Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 110−115
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1584

Telkom University News Topic Modeling Using Latent Semantic Analysis (LSA) Method on Online News Portal

Ihsan Ahsanu Amala, Donni Richasdy*, Mahendra Dwifebri Purbolaksono

School of Computing, Telkom University, Bandung, Indonesia
Email: [1]ihsanahsanuamala@student.telkomuniversity.ac.id, [2,*]donnir@telkomuniversity.ac.id, [3]mahendradp@telkomuniversity.ac.id
Email Penulis Korespondensi: donnir@telkomuniversity.ac.id

Submitted: **11/05/2022**; Accepted: **23/05/2022**; Published**: 30/06/2022**

Abstract−In this day and age, the development of online news portals regarding news is quite easy to access, online news portals are information that explains an event that has occurred or is happening with electronic media intermediaries, as well as news about Telkom University which is quite easily accessible through online news portals. A system has been designed that is capable of modeling Telkom University news topics. Modeling news topics is very interesting to be used as research material because the process of understanding each individual on the topics contained in the news is different, therefore topic modeling is needed to find out what topics are news about Telkom University. In this study, a Latent Semantic Analysis (LSA) model has been designed to carry out a topic modeling process that aims to make it easier for readers to understand news topics related to Telkom University, Latent Semantic Analysis (LSA) is a mathematical method in finding hidden topics by analyzing the structure semantics of the text. After doing several research scenarios, the best coherence score was 0.524 with a total of six topics.

Keywords: News; LSA; Topic Modeling; Topic Coherence; Telkom University

1. INTRODUCTION

News is a report of an incident, generally news is distributed through many different media, such as word of mouth, print media, electronic media and other media. News provides information that increases the knowledge of readers or news recipients [1]. News is usually published by a mass media agency, mass media can be in the form of electronic media, one of which is electronic media, namely online news portals, online news portals are electronic media containers whose use requires the internet to access the news portal. This research will focus on news related to Telkom University on online news portals. Telkom University is a private university located in Bandung Regency, West Java Province. To make it easier for the Telkom University marketing team to know what topics are news related to Telkom University, making it easier for the Telkom University marketing team in branding Telkom University to the general public, especially prospective new students, a topic modeling will be designed that is able to automatically model the topic. news that are interrelated or have relevant words.

Topic Modeling is a field in machine learning, especially in the field of natural language processing which is unsupervised learning, meaning that the data used does not need to have a label [2][3], the task that will be carried out by topic modeling is scanning a document, detecting word patterns and phrase. Then automatically group word groups and similar topics that best characterize the word set or document. Topic modeling generates automatic theme conclusions from a collection of texts that are formed into a corpus, quite difficult to do manually from a large corpus [4]. The basic task of topic modeling is to group words into a document by identifying the words and patterns that exist in the document so that they characterize a particular topic [5].

To overcome the problem of topic modeling automatically, a topic modeling will be built using the Latent Semantic Analysis (LSA) method, there are many methods that can be used to model topics such as LDA, NMF and LSA [6]. LSA is one of the best models in the unsupervised learning paradigm with a combination of statistical and algebraic methods. This method reveals the latent structure of words, sentences and texts through the Singular Value Decomposition (SVD) algorithm [7]. LSA utilizes SVD to reduce the dimensions of the term matrix, this reduction is carried out to reduce the sparse matrix, noisy matrix and also redundant matrices in many dimensions, dimension reduction can be done using truncated SVD (Singular Value Decomposition). SVD captures the semantic structure that underlies a collection of document representation matrices [8].

As for several previous studies that have been carried out related to topic modeling, research conducted [8] Analyzing latent semantics based on single value decomposition using data in the form of topics and topic descriptions obtained from various different articles within the scope of NLP, and getting the results that the Algorithm Latent Semantic with optimal value using single value decomposition gives query results that are semantically correlated in the corpus. Research [9] conducted a comprehensive comparative study of LSA with the TF-IDF approach which is commonly used for text classification and proved that LSA produces better accuracy in classifying texts. This paper proposes a method of using the concept of entropy, which will further increase the accuracy of text classification with a dataset consisting of 870 news articles belonging to 10 different news categories with the results obtained, namely the accuracy obtained when using LSA without TF-IDF is lower than with LSA using TF-IDF which added the concept of entropy.

