Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 6 (2022) 998 – 1005 ISSN Media Electronic: 2580-0760

Topic Modeling for Support Ticket using Latent Dirichlet Allocation

Wiranto[1], Mila Rosyida Uswatunnisa[2]
[1, 2]Department of Informatics, Faculty of Information Technology and Data Science, Sebelas Maret University
[1]wiranto@staff.uns.ac.id, [2]milarosyidauswatunnisa@student.uns.ac.id

Abstract

In the business world, communication over customers must be built properly to make it easier for companies to find out what customers want. Support ticket is one of the business instrument for communication between the customers and the companies. Through a support ticket, customers can respond, complain or ask questions about products with a support team. Increasing the business process of the companies will be increasing the support ticket volume that should be handled by support team. It also has a value for analysis to get business intelligence decision. With that chance, an efficient data processing method is needed to find topics are being discussed by customers. One way that can be used to solve this problem is Topic Modeling. This research uses several parameters the number of topics, alpha value, beta value, iteration, and random seed. With this combination of parameters, the best results based on evaluation of human judgement and topic coherence with 5 topics, an alpha value of 50, a beta value of 0.01, 100 iterations, and 50 random seeds. The five topics interpretation consists of hosting migration, error problems in wordpress, domain email settings and domain transfer, ticketing and transaction processing. The total of 5 topics has a coherence value of 0.507897.

Keywords: Topic Modeling, Support Ticket, Latent Dirichlet Allocation.

1. Introduction

In the business world, communication with customers must be built properly to make it easier for companies to find out what customers want. Communication can be in chat, email, support ticket, and telephone. The support ticket is a company's communication in the form of text used to interact directly between customers and the support team through a digital platform [1].

The advantages of support tickets are that all communications between customers and the support team can be recorded [2]. Increasing number of customers will allow an overflow of support tickets to be received by the support team. So, it will be impossible for the support team to read every information on the support ticket individually to find out the information or problems that occur.

Machine learning technology is constantly developing, especially in obtaining the information contained in a sentence. Machine learning text analysis is grouped into 2 approaches, namely supervised and unsupervised approaches [3]. The unsupervised approach does not require a labelling process [4]. Techniques that can process data text are Clustering and Topic Modeling.

The clustering is based on text representation and distance between centroid and text. While the Topic Modeling method extracts topics from a set of documents based on statistical techniques, and each topic is defined as the distribution of a set of words [5]. Topic Modeling is a technique for finding unstructured data topics [6].

The application of Topic Modeling can be made in several methods, including Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM) [7]. Latent Dirichlet Allocation (LDA) method is suitable for large amounts of unstructured data by adding alpha (α) and beta (β) parameters to generate a probability distribution [8]. Latent Dirichlet Allocation (LDA) is a probabilistic generative model whose output is a topic consisting of words and their probabilities. Topic Modeling is used in various processing such as topic extraction, meaning disambiguation of words, text classification, and information retrieval [9].

Research on Latent Dirichlet Allocation has been carried out by previous researchers in topic modeling for road traffic information in Yogyakarta city based on twitter message [12]. The experiment determines the

Accepted: 26-09-2022 | Received in revised: 13-12-2022 | Published: 29-12-2022

initial topic which consists of 10, 20, 30, 40, 50, 60, 70, 80, and 90. The results of experiment are evaluated by the perplexity value of document and the perplexity value of each word to find out good topics. Based on the collaboration of two perplexity types, the value 0.36 indicates 60 topics that divided into 3 concept, namely the most discussed topics, topics that discuss monitoring traffic conditions and topics in the form of announcements on traffic conditions.

The Latent Dirichlet Allocation (LDA) technique for analyzing topic of online reviews for two competitive products also has been carried out by other researchers [24]. The data is review of wireless mouse and diffuser products on amazon.com website. The results obtained were evaluated based on the value of Jensen-Shannon Divergence (JSD), perplexity and Topic Coherence. For the first product, there are 10 positive topics consisting of cost performance, hand feel, size, ease of use, button and scroll wheel, USB receiver, wireless, battery, shipping speed, customer service and 7 negative topics consisting of cost performance, hand feel, button, scroll wheel, battery, USB receiver, replacement. As for the second product, there are 10 positive topics consisting of cost performance, hand feel and size, price, ease of use, bought before, USB receiver, wireless, battery (life), battery (switch), button and scroll wheel, and 6 negative topics consists of break, size, button, scroll wheel, USB receiver, wireless.

