Topic Modeling Using Latent Dirichlet Allocation (LDA) and Sentiment Analysis for Marketing Planning Tiket.com

Berlin Helmi Puspita[1], Muhammad Muhajir[1,*], Hafizhan Aliady[2]

[1]*Departement of Statistics, Faculty of Mathematics and Science, Islamic University of Indonesia, Kaliurang Street 14,5 Kilometers*
[2]*PT. Global Tiket Network (Tiket.com)*
[*]*Corresponding author: mmuhajir@uii.ac.id*

ABSTRACT

Tiket.com is a company that provides online ticket booking services in Indonesia. Tiketcom wants to improve services by knowing content that is widely discussed by the public and positive and negative comments on Tiketcom. Therefore an analysis will be done using Twitter accounts with the Latent Dirichlet Allocation (LDA) method which aims to find patterns in a document that raises various topics from text data and sentiment analysis to find out positive and negative comments on Tiketcom. The data used is tweet and retweet the users Twitter to Tiketcom accounts starting from 17 November 2018 to 4 March 2019. Obtained as many as 20 topics in the text data and taken 5 topics with the highest coherence value to obtain a topic model. After analyzing the LDA it was found that 5 topics that were widely discussed were promo discount tickets provided by Tiketcom. In sentiment analysis 21.1% of negative tweets were obtained, mostly discussing disruption to ticket reservations and 15.4% positive tweets mostly discussing vouchers given by Tiketcom to their customers.

Keywords: *Tiketcom, Topic Modeling, Latent Dirichlet Allocation, Sentiment Analysis.*

1. INTRODUCTION

Tiketcom or also known as PT Global Tiket Network is an online ticket sales application that has been established since August 2011. In the application, Tiketcom provides a choice of services, namely hotel reservations, airplane tickets, train tickets, car rentals, and event ticket purchases. In addition to these services, there also available information about tours in Indonesia, tips and tricks on vacation packages in several places (Anonim, 2019).

Currently, Tiketcom is not the only online travel agency company in Indonesia, there are 4 other competitors. From the 5 online travel agent companies, two companies dominate the market today, namely Tiketcom and Traveloka, whose success is measured through the number of pageviews (Bhapkar, 2013). Based on the data obtained, there is a decrease in visitors to Traveloka, while the visitors to the Tiketcom website tend to be stable. However Tiketcom also must to maintain this stability and even have to increase. Thus, as an online-based company, Tiketcom must be able to maintain the quality of the website, therefore it is important to explore the factors a website must-have. One very important factor is understanding the type of content in the information provided to get a positive response and attract the attention of the users (Cao, Zhang, & Seydel, 2005).

Based on the above factors, Tiketcom wants to analyze to increase website visitors, sales, and marketing planning. The method used for this analysis is Topic Modeling using the Latent Dirichlet Allocation (LDA) algorithm and Sentiment Analysis. The selection of the Topic Modeling method with the LDA algorithm is because LDA uses a hierarchical model so that it is more stable and better at processing large amounts of data which is useful for determining topics in owned documents (Liu & dkk, 2013). Besides sentiment analysis was also carried out to obtain more specific results regarding the positive and negative things written about Tiketcom. The data used comes from tweets and retweets against the Tiketcom account. The use of Twitter as a medium to obtain text data based on information on Twitter is more widely spread, Twitter is a place to post all comments or things that are felt by everyone, and its easy use (Zarrella , 2010).

1.1. Literature Review

Research conducted by Hermanto (2016) shows that the Naive Bayes method is better than random forest because the naive bayes Classifier method only requires a small amount of training data to estimate the parameters required in classification. However, the results of the analysis using the Naive Bayes method are less specific than the results of the text classification using the Latent Dirichlet Allocation (LDA) method. The results obtained in Naive Bayes only show general topic classifications, while the LDA displays words related to the resulting topic.

There are several algorithms in document clustering such as K-means, Hierarchical Agglomerative Clustering, Singular Value Decomposition (Sahu & Srivastava, 2016). The three algorithms form several clusters of the document without human assistance. However, the problem faced by these three methods is that the main topic in the cluster cannot be generated, in contrast to the Latent Dirichlet Allocation (LDA) method which can bring up the main topic in the cluster (Sun, 2014) (Herwanto, 2018).

The Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) methods are text mining methods that ignore word order in the analysis process, this is a problem because certain terms will have very different meanings if they are looked at through the sequence, thus allowing grouping of terms. into a topic to be inappropriate. To overcome this problem, LDA is used as a method of grouping terms into certain topics by paying attention to the order of words in the formation process through the mixed model mechanism. (Zulhanif & et al, 2017) (Hoffman, 1999).

2. BASIC THEORY

2.1. Text Mining

Text Mining is a knowledge intensive process where users interact and work with a set of documents using several analytical tools (Feldman & Sanger, 2007). The purpose of text mining is to determine patterns and find new information so that it can be used to process, organize, and analyze large amounts of unstructured text (Rosadi, 2014). The text mining process requires several initial stages to prepare text data to be more orderly and structured as follows :

1.1.1. Text Preprocessing

This stage is a process of semantic analysis (the truth of meaning) and syntactic (correctness of structure) of text data with several processes, namely:

a. Case folding
b. Tokenizing
c. Stemming
d. Tagging
e. Feature Selection

1.1.2. Text Representation

This stage is useful for converting text data into representative data for easier processing.

1.1.3. Analyzing

This stage is the analysis stage which will then be carried out after the text mining process.

2.2. Term Frequency-Inverse Document Frequency (TF-IDF)

The Term Frequency-Inverse Document Frequency method is the weighting of relationships and a statistical measure used to evaluate how important a word is in a document (Grossman & Ophir, 1998). The following is the TF-IDF weighting formula.

$$W = TF \ x \ IDF \qquad (1)$$

with TF, namely the number of words that appear in each document and the IDF

$$IDF = log \ log \left(\frac{D}{DF}\right) \qquad (2)$$

where D is the number of all documents, DF is the number of documents that contain that word.

2.3. Topic Modeling

The concept of topic modeling consists of several entities, namely words, documents, and corporations. Words are considered as the basic unit of discrete data in a document which are defined as items of vocabulary indexed for each unique word in the document. Documents are structures of many words. Meanwhile, the corpus is a collection of documents and is the singular form of the corpora. While the topic is a distribution of several vocabulary words that are fixed. The purpose of topic modeling is to determine topics automatically from a set of documents (Blei, 2012). Here is one method of Topic Modeling that is currently popular.

2.3.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a general model of probability for a discrete set of data such as a corpus. LDA can be used to summarize, cluster, link very large data because LDA generates a weighted list of topics for each document (Hindle & Campbell, 2015).

According to Li and Jain (1998) there is a document formation process by assuming D as a corpus which is a collection of documents (d) as follows:

1. Choose a random topic that has the following distribution:

$$\varphi^{(k)} \sim Dirichlet(\beta), \qquad \text{for k=1,..,K}$$

2. Randomly select a distribution of topics for the document (d).

$$\theta_d \sim Dirichlet(\alpha), \qquad \text{d € D}$$

3. For every word on the document :

   i. Randomly select a topic on the topic distribution.

   $$z_i \sim Dirichlet(\theta_d)$$

   ii. Choose a word randomly from the corresponding topic of the vocabulary distribution.

   $$w_i \sim Dirichlet(\varphi^{(z_i)})$$

with :

K = numbers of latent topics in the document

$\varphi^{(k)}$ = discrete probability distribution on vocabulary representing the class-k topic distribution

θ_d = distribution of the d-document of the available topics

z_i = topic index on the class-i word

w_i = word class-i

α, β =parameters for the Dirichlet distribution

Blei presented the LDA method with a probabilistic method visually as shown in Figure 1.

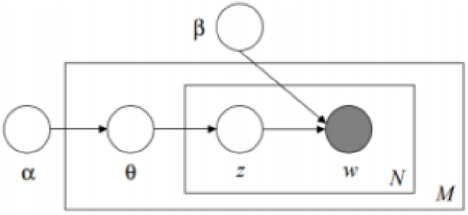


Figure 1. LDA visualization by Blei. Source: Blei (2012)

2.4. *Sentiment Analysis*

Sentiment analysis is the process of understanding, extracting, and processing textual data automatically to get the information contained in a sentence so that it becomes useful information. Sentiment analysis also analyzes some data to determine human emotions. Sentiment analysis is also used to predict sentiment polarity based on user sentiment data (Pang, Lee, & Vithyanathan, 2002).

