

Trend Topic Analysis using Latent Dirichlet Allocation (LDA) (Study Case: Denpasar People's Complaints Online Website)

Aulia Rizki Destarani¹, Isnandar Slamet², Sri Subanti³
^{1,2,3}Universitas Sebelas Maret, Surakarta 57126, Indonesia

ARTICLE INFO

Article history:

Received May 10, 2019
Revised July 13, 2019
Accepted July 22, 2019

Keywords:

Trend Topic
Topic Models
LDA
Gibbs sampling
Denpasar

ABSTRACT

According to the publication of the Central Bureau of Statistics 2017, the population of Denpasar people has increased to 914,300 people. The Increasing number of the population raises various problems that must be faced by the Denpasar's Government. The variety of problems is in line with the increase in complaints data posted through Denpasar people's complaints online website, which made it difficult to know the main topics of the problems. The purpose of this research is to find the main topics of complaints Denpasar residents quickly and efficiently. The method used to achieve the objective of the research is Latent Dirichlet Allocation topic models with Gibbs sampling parameter estimation. The number of topics obtained through the highest log-likelihood value -42,528.84, the value is in the number of topics 19. The trending topic was based on the highest topic probability, topic 4, with a topic probability value 0.055. Based on these results, the trend of a topic is on topic 4 which can be interpreted that many residents of Denpasar complained about damaged roads and requested to fix the roads.

*Copyright © 2019. Published by Universitas Ahmad Dahlan.
All rights reserved*

Corresponding Author:

Aulia Rizki Destarani
Department of Statistics,
Universitas Sebelas Maret,
Sutami Street 36A, Surakarta 57126, Indonesia
Email: aularizkide@gmail.com

1. INTRODUCTION

Denpasar is the capital of Bali Province, which is declared as the second-largest city in Eastern Indonesia. Denpasar position as the provincial capital has a multifunctional nature as a centre of government, business, and tourist destinations both from foreign and domestic tourists in the Bali region. Denpasar Central Bureau of Statistics stated that the population of Denpasar City is 788,587 people, while in 2017 the population of Denpasar City increased by 914,300 people. Domestic tourists motivated the selection of Denpasar City. Who was interested in visiting Denpasar, so it would be increased in the problems that arose in the community. Public service is needed to accommodate all complaints from Denpasar people on problems that arise. The original form of work by these needs, the Government of Denpasar, in collaboration with Satuan Kerja Perangkat Daerah (SKPD) created an application-based platform that provides Denpasar people complaints online website who named Pro Denpasar. The enthusiasm of the community after Pro Denpasar was formed by the amount of response data uploaded on the Pro Denpasar website. The variety of complaints from Denpasar residents is in line with the agreement of data collected. This explosion of data causes difficulty in discussing any topics discussed in Pro Denpasar. Manually extracting topics is not possible because it requires a long time. Therefore a method that can be obtained from the topic discussed with the topic of topic modelling is needed.

Topic models are one type of statistical model that aims to find a topic that appears abstractly from a collection of documents. Topic models are a useful exploratory tool. The topics provide a summary of the corpus that is impossible to obtain by hand, the per-document decomposition and similarity metrics provide a lens through which to browse and understand the documents. Topic modelling is often used in text mining to find hidden semantic structures, namely latent variables, while latent variables are hidden variables in observed variables, namely documents. One method used in topic modelling is Latent Dirichlet Allocation (LDA). That model determined from a discrete data collection. The data, such as documents that can be categorized as text or corpus data. One method for estimating the LDA parameter model is Gibbs sampling. Many other researchers use the Gibbs sampling algorithm as inference for various topic modelling.