In another study, an analysis of anonymous social media posts was carried out by other researchers in identifying the topics discussed by students [25]. The dataset used is the National University of Manila facebook status and comments. The results of experiment are evaluated using topic coherence. Based on the highest topic coherence score, the main topics discussed by students were about enrollment problems and class problems.

Further related research was carried out by Sethasathien, et al, in data analytics on research project data [26]. This research aims to extract information from the Thailand research project. For evaluation the results of experiments, researchers used perplexity value and Topic Coherence. It was found that the research trending topics were on Agricultural and Social Science which consisted of 18 topics.

This research will examine how Topic Modeling with the Latent Dirichlet Allocation method can be applied to support tickets in Indonesia's domain hosting provider.

2. Research Methods

This research was conducted in four stages, including collecting customer support ticket dataset and then doing text preprocessing to eliminate unnecessary words. After the text is clean, the next step is to implements Topic Modeling using Latent Dirichlet Allocation. From the Topic Modeling obtained will be analyzed the results and evaluation. The stages of this research are represented in Figure 1.

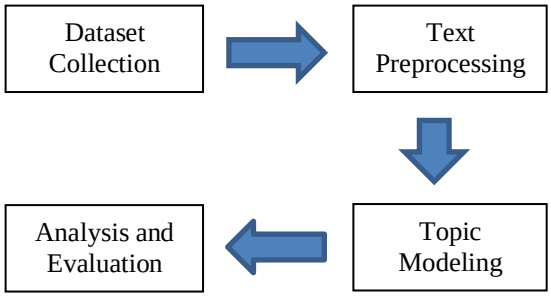

Figure 1. The Flow of Research Methods

2.1 Dataset Collection

The dataset of support ticket collected from the one of Indonesia's domain hosting providers. Support tickets are a communication instrument between customers and the company's support team [20]. Support tickets have a structure or format similar to chat and email. The support ticket structure includes greetings (customers write greetings), reports (customers report problems or communication content), and closings (customers write closing messages) [21].

Based on its contents, support tickets are divided into 3 types, including issues containing technical problems, requests containing customer requests, and feedback containing feedback that occurs between the customer and the support team [1]. Therefore, with ticket support, customers can immediately submit all responses in comments, suggestions, and complaints both technical and non-technical towards the product to the support team.

This is because customers have the right to freely provide reviews or responses to products offered [22]. The response sent by the customer via the support ticket, will be quickly replied by the support team.

2.2 Text Preprocessing

Text mining is used to find and process information in large amounts of text and unstructured data, and identify patterns and relationships between these patterns [10]. Text mining has several processing stages, such as text preprocessing, feature extraction, and mining [11]. The text preprocessing step consists of case folding, filtering, tokenization, normalization, and stemming. The feature extraction stage is the process of giving weight to words so that they can be processed.

Case folding is the change of capital letters to lowercase letters in sentences. Case folding aims to make sentences structured and easy to process.

Filtering is removing words that are considered unimportant, such as conjunctions and punctuation

DOI: https://doi.org/10.29207/resti.v6i6.4542

marks, attributes (tags, emoticons). This research also carried out deletion of email addresses, URL addresses, and extracting URL's containing guidance on support tickets. The filtering process is often referred to as the stopword removal process. Collection of meaningless words will be collected in a stopword dictionary stored in a file with .txt format.

Tokenization is a process in text preprocessing that to change sentences into tokens. Normalization is the process of converting abnormal words into normal or standard words.

Stemming is the process of removing prefix or suffix affixes to a word without removing the root word. In this study, using the stemming principle proposed by Nazief and Adriani is based on Indonesian words' morphology [23]. The steps of text pre-processing are represented in Figure 2.

