IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.16, No.4, October 2022, pp. 335~344
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.74383

Topic Modeling on Online News Portal Using Latent Dirichlet Allocation (LDA)

Mohammad Rezza Fahlevvi*[1], Azhari[2]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: *[1]**m.rezza@mail.ugm.ac.id**, [2] arisn@ugm.ac.id

Abstrak

Jumlah berita yang ditampilkan di portal berita online seringkali tidak menunjukkan topik apa yang sedang dibahas, tetapi berita tersebut dapat dibaca dan dianalisis. Di Berita Anda dapat menemukan topik utama dan tren yang sedang dibahas. Anda memerlukan cara yang cepat dan efisien untuk menemukan topik yang sedang trending di berita. Salah satu metode yang dapat digunakan untuk menyelesaikan masalah ini adalah pemodelan topik.

Pemodelan tema diperlukan untuk memungkinkan pengguna memahami perkembangan tema modern dengan mudah dan cepat. Salah satu algoritma dalam pemodelan topik adalah Latent Dirichlet Allocation (LDA). Tahapan penelitian ini dimulai dengan melakukan pengumpulan data, preprocessing, pembentukan n-gram, representasi kamus, pembobotan, validasi model topik, pembentukan model topik dan hasil pemodelan topik.

Berdasarkan hasil evaluasi topik didapatkan nilai terbaik pemodelan topik menggunakan coherence terkait jumlah passes dan jumlah topik menghasilkan 20 passes, 5 topik dengan nilai 0,53 coherence value dan dapat dikatakan cukup stabil berdasarkan nilai standar coherence value.

Kata kunci— *Portal Berita, Pemodelan Topik, Latent Dirichlet Allocation, Coherence Value*

Abstract

The amount of News displayed on online news portals Often does not indicate the topic being discussed, but the News can be read and analyzed. You can find the main issues and trends in the News being discussed. It would be best if you had a quick and efficient way to find trending topics in the News. One of the methods that can be used to solve this problem is topic modeling.

Theme modeling is necessary to allow users to easily and quickly understand modern themes' development. One of the algorithms in topic modeling is the Latent Dirichlet Allocation (LDA). This research stage begins with data collection, preprocessing, n-gram formation, dictionary representation, weighting, topic model validation, topic model formation, and topic modeling results.

Based on the results of the topic evaluation, the The best value of topic modeling using coherence was related to the number of passes. The number of topics produced 20 keys, five cases with a 0.53 coherence value. It can be said to be relatively stable based on the standard coherence value.

Keywords—*News Portal, Topic Modelling, Latent Dirichlet Allocation, Coherence Value*

Received April 26th,2022; Revised October 13th, 2022; Accepted October 30th, 2022

1. INTRODUCTION

The development of the information age demands the ability of each individual to locate and use information effectively and efficiently. The rapid growth of information technology has dramatically changed the form and medium of information expression. Today information is available in print and various electronic media/forms that can be accessed via the Internet. Undoubtedly, this information has become an essential commodity in today's modern world. There are various information providers, one of which is an online news portal. Online news portals are one of the most famous and unique types of mass media. Online news portals are characterized by modern, real-time, convenient information technology networks using computer equipment. One of the tasks of online news portals is to provide readers/users with engaging and fast updates to receive complete and up-to-date information. Of the many news items posted online in Indonesia, the topic under discussion is often not displayed, but news items can be used for analysis rather than just reading. In the News, you can find the main topics and trends being Discussed. This allows you to infer hidden information that can be used as evaluation material or information. Online news portals have so many messages that it's impossible to read them and draw conclusions. So we need a quick and efficient way to determine what topics are trending in the News. One way to solve this problem is topic modeling.

Topic modeling is used to obtain topics from various news stories on online news portals. Topic modeling can be described as finding groups of words (topics) from documents that can well represent the information contained in the document [1]. One of the algorithms in modeling such topics is the Latent Dirichlet Allocation (LDA). LDA is a mechanism used for topic extraction [2]. LDA has been widely developed in research on analyzing a text or document topic. Research related to using the LDA method is many ways to analyze trends in a topic with various sources. It can access Google Scholar, Wikipedia, Twitter, Instagram, Facebook, journals, and so on.