This research was done concerning previous research. Myeong-Ha, et al. propose a trend analysis method using LDA. The method consists of several steps such as the construction of the topic tables, results of topic modelling, extraction of the keyword table in the topic, normalization of the occurrence rate by keyword, extraction of the graph of the trend by topic [1],[2]. Trend analysis is carried out based on topics using extracted results that combine LDA and top 10 keywords from each document and construct them as a topic table. Trend detection methods introduced so far have not used the location matching feature, thus lack integrating geographic locations information with the analyzed trend. Jayadharsini [3], use LDA to deliver trend notifications to customers who serve as a prominent way of advertising and target marketing. Zou [4], applied Latent Dirichlet allocation (LDA) model to extract 50 main topics and conducted trend analysis to explore the temporal popularity of drug safety data over the years. Yang and Rim [5], propose a novel topic model called Trend Sensitive-Latent Dirichlet Allocation (TS-LDA) that can efficiently extract latent topics from contents by modelling temporal trends on Twitter over time. Dan et al. made a predictive model for review of online hotel case using LDA method [6]. In another case, LDA was conducted data compression to derive latent shopping interest through website browsing habits [7]. Zhang et al. used a topic model for assigning metagenomic reads to different species or taxonomical classes [8]. The LDA model was applied to the reads and generated several hidden topics. Based on previous research, in this research applying LDA with parameter estimation to analyze trending topics from Denpasar residents complaint data.

2. RESEARCH METHOD

2.1 Text pre-processing

Text mining is the application of the concept of data mining techniques in searching for text patterns. The aim is to find useful information with specific objectives. The initial stage of text mining is text pre-processing, which aims to prepare text into structured data and can be processed at the next stage. Following are the stages in text pre-processing:

1. Deleting not important punctuation.
2. Case folding, which is the stage of converting text into lowercase letters as a whole.
3. Tokenization, which is the stage of cutting the input string based on each word that composes it.
4. Normalization of non-standard words becomes the correct word.
5. Removal of stopwords.
6. Feature selection, the stage of reducing the word dimension.

2.2 Determine number of topics for Latent Dirichlet Allocation model

Before forming the LDA model, steps must be taken is to determine the number of topics. The choice of the number of topics is necessary because it will affect the quality of the topics produced. Beyond data structure and implementation, the papers with trending topic employing LDA topic models can be distinguished along another dimension, namely, the way they evaluate results. All papers examined utilize quantitative techniques. The choice of the number of topics is a reference to evaluate the model. The range of these includes H-Scores, Cosine Similarities [9], the Kullback–Leibler Divergence [10], the Log Marginal Density [11], the Log-Likelihood [12], Perplexity [13], RMSE (Root Mean Square Error) [14]. Also, one can validate that topics capture a single concept by “reading several example documents” [15].

2.3 Latent Dirichlet Allocation (LDA)

LDA (Latent Dirichlet Allocation) is a flexible, unsupervised machine learning approach to soft-cluster big data [5]. The model assumes that textual documents consist of topics, which in turn consist of words from a wordlist [16]. LDA belongs to the family of mixed membership models [17]. LDA posits that each document d is a multinomial distribution θ_d over K topics, each of which is a multinomial distribution ϕ_k over the vocabulary. Here is the step of LDA's generative process :

1. For each topic $k \in [1, K]$
 - a. Draw word distribution $\phi_k \sim \text{Dir}(\beta)$
2. For each document $d \in [1, D]$
 - a. Draw topic distribution $\theta_d \sim \text{Dir}(\alpha)$
 - b. For each word $n \in [1, N_d]$
 - i. Draw topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - ii. Draw word $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$.

The global latent variables are topics $\{\phi_k\}_{k=1}^K$ and the local latent variables for each document d are topic proportion θ_d . Assigning word n of training document d to topic k is

$$p(z_i | z_{-i}, \alpha, \beta) \propto \frac{\beta_v + n_{k,v}}{(\sum_{v=1}^V \beta_v + n_{k,v})} \times \frac{\alpha_k + n_{d,k}}{(\sum_{k=1}^K \alpha_k + n_{d,k})} \quad (1)$$

Where $n_{d,k}$ is the number of tokens in the training document d assigned to topic k , and $n_{k,v}$ the number of times word type v assigned to topic k . Marginal counts are denoted by \cdot . At each training iteration i , we estimate the distribution over words $\hat{\phi}_k(i)$ of topic k as :

$$\hat{\phi}_k(i) = \frac{\beta_v + n_{k,v}(i)}{(\sum_{v=1}^V \beta_v + n_{k,v}(i))} \quad (2)$$

where the counts $n_{k,v}$ is taken at training iteration i .