Research [4] using LDA and NMF methods to find topics that are widely discussed in the news of two consecutive elections in South Africa by obtaining results provides important insights related to two-term elections. Research [10] using the LDA method with the objective of knowing the modeling of Indonesian news topics with measurements based on manual interpretation of the relevance between components in one topic, the results obtained are several keywords for pre-match analysis topics. The measurement generally used in topic modeling is topic coherence, topic coherence is a method of evaluating the number of relevant topics and has the best coherence, such as research [11] which obtained a coherence score of 0.5 on average with the method used, namely LDA and The dataset used is abstract and fulltext from a research article.

This study focuses on modeling Indonesian news texts related to Telkom university, with the aim of implementing LSA and obtaining LSA evaluations based on the most optimal number of topics by using topic coherence measurements based on several test scenarios that have been carried out, so that the results are in the form of modeling. news topics related to Telkom university in order to find out topics that are news on online news portals.

2. RESEARCH METHODOLOGY

2.1 System Design

The system to be built is a news topic modeling system related to Telkom University using the Latent Semantic Analysis (LSA) model. The model development process begins with collecting datasets on online news portals, then a text preprocessing process is carried out which is useful for making the data structured and ready to be processed. by the model, after text preprocessing is done, the next process is making the LSA model, and the last process is evaluation using topic coherence to get the best number of topics and evaluation of the LSA model that has been built. Figure 1 shows an overview of the research flow.

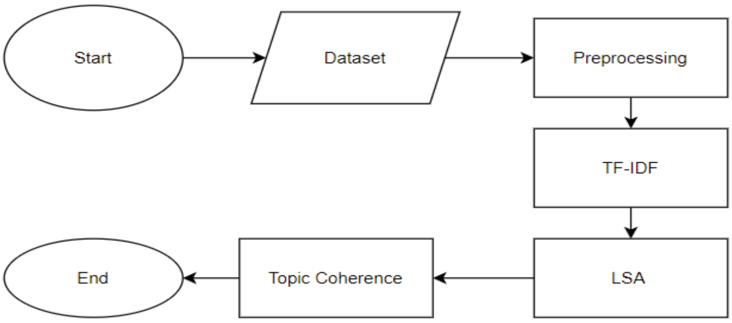

Figure 1. Research Flow

2.2 Dataset

The dataset used in the research is a title news dataset sourced from popular news portals [12], with news ranges between september 2019 to june 2021, the dataset acquisition process starts from scraping data on news portals using Scrapy tools which are open source. The total dataset that has been collected is 408 different news titles related to Telkom University. To find out whether the news dataset that has been scraped is related to Telkom University news, a WordCloud process is carried out to find out what words appear most often, to reflect the news dataset as a whole. With the results of the WordCloud shown in Figure 2.

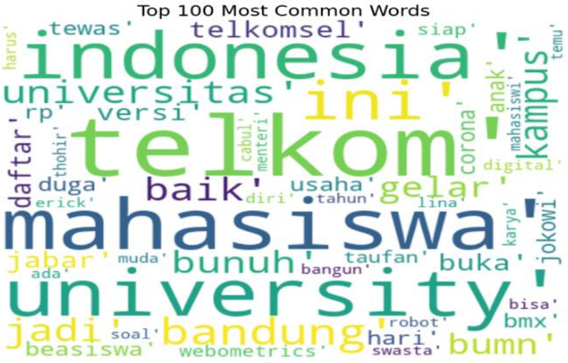

Figure 2. Words that Often Appear in The News Dataset

2.3 Preprocessing

Text preprocessing is the stage for making unstructured text data into structured text so that it gives better results when the dataset is processed by the model. There are several text preprocessing processes:

a. Lowercasing
   In this process all text that will be used in model processing is converted into lowercase letters, this aims to provide consistency to the words

b. Noise Removal
Noise Removal is the stage of removing characters in the dataset that can interfere with data processing in the model later, these characters can be in the form of punctuation marks, hashtags, URLs and similar characters.

c. Stopwords Removal
Stopwords removal is the stage of removing words that have no meaning, these words often appear in documents but do not have meaningful information for the model.

d. Stemming
Stemming is a process of returning formed words to their basic words, stemming is done by removing affixes to words, stemming is a process that is included in the normalization stage of the text so that the text is more consistent.

e. Tokenizing
Tokenizing is the process of parsing text into the smallest units in sentences known as tokens, tokenization aims to extract meaning from a text.