There is a related study [3] that analyzes drug safety trends. The results of this study display popular research topics based on year, topic distribution, clustering, and sources obtained in this study from only one journal and taken from each abstract. The related research [4] aims to model information topics that can automatically classify social media messages into issues that arise from modeling results. The source of the research data was taken from Twitter. Subsequent research that analyzes information content in the form of News is produced with a considerable number of births in the media daily. This study aims to model a combination of document clustering techniques[5]. Related research then analyzes how to view e-commerce social media content. The source of the data obtained is Instagram. The purpose of the study looked at the topics discussed by looking at the overall positive and negative sentiments contained in shopee Instagram e-commerce[6].

The LDA model is generative A probabilistic model of the corpus. The basic idea is a document represented as a mixed model of various topics called latent; each issue is characterized by the word [7]. The latent (hidden) variables explain the probability model's observed (observed) variables. The observation variable is the document, while the latent variable is the specified topic of each word in the paper. Complex distributions make estimating posterior distributions for LDA models complicated manually [8]. Based on this background, the proposed research will focus on modeling the topic. Topic modeling provides an informative topic description that users can immediately accept. That way, users can easily and quickly understand the latest issue developments. This study uses the LDA topic model method, which has proven to be one of the effective unsupervised learning methodologies for finding different topics in a document set. Cases from data processing will then be evaluated on the issue of perplexity and coherence, a test of the relationship between the descriptions of the probability of words found with each other in composing a problem [9]. The results of topic modeling analysis using LDA are expected to help understand what topics are trending on online news portals tempo.co more concisely.

2. METHODS

2.1 System Analysis

Latest information. This system was developed for the topic modeling of an online news portal. Online News portals are a popular form of mass media that, among other things, have a distinctive role in providing exciting and up-to-date information for readers/users in an exclusive way. Among the many news items offered by Indonesian online news portals, the discussed topic is often not displayed, but the News is not just for reading. It can be used for analysis. In the News, you can find discussed the most important topics and trends. This way, the hidden information can be derived and used as evaluation material or information. Online news portals have so many messages that it's impossible to read them one by one. So it would help if you had a quick and efficient way to find trending topics in the News. One method that can be used to overcome these problems is topic modeling. Topic modeling aims to find the main issues related to News in online news portals.

Topic modeling is desirable in online portals to provide informative topic descriptions that users (in this case, news readers) can directly accept. This allows users to easily and quickly track the development of the latest topics. This study uses the LDA topic model method. It has proven to be one of the most effective unsupervised learning methods for finding different issues in a collection of documents [7]. The problems resulting from the data processing are subjected to a topic consistency test, that is, an examination of the relationship of the probabilities of the words found when compiling the topics [9]. The results of our topic modeling analysis using LDA are expected to help us better understand the trending topics of the online news portal tempo. co. The study used news data from an Indonesian online news portal, tempo. Co, from May 2018 to July 2019, sent 11,466 messages. Determining the central theme in theme modeling uses the LDA algorithm. A rating is given to measure the performance of his LDA method on documents based on perplexity and topic coherence.

2. 2 System Architecture

This research has two main processes, namely, data collection using scrapping techniques, preprocessing, which has five stages, n-gram, dictionary representation, weighting, and LDA process consisting of topic validation and the formation of a topic model. The processes that occur in this system can be seen in Figure 1.

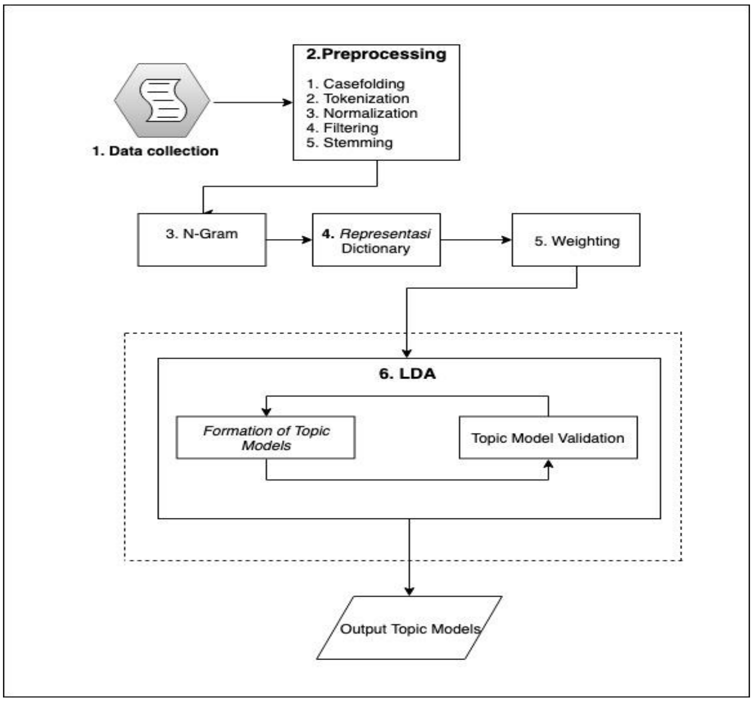


Figure 1 System Architecture

The following is a systematic explanation in Figure 1:

1. Data Collection, data were taken using the scrapping method. After the data is obtained, then it is saved into CSV. Then the data will be processed to the next stage, namely preprocessing.
2. Preprocessing, the data obtained is then carried out to clean the disturbance of disturbing characters and words that are not important. This process is called the preprocessing process. The preprocessing process includes case folding, tokenization, normalization, filtering, and stemming.
3. N-Gram, the data that has passed the preprocessing stage will then pass the feature selection stage using TF-IDF
4. Dictionary Representation, the data that has passed the preprocessing stage will then give the feature selection stage using TF-IDF.
5. Weighting, the data that has passed the preprocessing stage will then give the level of feature selection using TF-IDF.
6. LDA, When modeling a topic using LDA, you need to enter two parameters: the path or iteration parameter and the number of topic parameters.

2. 3 Data Collection

Data collection was obtained from tempo. co by scrapping. The scrapping process is carried out by taking data in text, generally HTML or XHTML type. This study uses the python library, Beautiful Soup. Data is taken in the form of news data on online news portals in Indonesia, namely tempo.co starting from May 2018 to July 2019, totaling 11,466 news.

2. 4 Preprocessing

Preprocessing is intended to process the scrapping data from the comment form and separate it into each constituent word. In addition, the data preprocessing process is an essential step in the data mining process. Aim for the raw data used for further processing [10]. After that, the data is extracted into crucial words to get the most critical part of a document, represented as an array of terms. Forming a variety of dishes aims to prepare and form data into a document arrangement by a predetermined structure and is then ready for processing. There are several sub-processing stages: case folding, tokenization, normalization, filtering, and stemming.

2. 5 N-Gram

N-gram is a contiguous sequence of items from a given line of a text or speech [11]. Ngram is a method that is applied for word or character generation. This n-gram method takes n-character fragments of several n words that are continuously read from the source text to the end of the document. The tokenization can be in the form of unigram, bigram (2-gram), trigram (3- gram) to N-gram. Unigram or 1-gram divides documents into tokens consisting of 1 word. The value of N on the N-gram indicates the distribution of tickets consisting of N words.

2. 6 Representatives Dictionary

Topic modeling requires changes to data in the form of dictionaries and corpora. A glossary is an a data format containing a unique indexed set of words, making it easy to view the comments in your model. A canon is a data format in the form Of a term matrix document later used to perform modeling experiments.

2. 7 Weighting

This level gives weight to each word in the content that changes in the form of vector space. This study uses the term Frequency-Inverse Document Frequency (TF-IDF) to transform words into vectors. Term frequency (TF) counts the number of occurrences in each word (time) and each document. In contrast, inverse document frequency (IDF) distributes values across the collection of affected documents to reduce the number of individual words that should not fluctuate considerably. TF-IDF results in the multiplication of the TF value and the IDF value.

Moreover, according to [12], TF-IDF is also the most popular and effective feature extraction method for reducing the spatial dimensionality of text features. TFIDF is commonly used to measure the importance of words in text sentences containing comments TF-IDF is a product of TF and IDF. Our TF-IDF weighting was done to determine the term weights in the tempo messages. Co, the word weights representing the information obtained are used as features for topic modeling. Extra care is taken to avoid excess functions when computing the TF-IDF.

2. 8 Topic Model Validation

The topic validation phase aims to ensure that the topic model that results from the modeling topic carried out in the document is correct, both in the form of topics and words in the topic. The things that are analyzed in the validation stage of the model topic are:

- Appropriate number of iterations to form the theme of the model
- The number of suitable topics depends on the distribution of impotence and coherence.
- A probability distribution for each training document associated with the formed topic model.