At each iteration j in test chain i , we can estimate the topic proportion vector $\hat{\theta}_d(i, j)$ for test document d as :

$$\hat{\theta}_{d,k}(i, j) = \frac{\alpha_k + n_{d,k}(i, j)}{(\sum_{k=1}^K \alpha_k + n_{d,k}(i, j))} \quad (3)$$

where $n_{d,k}(i, j)$ is taken using sample j of test chain i .

2.4 Research data

Data is obtained by scrapping using Data Miner through the Pro Denpasar website. Data Miner is useful as a process of retrieving text data by scrapping on the desired website. The advantage of scrapping data using Data Miner is the creation of time efficiency because it does not need to apply code to a program data is taken by adjusting the recipe on the website.

2.5 Step of research

There are several stages to carry out this research. These stages are based on the research of Steyvers and Griffiths regarding the flow of the Gibbs sampling algorithm on LDA, here are the steps to get a LDA Model :

1. Scrapping data on the Pro Denpasar website with Data Miner.
2. Do the text pre-processing that aims to reduce the dimensions of the word in the data.
3. Turning data into a corpus in the form of document term matrix.
4. Do several steps to get an LDA model, here are the steps :
 - a. Determine the number of topics by looking at the highest log-likelihood value.
 - b. Take an initialization for the z topic randomly from $\{1, \dots, K\}$
 - c. Generate a topic probability distribution for each document. The distribution is known to have Dirichlet distribution with parameters, Dirichlet distribution is chosen as a prior distribution because the probability of the emergence of a topic illustrates the probability of the emergence of a probability vector in each document or it can be said that each topic that appears in each document is a probability vector.

$$\pi(\theta|\alpha) = \prod_{d=1}^D \left[\frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k-1} \right] \quad (4)$$

With

$$B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)} \quad (5)$$

- d. Generate probability distributions of words that are on the topic specified in each document. Distribution of words to the topic has Dirichlet distribution with parameters, Dirichlet distribution is chosen as a prior distribution because the probability of occurrence of words describes the probability of probability vectors in the topic specified in each document or it can be said that each word appears in the topic specified in each document is a probability vector. Where is the probability of v words on topic k that has been determined and written as follows:

$$\pi(\phi|\beta) = \prod_{k=1}^K \left[\frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\beta_v-1} \right] \quad (6)$$

With

$$B(\beta) = \frac{\prod_{v=1}^V \Gamma(\beta_v)}{\Gamma(\sum_{v=1}^V \beta_v)} \quad (7)$$

- e. Generate the probability distribution of topics on topics that have been determined in each document based on the words that appear in it. The distribution has a Multinomial distribution. Multinomial distribution is chosen because it describes the probability of the emergence of a topic between topic 1 to topic k that has been determined in each document.

$$\pi(z|\theta) = \prod_{d=1}^D \left[\prod_{k=1}^K \theta_{d,k}^{n_{d,k}} \right] \quad (8)$$

Where;

$$n_{d,k} = \sum_{v=1}^V n_{d,k,v} \quad (9)$$

Value of $n_{d,k}$ is the number of occurrences of the topic k determined from the words in the document d.

- f. Generate a probability distribution for words in the corpus (w) for the chosen topic. The probability distribution has a Multinomial distribution. The choice of Multinomial distribution is because it describes the probability of the appearance of a word between the words that appear in the corpus :

$$\pi(w|z, \phi) = \prod_{k=1}^K \left[\prod_{v=1}^V \phi_{k,v}^{n_{k,v}} \right] \quad (10)$$

Where;

$$n_{k,v} = \sum_{d=1}^D n_{d,k,v} \quad (11)$$

- g. Generate the probability distribution, then get the shared posterior distribution which is the LDA probability model :

$$P(w, z, \theta, \phi|\alpha, \beta) = \frac{P(w, z, \theta, \phi, \alpha, \beta)}{P(\alpha, \beta)} \quad (12)$$

$$= \left[\prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k + n_{d,k} - 1} \right] \left[\prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\beta_v + n_{k,v} - 1} \right]$$

5. Interpret topics on LDA Model.
6. Analyze the trend topic based on the highest topic probability. The probability of each topic for the entire document is based on equation (3).

3. RESULTS AND DISCUSSION

In this research, data were obtained by swiping the Pro Denpasar website using Google Chrome Data Extension Miner. Data was collected in the February 2018 to April 2018 period, and 508 complaints column data were obtained. The optimum LDA model is obtained by selecting a good number of topics. In this research, the optimal LDA Model is obtained with 19 Topics.