Table 1 shows the preprocessing process carried out along with the results before and after the preprocessing process

Table 1. After and Before Preprocessing Example

Process	Before	After
Lowercasing	Telkom University	telkom university
Noise Removal	siapa universitas terbaik?	siapa universitas terbaik
Stopwords Removal	pergi ke pasar	pergi pasar
Stemming	menahan	tahan
Tokenizing	rita suka boneka	[‘rita’, ‘suka’, ‘boneka’]

2.4 TF-IDF (Term Frequency-Inverse Document Frequency)

Term Frequency – Inverse Document Frequency (TF-IDF) is a combination of two different words, namely Term Frequency and Inverse Document Frequency, TF is used to measure how often a terminology appears in a document, IDF is used to give small weight to words that have large frequency and a large weight for words that have a small frequency [13]. The basic purpose of using TF-IDF is to measure how relevant a word is in a sentence and to weight the word. The use of TF-IDF in the LSA model resulted in higher accuracy compared to the use of LSA without TF-IDF[4]. With the following equation (1):

$$w_{ij} = tf_{(ij)} * idf_j, where\ Idf_j = log\frac{N}{df(t)} \tag{1}$$

Where w_{ij} is the weighting of term t_j for d_j documents, tf_{ij} is the weight of how often a term t_j is in a document, N is the number of corpus, and $df(t)$ is the occurrence of words in N documents.

2.4 LSA (Latent Semantic Analysis)

LSA utilizes SVD (Singular Value Decomposition) to reduce the dimensions of the term matrix, this reduction is carried out to reduce sparse matrices, noisy matrices and also redundant matrices in many dimensions, dimension reduction can be done using truncated SVD (Singular Value decomposition). SVD captures the semantic structure that underlies a collection of document representation matrices [8]. Singular Value Decomposition (SVD) is a method that determines the latent structure of words, sentences or text based on a combination of statistical and algebraic methods[14].The basic idea of SVD is to find the most valuable information and use lower dimensions to represent that valuable information. With the SVD equation as follows(2):

$$A = USV^T \tag{2}$$

$$U \in R^{(m\,x\,k)}, V \in R^{(n\,x\,k)} \tag{3}$$

Where A is the original matrix of size mxn, m is the Terms, n is the document, k is the number of topics, U is a matrix that represents the document vector found in documents with size m x k, V is a matrix that represents the vector found topics with size n x k, and S is a matrix that represents the diagonal of the matrix with size k x k

2.5 Topic Coherence

Topic coherence is a method for evaluating the optimal number of topics in topic modeling. Measurement through topic coherence is very good in comparing the number of topics that are in accordance with human-interpretability [15]. Topic coherence provides an overview of the optimal value of the number of topics by explaining the interpretation of the relationship between the number of certain topics known as the coherence score. In this study using the Cv method, this method uses coherence calculations by taking the number of co-occurrence for a given word using a sliding window [15], co-occurance calculation is obtained by normalized pointwise mutual information (NPMI) of each top word. With the following calculations (3):

$$NPMI(wi, wj) = \sum_{j}^{N-1} \frac{log\frac{P(wi,wj)}{P(wi),P(wj)}}{-logP(wi,wj)} \qquad (4)$$

Where P(wi) is the probability of the random occurrence of wi in the document, P (wi, wj) is the probability of the two words wi and wj appearing in the document randomly. N is the highest possible choice of the words w1, w2, …, wn [11].

3. RESULT AND DISCUSSION

The research has been completed by carrying out four research scenarios to get the best LSA model results in modeling Telkom University news topics, at this stage experiments will be carried out on the stemming and stopwords process, where the dataset has been preprocessed outside of stemming and stopwords which will be the experimental process.