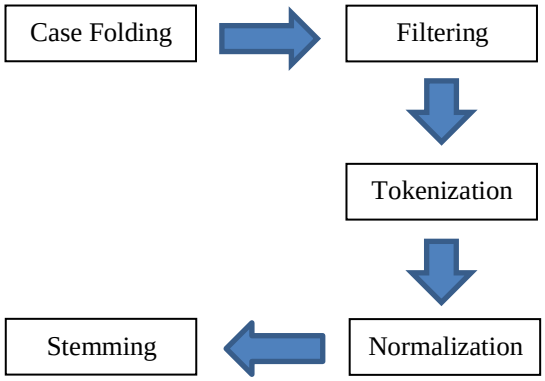

Figure 2. The Steps of Text Pre-Processing

2.3. Topic Modeling

Topic Modeling is a text mining technique to find topics in an unstructured set of documents [6]. Topic Modeling does not require a labelling process in documents [12]. Topic Modeling to identify a set of words from large data to generate topics based on the probability value distribution of each word in a document [13].

This stage uses Latent Dirichlet Allocation (LDA) approach. It is defined as a generative probabilistic model carried out on a set of data to perform Topic Modeling, where each topic is the distribution of each word [6]. The generative process includes 2 stages.

The first is to choose topics randomly based on the distribution of topics in each document. Second, for each word in the document, a random selection of topics is carried out based on the distribution of topics in the first step and word selection based on the distribution of words related to the selected topic. The LDA equation is written as formula 1.

$$P(W,Z,\theta|\alpha,\beta) = P(\theta|\alpha)\prod_{t=1}^{N} P(Z_n|\theta)P(W_n|Z_n,\beta) \quad (1)$$

W_n is the n^{th} word in the document, Z_n is topic for the n^{th} word in the document. Θ is distribution of topics in the document, φ is word distribution in the topic, α is topic distribution parameters in the document and β is word distribution parameters in the topic.

To get the probability value of a topic in a document and a word in the topic, need an approach namely Gibbs Sampling. Gibbs Sampling will perform repeated topic sampling of the data. Gibbs Sampling is used to train the system to interpret and generate information to get the probability value of words in the topic (φ), and the probability value of the topic in the document (θ).

The mathematical equation for Gibbs Sampling is defined in formula 2 and formula 3. [14].

$$\theta'^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\,\alpha} \quad (2)$$

$$\varphi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\,\beta} \quad (3)$$

C^{DT}_{dj} is he number of topic *j* defined in document *d*, C^{DT}_{dk} is the number of word defined as topics *k* in document *d* in the Gibbs Sampling process. C^{WT}_{ij} is the number of word *i* defined as topics *j* in a document, and C^{WT}_{kj} is the number of word *k* defined as topik *j* in a document in the Gibss Sampling process. Where w is total number of words in the corpus, α is value of alpha, β is value of beta and T is total number of topic.

2.4 Analysis and Evaluation

There are three approaches in Topic Modelling evaluation, such as human judgement, quantitative metric [15], and a mixture of human judgement and quantitative metric [16].

In machine learning, human judgement is method of evaluating an unsupervised learning model for exploration of text data. Topic Modeling is one of the unsupervised learning models. Topic Modeling aims to find topics in large text data. The result of Topic Modeling is the distribution or collection of words from a text data.

With human judgement, the word distribution of each Topic Modeling result can be interpreted [16]. In interpreting a topic, the human judgement method has 2 approaches consisting of word intrusion and topic intrusion [17]. A word intrusion approach evaluates whether a topic has human identifiable semantic coherence. Meanwhile, the topic intrusion approach evaluates whether the relationship between the document and the topic makes sense.

DOI: https://doi.org/10.29207/resti.v6i6.4542

Topic Coherence is a method used to evaluate the number of topics in Topic Modeling [18]. The word collection produced by Topic Modeling will be assessed based on the level of coherence like human interpretation [19]. Topic Coherence measures the value of a topic by measuring the level of semantic similarity between the words contained in the topic.

3. Results and Discussions

3.1 Results of Data Collection

The dataset collected is in .csv format on PT. Delta Neva Angkasa (DomainNesia) starting from January 2020 to February 2020. The amount of data is 16,000 raw text data. The example of data obtained is as shown in Table 1.