3.1. Word frequency visualization

Visualization of words frequency in the text data will be displayed in the following wordcloud :

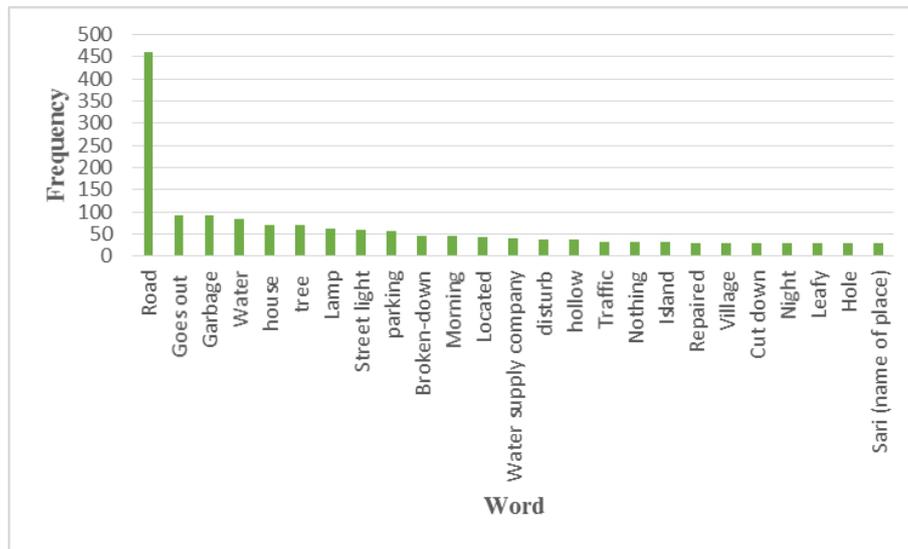


Fig. 1. The 25th most frequent words of Denpasar's inhabitant grievance data

[Fig. 1.](#) only shows the words that appear most often at least 10 times in text data. Based on [Fig. 1.](#) it was found that the most appear frequently words in Denpasar's inhabitant grievance data is the word “road”.

3.2 Text pre-processing

LDA is one of the methods included in the text mining category. All methods categorized into text mining need to be pre-processed. It aims to reduce the dimension of words in the data so that the data only contains leading and representative words. The first step, remove unnecessary punctuation such as commas, periods, question marks, exclamation marks, email addresses, website addresses, and emoticons. The next step is to change the text into lowercase letters as a whole. It aims to synchronize data so that it is easily analyzed. Then do tokenization, which is cutting the input string based on each word that composes it and normalizing non-standard words into standard words. Next step is removing non-essential words called stopwords, and researchers need to compile non-essential words based on previous research called stoplist. The final step do feature selection process that aimed at reducing the dimensions of the word. [Fig. 2.](#) depicts the data after and before the pre-processing stage.

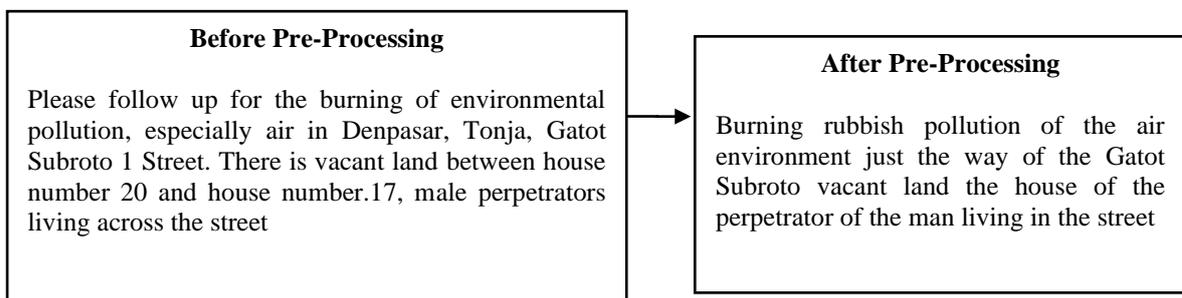


Fig. 2. Text Pre-Processing on the Data

3.3 Determine the number of topics

The beginning process of analyzing trending topic by topic modelling LDA is determining how many topics are right for the data. If the number of topics is chosen wrong, this can result in inadequate performance. In this research, the number of topics is obtained by looking at the plot between log-likelihood. The number of topics plot of number topics is represented in [Fig. 3.](#)