Perplexity is a method used to test the accuracy or suitability of information from documents to the topics produced. Perplexity takes n samples from N population data to be tested whether the n samples have topic suitability with topic groups in N populations [13]. One form of topic evaluation is topic coherence which emphasizes the ease of interpretation in humans, where a set of words produced by the topic model is assessed based on the level of coherence or the level of simplicity in human understanding [9]. Existing topic coherence methods calculate coherence scores based on the semantics of words related to the topic [14]. This measurement helps differentiate between topics that can be asked and statistics-related topics.

2. 9 Formation of Topic Models

At the stage of the Topic modeling process with LDA, the primary step is forming a model using the Gensim library. Then the model is evaluated by evaluating perplexity and coherence. In developing the experimental model carried out on the input parameters. The model search results will be used to get any topic arising from the document's analysis. After there is a topic model, an evaluation of perplexity and coherence is carried out in the library in python. Models that show more minor and unchanging results will be selected as temporary models to evaluate confusion. For coherence, models that show more extensive and stable results will be chosen as models.

In forming the topic model using LDA, there is a process in getting the distribution of words in the topic using the Gibbs Sampling algorithm where the algorithm starts by randomly assigning topics to each word, with each word having a probabilistic weight according to the frequency of occurrence in the document. The following is the calculation of the Gibbs Sampling Algorithm on the LDA.

Parameters used:

alpha	: 0.1	*threshold*	: 0.001
beta	: 0.01	iteration	: 1

number of topics: 2

Table 1 *Bag of Word* (BoW)

Bag Of Word		
Id_token	doc	word
1	1	commission_eradication
2	1	billion
3	2	bribe
4	2	money
5	2	project
6	2	criminal
7	3	corruption
8	3	law
9	3	police

The data used as calculations can be seen in Table 1. Vocabulary formation is taken from the words in the document by taking a unique word.

Table 2 *Vocabulary*

Vocabulary	
id	***word***
0	commission_eradication
1	billion
2	bribe
3	money
4	project
6	criminal
7	corruption
8	law
	police

Table 2 shows the vocabulary or dictionary used. Data Dictionary is a catalog of facts about data and information needs of an information system.

Table 3 Random Topic Initialization

id_token	*doc*	*word*	topik (*random*)
0	1	commission_eradication	2
1	1	billion	2
2	2	bribe	1
3	2	money	1
4	2	project	2
5	2	criminal	2
6	3	corruption	2
7	3	law	2
8	3	police	2

The first step is to initialize topics for each token randomly. If there are two topics, then initialize topic one and topic 2. This initialization can be seen in Table 3.

Using collapsed Gibbs sampling, the iteration will stop if it meets one of the conditions, namely:

1. Iteration reaches a predetermined number of iterations.
2. Predetermined threshold as the value if the difference is the difference from perplexity the i-th iteration with the perplexity value of the 1st iteration (i-1) is less than the predetermined threshold. Annotated like ((perplexity i – perplexity (i-1)) < threshold).

Calculate topic word probability (PWZ)

PWZ word (commission_eradication):

- Word (commission_eradication), topic1

$$\varphi_{j,t} = p(w = t|z = j) = \frac{n_j^{(t)} + \beta_t}{\sum_{t=1}^{v} n_j^{(t)} + w\,\beta_t} \qquad (2)$$

$$= \frac{n_1^{(1)} + \beta_1}{\sum_{t=1}^{v} n_1^{(1)} + w\,\beta_1} = \frac{0+0.01}{2+(8*0.01)} = 0.004808$$

- Word(commission_eradication), topic 2

$$\varphi_{j,t} = p(w = t|z = j) = \frac{n_j^{(t)} + \beta_t}{\sum_{t=1}^{v} n_j^{(t)} + w\,\beta_t}$$

$$= \frac{n_2^{(1)} + \beta_1}{\sum_{t=1}^{v} n_2^{(1)} + w\,\beta_1} = \frac{1+0.01}{7+(8*0.01)} = 0.142655$$

The calculation of all words and produce results $PWZ_{j,t}$:

Table 9 Probability of Word Topics (PWZ)

topic	commission_eradication
1	0.0048
2	0.1427

3. RESULT AND DISCUSSION

This chapter will explain the results of the topic modeling validation, the topic modeling results, and the topic modeling visualization obtained from the implementation process discussed in the previous chapter.

3. 1 Topic Model Validation

Before creating topic modeling, test to determine the number of topics and iterations (passes) used in the topic-building process. In determining the number of subjects and number of replicates (runs), the method used in the scoring process uses coherence score analysis. When forming an experimental model that is run with input parameters. Experimental results are used to obtain an evaluation value, i.e., coherence. Consistency evaluation selects the model as the one that shows more excellent results and does not change.