Table 1. Example of Support Ticket

No	Raw Text Data
1.	Dear domainnesia Beberapa waktu lalu saya sempat create tiket tetang ada orang yang merubah cpanel saya, ternyata 3 domain yang taruh di hosting ini di identifikasi sebagai web berbahaya. ketiga domain tersebut adalah: 1. baliparadisetours.com 2. wayansukerta.com 3. balipackagesescape.com saya sudah menghapus instalasi wordpress dan saya menemukan beberapa beberapa file php yg terselip di beberapa derectori dan saya sudah sertakan pada lampiran. Bebrapa hal yang saya tanyaka adlah: 1. Bagaimana hacker bisa masuk dan mengganti cpanel saya, bisa anda memberikan log nya? 2. Bagaimana mengebalikan sehingga web saya aman di kunjungi, saya sudah mengikuti petunjuk untuk keluar dari black list, namu sampai hari ini masih di blacklist situs2 saya. Saya menunggu penjelasan dari anda terimakasih Best regards wayan Sukerta
2.	Halo, Domain sudah kami pindahkan ke akun : juraganseblakladapisan@gmail.com Silakan cek kembali. Silakan hubungi kami jika membutuhkan bantuan lainnya. Terimakasih atas kepercayaan anda kepada DomaiNesia :) Salam, Adisty C Putri DomaiNesia Twitter: @DomaiNesia | Facebook: DomaiNesia Instagram : @DomaiNesia Suka dengan layanan DomaiNesia? Beri tahu orang lain dengan mengulas DomaiNesia di Google https://dnva.me/ulasdigoogle dan Facebook https://dnva.me/ulasdifacebook, terima kasih!

3.2 Results of Text Preprocessing

The first step in text preprocessing is case folding. This step is the change of capital letters to lowercase letters in sentences. Case folding aims to make sentences structured and easy to process. The example of this process result in this research can be seen in Table 2.

Table 2. Example of Case Folding

Before	After
Halo,	halo,
Berikut panduan untuk mengetahui kode epp domain anda :	berikut panduan untuk mengetahui kode epp domain anda :
https://www.domainesia.com /panduan/apa-itu-kode-epp/	https://www.domainesia.com/panduan/apa-itu-kode-epp/
Setelah itu bisa anda tunggu 1x24 jam, nanti akan dikirim otomatis kode eppnya ke email anda.	setelah itu bisa anda tunggu 1x24 jam, nanti akan dikirim otomatis kode eppnya ke email anda.
Silahkan hubungi kami jika membutuhkan bantuan lainnya. Terimakasih atas kepercayaan anda kepada DomaiNesia :)	silahkan hubungi kami jika membutuhkan bantuan lainnya. terimakasih atas kepercayaan anda kepada domainesia :)
Salam, Anjarini DomaiNesia Twitter: @DomaiNesia | Facebook: DomaiNesia | G+: DomaiNesiaPlus	salam, anjarini domainesia twitter: @domainesia | facebook: domainesia | g+: domainesiaplus
Suka dengan layanan DomaiNesia? Beri tahu orang lain dengan mengulas DomaiNesia di Google https://dnva.me/ulasdigoogle dan Facebook https://dnva.me/ulasdifacebook, terima kasih!	suka dengan layanan domainesia? beri tahu orang lain dengan mengulas domainesia di google https://dnva.me/ulasdigoogle dan facebook https://dnva.me/ulasdifacebook, terima kasih!

After case folding, the second step is filtering. This step is removing words that are considered unimportant, such as conjunctions and punctuation marks, attributes, tags, emoticons. This research also carried out deletion of email addresses, URL addresses, and extracting URLs containing guidance on support tickets.

The filtering process is often referred to as the stopword removal process. Collection of meaningless words will be collected in a stopword dictionary stored in a file with .txt format. An example of the filtering process in this research is explained in Table 3.

Table 3. Example of Filtering Result

Before	After
halo, berikut panduan untuk mengetahui kode epp domain anda : https://www.domainesia.com /panduan/apa-itu-kode-epp/	panduan mengetahui kode epp domain dikirim kode eppnya email hubungi membutuhkan bantuan kepercayaan domainesia domainesia domainesia domainesia layanan domainesia mengulas domainesia ulasdigoogle ulasdifacebook

DOI: https://doi.org/10.29207/resti.v6i6.4542

setelah itu bisa anda tunggu 1x24 jam, nanti akan dikirim otomatis kode eppnya ke email anda.

silahkan hubungi kami jika membutuhkan bantuan lainnya.

terimakasih atas kepercayaan anda kepada domainesia :)

salam,

anjarini

domainesia

twitter: @domainesia | facebook: domainesia | g+: domainesiaplus

suka dengan layanan domainesia? beri tahu orang lain dengan mengulas domainesia di google https://dnva.me/ulasdigoogle dan facebook https://dnva.me/ulasdifacebook, terima kasih!