3. METHODOLOGY

The data used in this research were 41087 data which were tweets and retweets against the Tiketcom twitter account which were taken by the crawling process from November 17th, 2018, to March 4th, 2019. Only one variable used was the “tweet_text” variable. Topic modelling was used to perform the analysis using Latent Dirichlet Allocation (LDA) and sentiment analysis.

4. RESULT

4.1. *Text Preprocessing*

Before further analysis is carried out, it is necessary to clean the text data with the following steps:

4.1.1. Case Folding

The first stage is carried out to change capital letters to lowercase letters, eliminating URLs, signs, and numbers. The following are the results obtained in Table 1 after carrying out the case folding stage on text data.

Table 1. Example of Case Folding Result

Tweet_text	**Clean_text**
As one of the first start-ups in Indonesia, https://t.co/Zdjh2A0WLj | @tiket is an application that is widely used by tourists to make ticket and hotel reservations. #tiketWonderfulIndonesia	As one of the first start-ups in Indonesia, ticketing is an application that is widely used by tourists to make ticket and hotel reservations #tiketwonderfulindonesia

4.1.2. Tokenizing

At this stage, every word that makes up a sentence is cut, so that later the words stand on their own. The following are the results of the tokenizing process in Table 2.

Table 2. Example of Tokenizing Results

Tweet 1		
As one of the first start-ups in Indonesia	@tiket is an application that is widely used by tourist	to make ticket and hotel reservations #tiketwonderfulindonesia

4.1.3. Feature Selection

This stage is the process of removing uninformative words in a sentence. These words are called stopwords. Stopwords include and, yes, I, which, with, therefore, by, and others.

So that if a document is carried out the feature selection stage will produce the output as in Table 3 as follows.

Table 3. Example of Feature Selection

Tweet 1	
As one of the first start-ups in Indonesia @tiket	is an application that is widely used by tourists to make ticket and hotel reservations #tiketwonderfulindonesia

4.2. *Weighting TF-IDF*

Based on the text data that has been obtained from the text preprocessing process, then weighting will be carried out on the text data using TF-IDF. Following are the results of the TF-IDF weighting in Table 4.

Table 4. Term-Frequency Results

No. Tweet	Index of Words			
	0	1	…	N
d1	1	2	…	0
…	…	…	…	…
d41086	0	0	…	0

Table 4 shows the TF value, where d1 is the 1st document and so on and then the word index is a different word. If a word has a value of 1, then the word appears 1 time in the document. After the TF value is owned, it is followed by calculating the DF value. DF is the number of documents containing the word. For example, a word with index 1 appears in 240 documents, then this value is called DF. If the DF value has been obtained, then look for the IDF value. The following is an example of calculating the IDF value for words with an index of 1.

$$IDF_j = log\ log(\frac{41086}{240}) = 2.2335$$

Then the calculation of the weight value for words with index 1 uses equation 1 as follows

$$W = 2\ x\ 2.2335 = 4.467$$

4.3. *Topic Coherence*

The results of the topic coherence bring up the coherence value which is useful for evaluating Topic Modeling, the better the model, the higher the coherence value. The following will display a coherence score table in Table 5 and the visualization of the graph in Figure 2.

Table 5. Coherence Score for Each Topic

Num Topics	Coherence Score	Num Topics	Coherence Score
1	0.693278	11	0.435694
2	0.581212	12	0.466943
3	0.492517	13	0.447322
4	0.456178	14	0.448882
5	0.575347	15	0.444101
6	0.495498	16	0.441446
7	0.56556	17	0.399199
8	0.466327	18	0.420927
9	0.475052	19	0.415531
10	0.445963	20	0.423318

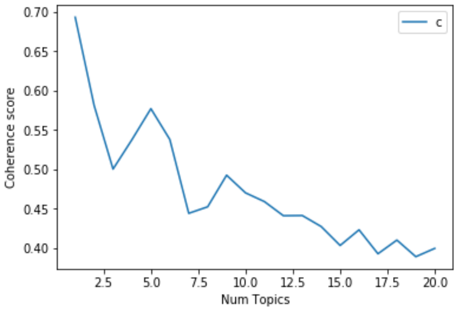


Figure 2. Visualization of the Coherence Score Graph for 20 Topics

Based on Table 5 and Figure 2, it is found that there are 20 topics in the document, with the highest score being topic 1. To continue in the next analysis, the researchers took 5 topics with the highest coherence score.