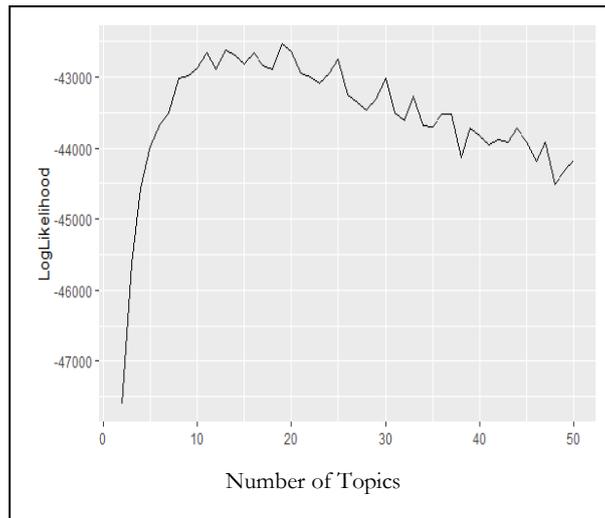


Fig. 3. Plot of the log-likelihood value versus the number of topics

Fig. 3. the highest log-likelihood value is -42,528.84 in the number of topics 19. The results will be used to build the LDA model with 19 topics. The table would be displayed the LDA topic models with 6 terms on each topic, and there are 19 topics on the model as follow:

Table 1. LDA Model with 19 Topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[1] tree	[1] community	[1] road	[1] road	[1] goes out
[2] cut down	[2] field	[2] little	[2] hollow	[2] street lights
[3] leaf	[3] puputan (name of place)	[3] to	[3] patch up	[3] repair
[4] sari (name of place)	[4] pro	[4] biak (name of place)	[4] broken	[4] slake
[5] wind	[5] goes out	[5] good	[5] hole	[5] little
[6] lush	[6]kenyeri (name of place)	[6] rai (name of place)	[6] patched	[6] checking
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
[1] can't	[1] road	[1] enter	[1] parking	[1] nothing
[2] renon (name of place)	[2] broken-down	[2] sidewalk	[2] car	[2] environment
[3] night	[3] house	[3] do	[3] disturb	[3] house
[4] sidakarya (name of place)	[4] repaired	[4] high	[4] house	[4] cable
[5] complaint	[5] report	[5] hole	[5] mountain	[5] gutter
[6]gelogor (name of place)	[6] sumerta (name of place)	[6] jammed	[6] transport	[6] mountain
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
[1] road	[1] water	[1] road	[1] road	[1] light
[2] morning	[2] water supply company	[2] morning	[2] grave	[2] traffic
[3] school	[3] house	[3] sanur (name of place)	[3] junior high school	[3] light
[4] children	[4] drainase	[4] broken	[4] tukad badung (name of place)	[4] red
[5] sign	[5] gutter	[5] jamming	[5] careless	[5] vehicle
[6] pipe	[6] sunday	[6] area	[6] endanger	[6] repair

Topic 16	Topic 17	Topic 18	Topic 19
[1] garbage	[1] island	[1] road	[1] village
[2] flood	[2] park)	[2] fix	[2] location
[3] place	[3] bali (name of place)	[3] residents	[3] permission
[4] operator	[4] telecommunaction company	[4] cause	[4] kaja (name of place)
[5] island	[5] point	[5] photo	[5] people
[6] garbage	[6]information	[6] located	[6] stin

Based on [Table 1](#), the LDA model is obtained, which contains a collection of words on each topic. The model formed is the LDA model with 19 topics and is composed of 6 words that have the highest probability of appearing on each topic. Arranged words on each topic are based on similarities, forming a typical topic of discussion. After analyzing the words on the LDA model, it was found the fact that not all words on the topic can be interpreted correctly because not all LDA models produce good convergence so that the words on the topic do not refer to the same discussion.. A topic said to be convergent if the distribution of the words in topic leads to a discussion of the same topic. Interpretations of each topic from the LDA model are stated in [Table 2](#).