The third step in text preprocessing is tokenization. It is a process to change sentences into tokens. An example of tokenization result can be seen in Table 4.

Table 4. Example of Tokenization Result

Before	After
panduan mengetahui kode epp domain dikirim kode eppnya email hubungi membutuhkan bantuan kepercayaan domainesia domainesia domainesia layanan domainesia mengulas domainesia ulasdigoogle ulasdifacebook	panduan, mengetahui, kode, epp, domain, dikirim, kode, eppnya, email, hubungi, membutuhkan, bantuan, kepercayaan, domainesia, domainesia, domainesia, layanan, domainesia, mengulas, domainesia, ulasdigoogle, ulasdifacebook

After tokenization, the next step is normalization. It is the process of converting abnormal words into normal or standard words. An example of the normalization process in this research is explained in Table 5.

Table 5. Example of Normalization Result

Before	After
[panduan, mengetahui, kode, epp, domain, dikirim, kode, eppnya, email, hubungi, membutuhkan, bantuan, kepercayaan, domainesia, domainesia, domainesia, domainesia, layanan, domainesia, mengulas, domainesia, ulasdigoogle, ulasdifacebook]	[panduan, mengetahui, kode, epp, domain, dikirim, kode, epp, email, hubungi, membutuhkan, bantuan, kepercayaan, domainesia, domainesia, domainesia, domainesia, layanan, domainesia, mengulas, domainesia, ulasdigoogle, ulasdifacebook]

The last step is stemming. It is the process of removing prefix or suffix affixes to a word without removing the root word. In this study, using the stemming principle proposed by Nazief and Adriani is based on Indonesian words' morphology. An example of a stemming process is shown in Table 6.

Table 6. Example of Stemming Result

Before	After
panduan, mengetahui, kode, epp, domain, dikirim, kode, epp, email, hubungi, membutuhkan, bantuan, kepercayaan, domainesia, domainesia, domainesia, domainesia, layanan, domainesia, mengulas, domainesia, ulasdigoogle, ulasdifacebook	panduan tahu kode epp domain kirim kode epp email hubung butuh bantu percaya domainesia domainesia domainesia domainesia layan domainesia ulas domainesia ulasdigoogle ulasdifacebook

3.3 Topic Modeling

In this research, we set a limit on the occurrence of N-gram 10 times. This is to find out how many words on N-gram affect a support ticket. The process of forming the N-gram is explained in the sentence "Sorry, our server is experiencing problems, this is still the process being handled by the support team", where the N-gram formed is shown in Table 7.

Table 7. Example of N-gram

No	N-gram Type	Formed Word
1.	Unigram	‘mohon’, ‘maaf’, ‘server’, ‘kami’, ’mengalami’, ‘gangguan’,’saat’, ‘ini’, ‘masih’, ‘proses’, ‘ditangani’, ‘oleh’, ‘tim’, ‘support’
2.	Bigram	‘mohon maaf’, ‘maaf server’, ‘server kami’, ‘kami mengalami’, ‘mengalami gangguan’, ‘gangguan saat’, ‘saat ini’, ‘ini masih’, ‘masih proses’, ‘proses ditangani’, ‘ditangani oleh’, ‘oleh tim’, ‘tim support’

Process of weighting text using the Bag of Words method uses MALLET in the generator library wrapper by Radim Rehurek in Python programming platform. The words in the dictionary will have the token id as the id of the word and the token frequency as the number of words in the document. An example of a Bag of Words shown in Table 8.

Table 8. Document Term Matrix Bag of Words

Document	Bag of Word
1.	[['pindah', 1], ['server', 1], ['gambit', 1], ['articuno', 1], ['server_gambit', 3]]
2.	[['transfer', 2], ['status', 1], ['informasi', 1], ['kena', 1], ['pending', 1]]

In this experiment of topic modeling, five parameters were determined randomly. These parameters are the number of topics, alpha value, beta value, iteration and random_seed. Parameter model is shown in Table 9.