3. 2 Determination of Amount of Iteration on Coherence Value

The method used in determining the number of iterations (passes) is to analyze the coherence values. The analysis is performed by modeling the theme with at least three different theme parameters with an initial pass value of 30. In this case, the parameters for the selected subjects are 5, 10, 15, and 20. The coherence value results emerging from each parameter of several subjects are then recorded and visualized in Figure 2.

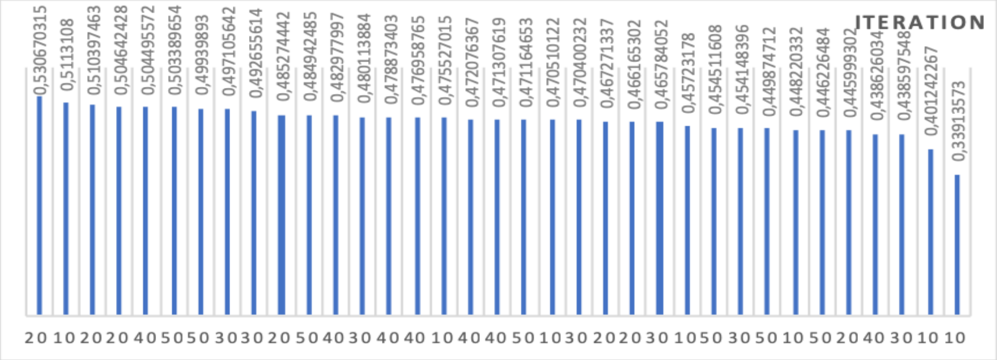


Figure 2 Amount of Iteration on Coherence Value

Based on the visualization of figure 2 determining the number of iterations, it can be seen that the coherence value has reached a stable tendency and gets the highest value at the 20th pass for all parameters of the number of caps with a coherence value of 0.53, so it can be concluded that the best iteration is based on the coherence evaluation is 20.

3. 2 Determination of the Number of Topics Against Coherence Value

Determination of the Some topics are done by analyzing the coherence value. Still, the analysis of the coherence value in the context of determining the number of issues is carried out by experimenting with the parameter number of topics in a broader range of values. In this case, the content of values used in the experiment is shown in the table below. Coherence value analysis in the context of determining the number of topics is done by running 30 times to get an average perplexity value that is accurate for each parameter of the number of

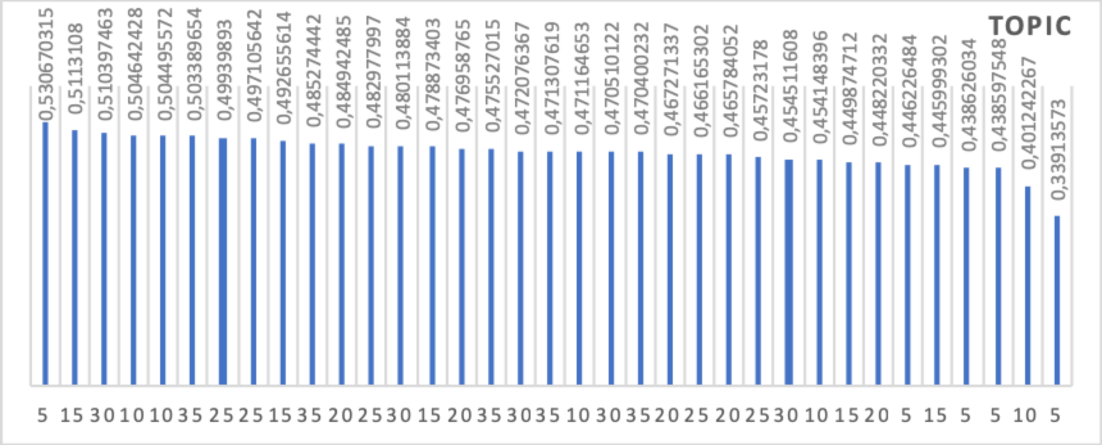


Figure 3 Number of Topics Against Coherence

Based on Figure 3, the highest coherence value is found in the number of topics five, which is 0.53, and the trend of coherence value increases for the number of topics that are getting higher. Hence, five topics are the best number based on the analysis of coherence values.