4.4. *Topic Modeling Using Latent Dirichlet Allocation (LDA)*

In determining the content on 5 topics with the highest coherence score, word associations are used as in Table 6 below.

Table 6. The Result of LDA Modeling on 5 topics with the Highest Coherence Score

Num Topics	Model LDA
1	0.052*"worldtour_kpop" + 0.046*"blackpink_coming" + 0.046*"meet_blinks" + 0.046*"ticketing_site" + 0.036*"inyourare_inyourareajakarta" + 0.023*"coming" + 0.023*"blackpink_coming_meet_blinks" + 0.023*"ticketing_site_inyourarea_inyourareajakarta" + 0.022*"meet" + 0.018*"kpop"
2	0.046*”siap” + 0.044*”siap_siap” + 0.039*”tiketcomotw_dapatkan” + 0.022*”tiketcomotw” + 0.021*”metro_tv” + 0.020*”tiketcomotw_dapatkan_diskon_tiketkemanapun” + 0.019*”siap_siap_samber_tiketmu” + 0.015*”tiketkemanapun” + 0.014*”tiketcomotw_tiketkemanapun” + 0.013*”ice_bsd”
3	0.017*”you_citiliveevent” + 0.012*”gfriendxclcinjakarta_brought” + 0.011*”more_info” + 0.011*”clc_cubeclc” + 0.010*”citiliveevent” + 0.009*”clc” + 0.009*”gfriendxclcinjakarta” + 0.009*”gfriend”
4	0.015*”terima_kasih” + 0.014*”beli” + 0.014*”samber_tiketmu” + 0.014*”diskon_tiketkemanapun” + 0.013*”ime_indonesia” + 0.011*”diskon” + 0.010*”promo” + 0.010*”harga” + 0.009*”min”
5	0.022*”penawaran_spesialnya” + 0.017*”kudus_relay” + 0.016*”cardholder” + 0.015*”kudus” + 0.014*”beli_mandiricard” + 0.014*”mandiridebit_dapatkan” + 0.014*”marathon” + 0.014*”beli_tiketnya” + 0.013*”relay” + 0.012*”week_beli”

The word association that appears on topic 1 leads to content regarding concert tickets for the Blackpink In Your Area Girlband concert sold by the Tiketcom application. Each word has a chance to appear in composing the topic, such as the word "worldtour_kpop" on topic 1 has a chance of appearing as much as 0.052 or 5.2%. Then for the second topic, it discusses a 50% discount on airplane and hotel tickets for the Surabaya, Medan, Makassar, and Yogyakarta areas. Then the 3rd topic discussed ticket discount offers with the theme "Tiketcom OTW". Furthermore, the fourth topic discussed the discount for purchasing tickets for the holy relay marathon. And the 5th topic discusses discounts on K-City Camp ticket purchases held by Citilive Event. The five topics both have content regarding discounts, it can be concluded that the public is interested in the discounts and vouchers provided by Tiketcom to its users.

4.5. *Sentiment Analysis*

Sentiment analysis was carried out to determine the positive and negative comments of the public on Tiketcom. The following are the results of the sentiment analysis obtained.