Table 2. Interpretation of LDA Model

Topic	Intrepretation of each topic	Interpretable
1	Discuss about the demand to cut a shady tree because there is a lot of wind in front of sari mart on tukad yeh aya street	Yes
2	Request to follow-up dead lighting problem for several days at the Puputan Badung Field, it dies in days	Yes
3	Request for arrangement of small gardens near the river biak island street 1 and biak island street 2. After the end of the year floods a lot of garbage has already been overgrown with grass and wild plants	Yes
4	Many holes have to be patched and requests to the government to immediately repair the road.	Yes
5	Discuss the outage of street lighting and requests to immediately check and make repairs.	Yes
6	-	No
7	Explaining reports of damaged roads in Sumerta village and requests for repair	Yes
8	-	No
9	Motorists not park properly on the mountain road of Indrakila, Mount street, Bromo Mountain street and Slamet Mountain street because they endanger road users	Yes
10	-	No
11	Request for realization of school safe zone signs on imam bonjol street priest due to the increasing intensity of passing vehicles.	Yes
12	PDAM water drains are blocked in several houses	Yes
13	Damaged road area in Sanur causes traffic jams in the morning	Yes
14	There are drivers driving carelessly and speeding in the junior high school tukad badung street, dangerous for students going to school	Yes
15	Request to repair red traffic lights because the timer is not working properly	Yes
16	Request to add garbage transport officers, because the depo sometimes opens and closes in receiving garbage. this will have an effect on the accumulation of waste and cause flooding	Yes

17	Telecom cable on Merpati street and Bali street Number 2 is very disturbing because it is stuck in a pillar of the door of the house. They request PT.Telkom immediately be able to tighten the cable	Yes
18	Request to repair the broken road in front of rent car on Tukad Badung street. the road is not photographed because the streets are crowded	Yes
19	-	No
Total Interpretable Topic		15
Total Uninterpretable Topic		4

Based on Table 2, there are 4 topics, that difficult to review, namely topic 6, topic 8, topic 10 and topic 19 because the words were not discussed in the same discussion. In this research, validating a topic is done by dividing the number of topics that can be transferred by the total number of topics created in the LDA model. Based on Table 2, the number of topics that can be accommodated is 15 topics. Acceptable 78.95% of the discussion topics can be asked correctly according to the complaint data column uploaded via the Pro Denpasar website. Based on the same calculation, it was found that 21.05% of the topics could not be interpreted correctly, because the words contained in the topic did not match the discussion in the data. To improve the model, this allows the researcher to make appropriate parameter settings, sensible topics, and to assess the stability of outputs and covariate effects. A standard procedure is to initialize random inferences, triggering a repetitive process

3.4 Trend topic analysis

An outstanding advantage of using LDA trend detection is that it can well capture events with narrow topical scope. To determine the data trending topic in Denpasar complaints on Pro Denpasar website was done by looking at the highest probability value of each topic. The probability topic values from topic 1 to topic 19 are presented in Table 3.

Table 3. Topic Probability Value of 19 Topics

Topic	Topic Probability	Topic	Topic Probability
1	0,0546	11	0.0524
2	0,050	12	0.054
3	0,0518	13	0.052
4	0,055	14	0.0519
5	0.053	15	0.0533
6	0.0508	16	0.0543
7	0.052	17	0.0518
8	0.0518	18	0.0522
9	0.053	19	0.0515
10	0.0515		

Based on Table 3, the highest probability topic is on topic 4 is a value of 0.055. The terms on topic 4 consist of "jalan (road)", "berlubang (hollow)", "menambal (patch up)", "rusak (broken)", "lubang (hole)", and "ditambal (patched)". The term related to the data in each complaint column can be interpreted that many Denpasar residents complained about damaged roads and demanded the Denpasar government to repair the road.

4. CONCLUSION

Based on the LDA model with 19 topics found that the trending topic is on topic 4 with the topic probability value is 0.055. The term that appears in topic 4 means that the damaged road in this city is a trend in Denpasar residents' complaints data. There are 78.95% of the topics in the LDA model that can be interpreted and match the content of complaint data on the website Pro Denpasar, and there are 21.05% topics that cannot be interpreted correctly. It happened because not all LDA models produce good convergence so that the words on the topic do not refer to the same discussion. To overcome this can be done by determining the iteration value, the first state random value and the seed in the computational for LDA model. The results from this topic can be used as a reference for the Denpasar City Government to focus its performance in repairing damaged roads in Denpasar for immediate repair. The impact of this research for

Denpasar residents is to know the topic of discussion on the complaint data from February 2018 to April 2018.