3. 3 The results of the formation of the LDA Model

After determining the evaluation used for the topic modeling process using coherence with 20 iterations, the number of topics five, and an evaluation value of 0.53, we will get the probability distribution of words in the 5 Topics (10 words per topic) of the model through the coherence value validation process can be seen in Table 6.2.

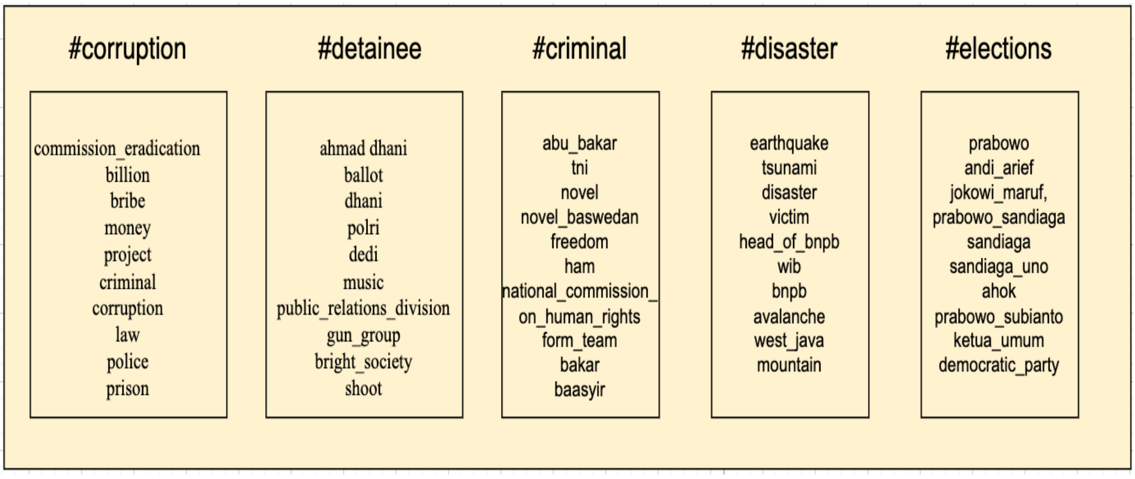
Table 9 Probability of Word Topics (PWZ)

Topic No	Topic Word Distribution
1	commission_eradication, billion, bribe, money, project, criminal, corruption, law, police, prison
2	Ahmad Dhani, ballot, Dhani, polri, dedi, music, public_relations_division, gun_group, bright_society, shoot
3	abu_bakar, tni, novel, novel_baswedan, freedom, ham, national_commission_on_human_rights, form_team, Bakar, baasyir
4	earthquake, tsunami, disaster, victim, head_of_bnpb, wib, bnpb, avalanche, west_java
5	prabowo, andi_arief, jokowi_maruf, prabowo_sandiaga, sandiaga, sandiaga_uno, ahok, prabowo_subianto, chairman, democratic_party

The topic modeling results produce word distribution that forms meaningful topics similar to word-topic distribution according to the number of topics chosen. Each word has a probability of each issue and the relevance of one topic to another. For example, the word tsunami has the probability of each issue with the highest probability on topic four. It is similar to the case of word distribution on topic four, such as earthquake, disaster, victim, and others.

3. 4 Definition of Topic

After the topic consists of the distribution of words, it is necessary to define the topic based on the linkages of words in each topic. This definition is done manually so that we can easily describe a topic. For example, the topic of tahanan_polri has words about ahmad_dhani with high probability and so on for determining topics according to word distribution. After the LDA topic model has been created, a document or article can be determined on the distribution of topics describing the document's collection of words. LDA uses a bag of words assumption, i.e., the order in which words appear in the document is ignored. This assumption is recognized as unrealistic, but it is reasonable because the purpose of LDA is only to find the semantic structure of the text. An example is shown in Figure 4.