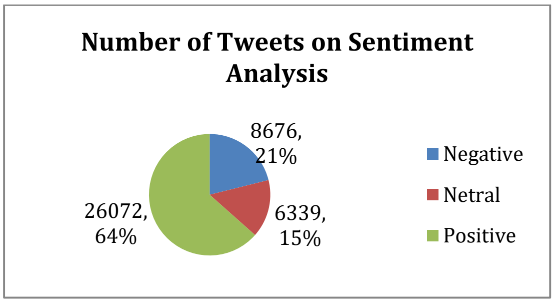


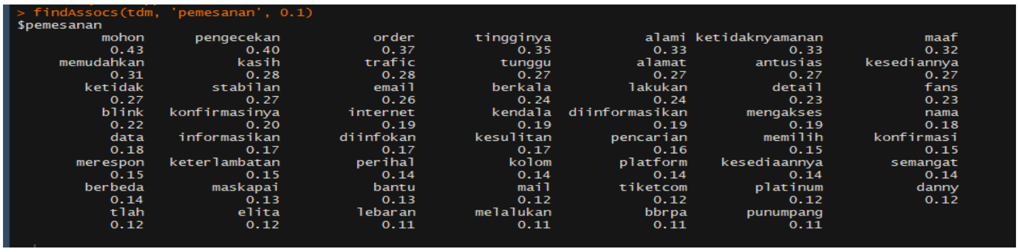
Figure 3. Sentiment Analysis Results for Positive, Negative and Neutral Comments

Based on Figure 3, there are 15% positive comments or 6339 tweets, 21% negative comments or 8676 tweets, and as many as 64% or 26072 are neutral tweets. To find out the essence of negative and positive comments, the following negative and positive word associations are used.

4.5.1. *Negative Word Association*

In negative comments, it is known that the meaningful word that appears the most is "pemesanan" so that the word is used to look for word associations and the following results are obtained.

Based on Figure 4, it is obtained the words related to the word "pemesanan", with a correlation of not less than 0.10. According to the negative word association above, it can be seen that the content or something that makes the public comment negatively about Tiketcom is the problem with ticket ordering on Tiketcom made by its users, such as late confirmation and so on.

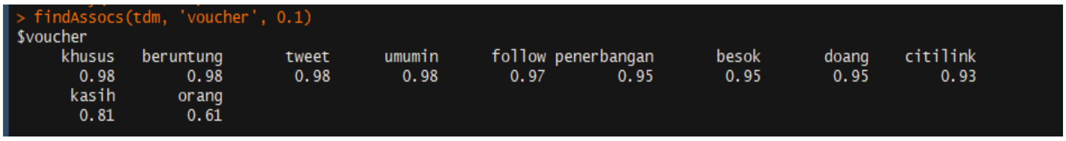


```
> findAssocs(tdm, 'pemesanan', 0.1)
$pemesanan
     mohon    pengecekan     order tingginya          alami ketidaknyamanan        maaf
      0.43          0.40      0.37      0.35           0.33            0.33        0.32
memudahkan         kasih    trafic    tunggu         alamat        antusias kesediannya
      0.31          0.28      0.28      0.27           0.27            0.27        0.27
   ketidak      stabilan     email   berkala        lakukan          detail        fans
      0.27          0.27      0.26      0.24           0.24            0.23        0.23
     blink konfirmasinya  internet   kendala diinformasikan       mengakses        nama
      0.22          0.20      0.19      0.19           0.19            0.19        0.18
      data  informasikan diinfokan kesulitan      pencarian         memilih  konfirmasi
      0.18          0.17      0.17      0.17           0.16            0.15        0.15
  merespon keterlambatan   perihal     kolom       platform    kesediaannya    semangat
      0.15          0.15      0.14      0.14           0.14            0.14        0.14
   berbeda      maskapai     bantu      mail       tiketcom        platinum       danny
      0.14          0.13      0.13      0.12           0.12            0.12        0.12
      tlah         elita   lebaran melalukan          bbrpa       punumpang
      0.12          0.12      0.11      0.11           0.11            0.11
```

Figure 4. Result of Negative Word Association

4.5.2. Positive Word Association

Meanwhile in positive comments, the word that appears the most is "voucher" so that the word is used to find the associated word association.



```
> findAssocs(tdm, 'voucher', 0.1)
$voucher
     khusus   beruntung       tweet      umumin      follow penerbangan       besok       doang    citilink
       0.98        0.98        0.98        0.98        0.97        0.95        0.95        0.95        0.93
      kasih       orang
       0.81        0.61
```

Figure 5. Result of Positive Word Association

Based on Figure 5 above with a correlation of not less than 0.10, the words related to the word "voucher" are *khusus*, *beruntung*, *tweet*, *umumin*, *follow*, *penerbangan*, *besok*, *doang*, *citilink*, *kasih*, *orang*. According to the word association, it is known that the thing or content that makes the public comment positively on Tiketcom is the voucher issued by Tiketcom for its users.

