067, 2015.

[9] J. Duan, Y. Ai and X. Li, "LDA topic model for microblog recommendation," in *International Conference on Asian Language Processing*, Suzhou, 2015.

[10] M. Allahyari and K. Kochut, "Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network," in *IEEE Tenth International Conference on Semantic Computing*, Laguna Hills, 2016.

[11] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling," in *The Sixth International Conference on Computer Science, Engineering and Information Technology*, Vienna, 2016.

[12] S. Moghaddam and M. Ester, "On the design of LDA models for aspect-based opinion mining," in *The 21st ACM International Conference on Information and Knowledge Management*, Maui, 2012.

[13] I. R. Putri and R. Kusumaningrum, "Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia," *Journal of Physics: Conference Series,* vol. 801, no. 1, pp. 1-6, 2017.

[14] K. W. Lim and W. Buntine, "Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon," in *The 21st ACM International Conference on Information and Knowledge Management*, Maui, 2012.

[15] G. Chakraborty, M. Pagolu and S. Garla, Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, North Carolina: SAS Publishing, 2013.

[16] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and R. A. Nisbet, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Oxford: Elsevier, 2012.

[17] B. Gruen and K. Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software,* vol. 40, no. 13, pp. 1-30, 2011.

[18] 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, no. 1, 2014.