DOI: https://doi.org/10.29207/resti.v6i6.4542

Table 9. Parameter Model

Parameter	Value
Number of Topic	2, 3, 4, 5, 6, 7, 8, 9, 10
Alpha	5, 10, 50
Beta	0.01
Iteration	100, 500, 1000
Random Seet	1, 10, 50

3.4 Analysis and Evaluation

After Topic Modeling was carried out, 243 models were obtained with predetermined parameter combinations. In each model in the results of this research, analysis and evaluation were carried out using 2 approaches, namely a qualitative approach with human judgement and a quantitative approach with the topic of coherence.

By looking at the set of words formed, it turns out that the results of the research with the parameters of the number of topics 5, an alpha value of 50, a beta value of 0.01, 100 iterations, and a random seed of 50 resulted in a collection of words related to each topic. Example of topic modeling result can be seen in Figure 3. The model with these five topics has a coherence value of 0.507897.

Figure 3. Set of Words in Five Topics

Topic 1 has a collection of words consisting of 0.04* "server" + 0.033*"cpanel" + 0.030*"akun" + 0.025* "pindah" + 0.025*"informasi" + 0.025*"website" + 0.017* "password" + 0.016*"guna" + 0.016*"ssl" + 0.016*"arah". This collection can be interpreted that the customer discusses migration hosting followed by how to install SSL on the server.

Topic 2 has a collection consisting of 0.021*"wordpress" + 0.020*"file" + 0.020*"error" + 0.017*"website" + 0.015*"email" + 0.014*"minta" + 0.013*"proses" + 0.013*"lampiran" + 0.011*"informasi_lengkap" + 0.010*"upload". This collection can be interpreted that customers discuss about errors that occur in WordPress.

Topic 3 has a collection consisting of '0.037*"akses" + 0.034*"email" + 0.018*"setting" + 0.018*"website" + 0.014*"cek" + 0.014*"kode_epp" + 0.013*"database" + 0.012*"buka" + 0.009*"dns" + 0.008*"tanya". This collection of words can be interpreted that the customer is discussing domain transfers accompanied by domain email settings.

Topic 4 has a collection of words consisting of 0.043*"cek" + 0.039*"ticket" + 0.039*"technical" + 0.030*"detail_ticket" + 0.027*"reviewdomainesia_ reviewfacebook" + 0.022*"baru" + 0.022*"status" + 0.018*"status_answered" + 0.016*"id" + 0.015*"cara". This collection of words can be interpreted regarding the content along with the status of the support ticket.

Topic 5 has a collection of words consisting of 0.040*"bayar" + 0.031*"transfer" + 0.027*"invoice" + 0.026*"bank" + 0.014*"bayar_jumlah" + 0.014*"bca" + 0.013*"aktif" + 0.013*"bank_mandiri" + 0.013* "bank_bca" + 0.013*"bank_tujuan". This collection of words can be interpreted regarding the payment transaction process through a support ticket.

To ensure the research results, we calculated topic coherence on the 4 parameters of the alpha value, beta value, iteration, and the combined random seed. Of these 4 parameters, 27 model combinations are produced.

The calculation of topic coherence aims to determine the highest topic coherence value in each model. The highest coherence topic will determine the number of topics used. The 3 highest topic coherence values were obtained. The three highest topic coherence values have indicators are shown in Table 10.

Table 10. Highest Topic Coherence Values

Parameter				Topics	Coherence Value
Alpha	Beta	Iteration	Random		
50	0.01	500	50	9	0.565449
50	0.01	1000	50	9	0.562123
50	0.01	100	50	9	0.540965

The coherence values in the three categories are almost the same. However, when these three values are evaluated with a human judgement approach, the three values have one topic that cannot be interpreted. The topic that cannot be interpreted occurs in the topic 9 with a collection of words as shown in Table 11.