5. CONCLUSION

Based on the results obtained after analyzing the tweet data on the Tiketcom account using Latent Dirichlet Allocation (LDA) and sentiment analysis, several conclusions were obtained, namely, the Topic Modeling method using Latent Dirichlet Allocation can be used to find out which topics are trending in the community through comments. which is brought up to improve marketing planning, sentiment analysis and word association can be used to find out specifically what triggers users to comment negative or positive. discounts and promos. Meanwhile, according to negative comments, it was found that users experienced some difficulties in ordering tickets, such as the length of time when Tiketcom confirmed the order's email.

REFERENCES

1. *Tiket.com careers*. Retrieved from https://www.tiket.com/careers/teams.
2. Bhapkar, N. (2013). 8 KPIs Your Content Marketing Measurement Should Include. Content Marketing Institute. Retrieved from https://contentmarketinginstitute.com/2013/02/kpis-for-content-marketing-measurement/
3. Blei, D. (2012). Probabilistic Topic Model . Communications of the ACM, April 2012. 55(4): 77-84. DOI: 10.1145/2133806.2133826.
4. Cao, M., Zhang, Q., Seydel, J. (2005). B2C E-commerce website quality. *An empirical examination*. Industrial Management & Data System. 105(5): 645-661. DOI: 10.1108/02635570510600000
5. Feldman, R., and Sanger, J. (2007). *The Text Mining Handbook*. Cambridge: University Pers. New York.
6. Grossman, D., Ophir, F. (1998). Information Retrieval: Algorithm and Heuristics. Kluwer Academic Publisher.
7. Herwanto, G. B. (2018). Document Clustering Dengan Latent Dirichlet Allocation dan Ward Hierarichal Clustering. Pseudocode. 5(2): 29-37.
8. Hindle, A., Campbell, J. C, Hindle, A., Stroulia, A. (2014). Latent Dirichlet Allocation: Extracting

Topic. The Art and Science of Analyzing Software Data. 139-159.

9. Hoffman, T. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the TwentySecond Annual International SIGIR Conference on Research and Development in. SIGIR-99.

10. Hong, L., and Davison, B. (2010). Empirical Study of Topic Modeling in Twitter. 1st Workshop on Social Media Analytics (SOMA'10).

11. Liu, B., Gao, Y., Xu, Y., Li,Y. (2013). A Two-Stage Approach for Generating Topic Models. *Advances in Knowledge Discovery and Data Mining*. Berlin: Springer, Berlin, Heidelberg. 221-232.

12. Melita , R., Amrizal, V., Suseno, H., Dirjam, T. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Consine Similarity pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web. Jurnal Teknik Informatika. 11(2): 149-164.

13. Pang, B., Lee, L., Vithyanathan, S. (2002). Thumbs Up? SentimentClassification Using Machine Learning Techniques. Proceedings of The ACL-02 conference on Empirical methods in natural language processing. Stroudsburg: Association for ComputationalLinguistic.79-86.

14. Rosadi, D. (2014). Pemodelan Topik untuk Media Sosial Menggunakan Latent Dirichlet Allocation. Yogyakarta.

15. Sahu, S., Srivastava, S. (2016). Review of Web Document Clustering Algorithm. 3rd International Conference on Computing for Sustainable Global Development. INDIA Com. 1153-1155

16. Sun, X. (2014). Textual Document Clustering Using Topic Models. 10th International Conference on Semantics. Knowledge and Grids. 1-4

17. Zarrella , D. (2010). The Social Media Marketing Book. Canada: O'Relly Media.

18. Zulhanif, Sudartianto, Tantular,B., Jaya, I.G.N.M. (2017). Aplikasi Latent Dirichlet Allication Pada Clustering Data Teks. Jurnal Logika. 7(1): 46-51.