#corruption	#detainee	#criminal	#disaster	#elections
commission_eradication	ahmad dhani	abu_bakar	earthquake	prabowo
billion	ballot	tni	tsunami	andi_arief
bribe	dhani	novel	disaster	jokowi_maruf,
money	polri	novel_baswedan	victim	prabowo_sandiaga
project	dedi	freedom	head_of_bnpb	sandiaga
criminal	music	ham	wib	sandiaga_uno
corruption	public_relations_division	national_commission_on_human_rights	bnpb	ahok
law	gun_group	form_team	avalanche	prabowo_subianto
police	bright_society	bakar	west_java	ketua_umum
prison	shoot	baasyir	mountain	democratic_party

Figure 4 Definition of Topic

4. CONCLUSIONS

The following conclusions can be drawn based on observations, investigations, and analysis of the results obtained. The results of LDA topic modeling on online news portals over some time have successfully formed topics containing information Or topic descriptions. The topic description for each story indicates that those topics were discussed or preached by tempo.co readers within a specific time frame. Given the number of issues specified in the message data, you can create a collection of words that make up the topics appropriately. Use the information that represents current content to help readers and relevant stakeholders understand news developments and popular subjects. Various experiments were conducted to model the case using the LDA method. From the results of this experiment, the best score uses the coherence of the number Of runs, and the number of generated topics is 20 runs, five cases with a coherence value Of 0.53, it can be said that it is fairly stable against the reference value.

REFERENCES

[1] Nair G. 2016. Text mining 101: Topic modeling Aug. 2019 [Online]. Available: http://www.kdnuggets.com/2016/07/text-mining-101-topic- modeling.html. [Accessed: 16-Feb-2019]

[2] Korzycki, M., Gatkowska, I., Lubaszewski, W., 2017. 2 - Can the Human Association Norm Evaluate Machine-Made Association Lists?, in Sharp, B., Sèdes, F., Lubaszewski, W. (Eds.), Cognitive Approaches to Natural Language Processing. Elsevier, pp. 21–40. https://doi.org/10.1016/B978-1-78548-253-3.50002-0.

[3] *C. Zou, "Analyzing research trends on drug safety using topic modeling*," Expert Opin. Drug Saf., vol. 17, no. 6, pp. 629–636, 2018.

[4] K. B. Putra and R. P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (LDA)," J. Tek. ITS, vol. 6, no. 2, pp. 4–9, 2017.

[5] I.Komputer, D. Ilmu, F. Matematik, P. Alam, and U. G. Mada, "Document Clustering Dengan *Latent Dirichlet Allocation Dan Ward*," vol. V, no. September, 2018.

[6] I. N. Kabiru, P. K. Sari, S. Prodi, and M. Bisnis, "Analisa Konten Media Sosial E-Commerce Pada Instagram Menggunakan Metode Sentimen Analysis Dan Lda-Based Topic Modeling (Studi Kasus : Shopee Indonesia) A*nalysis Of Content Social Media E-Commerce In Instagram Using Sentiment Analysis And Lda Based Topki*," vol. 6, no. 1, pp. 12–19, 2019.

[7] Krasnashchok, K., Jouili, S., 2018. Improving Topic Quality by Promoting Named Entities in Topic Modeling. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 247–253.

[8] Utami, K.,P., 2017. Analisis Topik Data Media Sosial Twitter Menggunakan Model Topik *Latent Dirichlet Allocation*, *Skripsi*, Program Studi S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor.

[9] Bhatia, S., Lau, J.H., Baldwin, T., 2017. An Automatic Approach for Document-level Topic Model Evaluation. Conference on Computational Natural Language Learning 206–215.

[10] Chandrasekar, P., Qian, K., 2016. *The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier, in 2016 IEEE 40th Annual Computer Software and Applications Conference* (COMPSAC). pp. 618–619. https://doi.org/10.1109/COMPSAC.2016.205.

[11] Hong, V.N., Nguyen, H., Hieu, D.N., Snasel, V., 2016. n -Gram-Based Text Compression. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2016/9483646

[12] Sun, S., Dai, Z., Xi, X., Shan, X., Wang, B., 2018. *Ensemble Machine Learning Identification of Power Fault Countermeasure Text Considering Word String TF-IDF Feature,* in 2018 IEEE *International Conference of Safety Produce Informatization* (IICSPI). pp. 610–616. https://doi.org/10.1109/IICPSPI.2018.8690443.

[13] Agustina, A. 2017. Analisis dan visualisasi suara pelanggan pada pusat layanan pelanggan dengan pemodelan topik menggunakan *latent dirichlet allocation* (LDA) studi kasus: PT. Petrokimia Gresik [skripsi]. Surabaya(ID): Institut Teknologi Sepuluh Nopember.

[14] Korencic, D., Ristov, S., Snajder, J., 2018. *Document-based Topic Coherence Measures for News Media Text. Preprint submitted to Expert Systems with Applications 1–44.*