In the first category with an alpha value of 50, a beta value of 0.01, 500 iterations, and a random seed of 50, there is a collection of words consisting of "tambah", "bayar_jumlah", "bank_tujuan", "setting", "kirim", "error", "pesan", "mail", "dns", and "mail_type". When evaluated using human judgement, the word collection cannot be interpreted, because there are several unrelated words (word intrusion). For example :

DOI: https://doi.org/10.29207/resti.v6i6.4542

- In the words "bayar_jumlah" and "bank_tujuan" it can be interpreted that the two words are about payment transactions.
- In the word "kirim", "pesan", "mail", "mail_type" about email.
- Meanwhile, the words "setting", "tambah", "error", "dns" regarding problems with dns settings.

It means that in one topic there are 3 interpretations. Such conditions in human judgement can be said to be bad or cannot be interpreted.

Table 11. All Words in Highest Topic Coherence Values

Parameter	Set of Words
50 0.01 500 50	'0.028*"tambah" + 0.027*"bayar_jumlah" + 0.024*"bank_tujuan" + 0.024*"setting" + 0.023*"kirim" + 0.022*"error" + 0.021*"pesan" + 0.018*"mail" + 0.017*"dns" + 0.017*"mail_tipe"'
50 0.01 1100 50	'0.028*"error" + 0.027*"tambah" + 0.027*"bayar_jumlah" + 0.024*"bank_tujuan" + 0.023*"setting" + 0.023*"server" + 0.022*"kirim" + 0.021*"pesan" + 0.018*"dns" + 0.017*"email"'
50 0.01 100 50	'0.028*"tambah" + 0.027*"bayar_jumlah" + 0.025*"error" + 0.024*"kirim" + 0.024*"bank_tujuan" + 0.023*"setting" + 0.021*"pesan" + 0.019*"dns" + 0.017*"mail" + 0.017*"server"' **(except "mail_tipe")**

Based on the analysis results and evaluation through between words is the most important thing. The quality of the best topic formation is determined based on the relationship between words according to the topic with the human judgement approach [14].

How high is the topic coherence value of a model if the model has one or more words cannot be interpreted in human judgement, then the model is still said to be bad [17]. With this case, so the best results will be found in the parameters used are 5 topics of topic, an alpha value of 50, a beta value of 0.01, 100 iterations, and a random seed of 50 with a topic coherence value of 0.507897.

4. Conclusion

Based on the research results, it can be concluded that the Latent Dirichlet Allocation (LDA) method in Topic Modeling can be applied to support tickets. In this research, this result can be used as a business decision-making instrument for the company. The best results will be found in the parameters used are 5 topics of topic, an alpha value of 50, a beta value of 0.01, 100 iterations, and a random seed of 50 with a topic coherence value of 0.507897.

The results chosen are determined by the highest score on the topic of coherence and by paying attention to human judgement. Because in Topic Modeling, human judgement is very influential in interpreting a collection of words into a topic. Interpretation of each topic on the number of topics 5 are, topic 1 the customer and the support team discuss hosting migration continued how to install SSL on the server. On topic 2, which is discussing problems with wordpress. Topic 3 discusses domain email and domain transfers. Topic 4 discusses about tickets. The last topic, topic 5, discusses customer transactions. For comparison, a suggestion for further research is to determine Topic Modeling using the Correlation Topic Model on the support ticket.

Reference

[1] C. C. A. Blaz and K. Becker, "Sentiment Analysis in Tickets for IT Support," in *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, May 2016, pp. 235–246.

[2] C. Werner, G. Tapuc, L. Montgomery, D. Sharma, S. Dodos, and D. Damian, "How Angry are Your Customers? Sentiment Analysis of Support Tickets that Escalate," in *2018 1st International Workshop on Affective Computing for Requirements Engineering (AffectRE)*, Aug. 2018, pp. 1–8, doi: 10.1109/AffectRE.2018.00006.

[3] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Jan. 2017, pp. 72–76, doi: 10.1109/ICoAC.2017.7951748.

[4] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Mar. 2017, pp. 216–221, doi: 10.1109/ICICCT.2017.7975191.

[5] J. A. Lossio-Ventura, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Clustering and topic modeling over tweets: A comparison over a health dataset," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 1544–1547, doi: 10.1109/BIBM47256.2019.8983167.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," p. 30, 2003.

[7] D. Chehal, P. Gupta, and P. Gulati, "Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations," *J Ambient Intell Human Computer*, Apr. 2020, doi: 10.1007/s12652-020-01956-6.