REFERENCES

- [1] D. M. Blei and J. D. Lafferty, "Topic Models," *Text Min. Classif. Clust. Appl.*, 2009 available at: [Google Scholar](#).
- [2] J. Tang, Z. Meng, X. L. Nguyen, Q. Mel, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," in *31st International Conference on Machine Learning, ICML 2014*, 2014 available at: [Google Scholar](#).
- [3] J. Jayadharshini, R. Sivapriya, and S. Abirami, "Trend square: An Android Application for Extracting Twitter Trends Based on Location," in *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, 2018, doi: [10.1109/ICCTCT.2018.8551056](#)
- [4] C. Zou, "Analyzing research trends on drug safety using topic modeling," *Expert Opin. Drug Saf.*, pp. 629-636 2018, doi: [10.1080/14740338.2018.1458838](#)
- [5] M. C. Yang and H. C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Syst. Appl.*, vol. 41, no. 19, pp. 4330-4336, 2014, doi: [10.1016/j.eswa.2013.12.051](#).
- [6] N. Dan N, R. Bellio , and T. Reutterer, "A note on latent rating regression for aspect analysis of user-generated content", in *Proceedings of the 33rd International Workshop on Statistical Modeling*, vol. 1, Hrsg. Statistical Modelling Society , pp. 63-68, 2018 available at: [Google](#)
- [7] N. Schröder, "Using multidimensional item response theory models to explain multi-category purchases". *Market ZFP*, vol. 39, no. 2, pp.27-37, 2017 available at: [Google Scholar](#)
- [8] R. Zhang, Z. Cheng, J. Guan, and S. Zhou, "Exploiting topic modeling to boost metagenomic reads binning," *BMC Bioinformatics*, 2015 available at: [Google Scholar](#)
- [9] Z. Wang, L. Li, C. Zhang, and Q. Huang, "Image-regulated graph topic model for cross-media topic detection," in *ACM International Conference Proceeding Series*, 2015, doi: [10.1145/2808492.2808569](#).
- [10] D. Cao, R. Ji, D. Lin, and S. Li, "Visual sentiment topic model based microblog image sentiment analysis," *Multimed. Tools Appl.*, vol. 75, no. 15, pp.8955-8968, 2016, available at: [Google Scholar](#).
- [11] J. Büschken, and G.M. Allenby, "Improving text analysis using sentence conjunctions and punctuation," *SSRN*, 2017 available at: [Google Scholar](#)
- [12] K.B. Park, and S.H. Ha., Mining user-generated contents to detect service failures with topic model," *Int J Comput Electr Autom Control Inform Eng*, vol. 10, no. 8, pp. 1491-1496, 2016, doi: [10.5281/zenodo.1126073](#)
- [13] Z. Yang, A. Kotov, A. Mohan, and S. Lu, "Parametric and non-parametric user-aware sentiment topic models," in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 413-422, 2015, doi: [10.1145/2766462.2767758](#).
- [14] H. R. Iqbal, M. A. Ashraf, and R. M. A. Nawab, "Predicting an author's demographics from text using Topic Modeling approach," in *CEUR Workshop Proceedings*, 2015, available at: [Google Scholar](#).
- [15] M. E. Roberts, B. M. Stewart, and D. Tingley, "Navigating the local modes of big data: The case of topic models," in *Computational Social Science: Discovery and Prediction*, 2016 available at: [Google](#).
- [16] E. M. Airoldi, D. M. Blei, E. A. Erosheva, and S. E. Fienberg, "Introduction to mixed membership models and methods," in *Handbook of Mixed Membership Models and Their Applications*, 2014, available at: [Google Scholar](#).
- [17] A. Galyardt, "Interpreting mixed membership models: Implications of erosheva's representation theorem," in *Handbook of Mixed Membership Models and Their Applications*, 2014 available at: [Google Scholar](#).