[8] A. Uteuov, "Topic model for online communities' interests prediction," *Procedia Computer Science*, vol. 156, pp. 204–213, 2019, doi: 10.1016/j.procs.2019.08.196.

[9] M. Allahyari and K. Kochut, "Discovering Coherent Topics with Entity Topic Models," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Oct. 2016, pp. 26–33 doi: 10.1109/WI.2016.0015.

[10] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge ; New York: Cambridge University Press, 2007.

[11] D. P. Langgeni and Z. A. Baizal, "Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection," p. 10, 2010.

[12] A. F. Hidayatullah and M. R. Ma'arif, "Road traffic topic modeling on Twitter using latent dirichlet allocation," in *2017*

DOI: https://doi.org/10.29207/resti.v6i6.4542

International Conference on Sustainable Information Engineering and Technology (SIET), Malang, Nov. 2017, pp. 47–52.
doi: 10.1109/SIET.2017.8304107.

[13] M. Hagras, G. Hassan, and N. Farag, "Towards Natural Disasters Detection from Twitter Using Topic Modelling," in *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, Nov. 2017, pp. 272–279, doi: 10.1109/EECS.2017.57.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Supplement 1, pp. 5228–5235, Apr. 2004
doi: 10.1073/pnas.0307752101.

[15] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014, pp. 530–539.
doi: 10.3115/v1/E14-1056.

[16] M. Omar, B.-W. On, I. Lee, and G. S. Choi, "LDA Topics: Representation and Evaluation," *Journal of Information Science*, p. 14, 2015.

[17] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," 2009.

[18] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, Aug. 2015.
doi: 10.1016/j.eswa.2015.02.055.

[19] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, Jun. 2010, pp. 100–108, Accessed: Oct. 03, 2020.
https://www.aclweb.org/anthology/N10-1012.

[20] L. Montgomery and D. Damian, "What do Support Analysts Know About Their Customers? On the Study and Prediction of Support Ticket Escalations in Large Software Organizations," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, Lisbon, Portugal, Sep. 2017, pp. 362–371
doi: 10.1109/RE.2017.61.

[21] G. Aalipour, P. Kumar, S. Aditham, T. Nguyen, and A. Sood, "Applications of Sequence to Sequence Models for Technical Support Automation," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 4861–4869, doi: 10.1109/BigData.2018.8622395.

[22] A. Angelpreethi and S. B. R. Kumar, "An Enhanced Architecture for Feature Based Opinion Mining from Product Reviews," in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, Tamil Nadu, India, Feb. 2017, pp. 89–92.
doi: 10.1109/WCCCT.2016.30.

[23] Nazief, Bobby dan Mirna Adriani, "Confix-Stripping : Approach to Stemming Algorithm for Bahasa Indonesia", in 2004, Faculty of Computer Science University of Indonesia.

[24] Wang, W., Feng, Y., Dai, W., 2018. Topic Analysis Of Online Reviews For Two Competitive Products Using Latent Dirichlet Allocation. Electron. Commer. Res. Appl. 29, 142–156. https://Doi.Org/10.1016/J.Elerap.2018.04.003

[25] Valencia, J.D.M., Laure, A.J.T., Centino, N.M.R., Fabito, B.S., Imperial, J.M.R., Rodriguez, R.L., De La Cruz, A.H., Octaviano, M.V., Jamis, M.N., 2019. Understanding Anonymous Social Media Posts Using Topic Modeling, In: 2019 IEEE 11th International Conference On Humanoid, Nanotechnology, Information Technology, Communication And Control, Environment, And Management (Hnicem). Presented At The 2019 Ieee 11th International Conference On Humanoid, Nanotechnology, Information Technology, Communication And Control, Environment, And Management, IEEE, Laoag, Philippines, pp.1-4.
https://Doi.Org/10.1109/Hnicem48295.2019.9072791

[26] Sethasathien, N., Prasertsom, P., 2020. Research Topic Modeling: A Use Case For Data Analytics On Research Project Data, In: 2020 1st International Conference On Big Data Analytics And Practices (Ibdap). Presented At The 2020 1st International Conference On Big Data Analytics And Practices (Ibdap), pp. 1–6.
https://Doi.Org/10.1109/Ibdap50342.2020.9245451

DOI: https://doi.org/10.29207/resti.v6i6.4542