The scenario evaluation process uses topic coherence as a measure of which scenario is the best, by comparing the coherence score of each scenario on each topic, the best research scenario will be found, the coherence score taken is up to three digits behind the comma Number of topics with the best scenario will be used to model the topic in determining what topics will be reported by Telkom University

3.1 Result

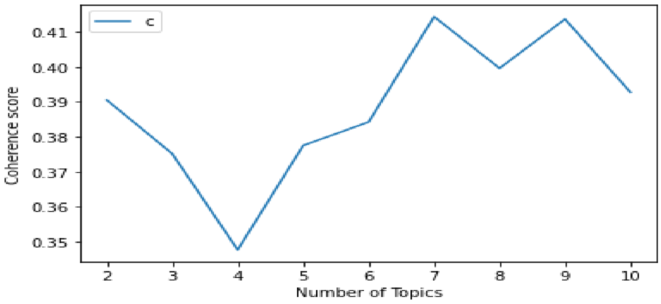

Figure 3. Coherence Score Without Stemming and Stopwords

Figure 3. Shows the first research scenario is the implementation of the LSA model with news title text data without stemming or stopwords removal processes, the number of topics with the best coherence value is seven with a coherence score of 0.414

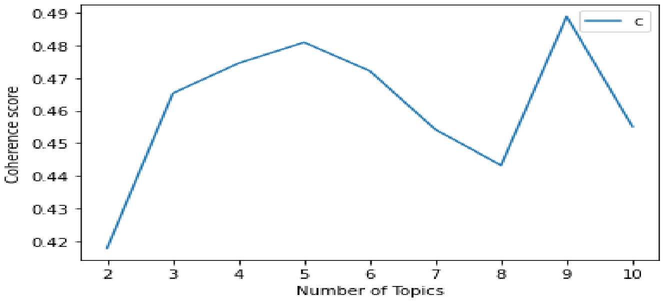

Figure 4. Coherence Score Without Stemming but Implementing Stopwords

Figure 4. Shows the second research scenario is the implementation of the LSA model with news title text data, no stemming process is carried out but a stop words removal process is carried out, the number of topics with the best coherence value is nine with a coherence value of 0.488

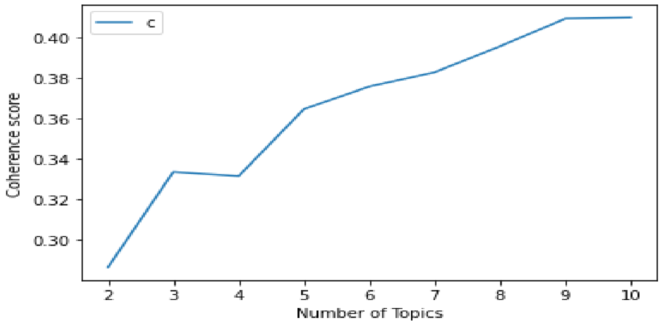

Figure 5. Coherence Score with Stemming but Without Stopwords Removal

Figure 5. Shows the third research scenario is the implementation of the LSA model with the text of the news title data, a stemming process is carried out but no stopwords removal process is carried out, the number of topics with the best coherence score is nine and ten with a coherence score of 0.409.

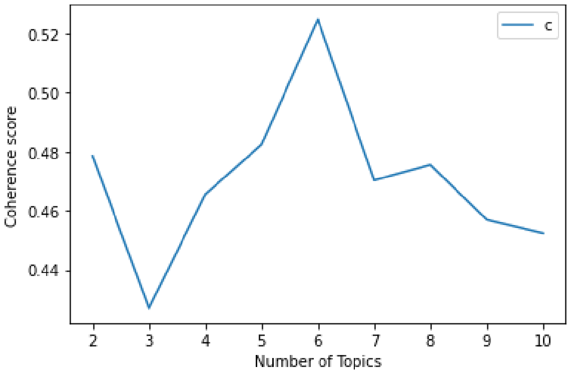

Figure 6. Coherence Score with Stemming and Stopwords

Figure 6. shows the last research scenario, namely the implementation of the LSA model with news title text data, was carried out by a stemming process and a stopwords removal process, the best number of topics was six with a coherence score of 0.524.

3.2 Discussion

To see more clearly the comparison of the best scenario test results from the four tested scenarios, it can be seen in the following table

Table 2. Comparison of the coherence score topic coherence in the test result scenario

Number of Topics	No Stemming and Stopwords Removal	No Stemming but Implement Stop Words Removal	Stemming but Not Impelement StopWords Removal	Stemming and Impelement StopWords Removal
2	0.390	0.417	0.286	0.478
3	0.375	0.465	0.333	0.426
4	0.347	0.474	0.331	0.465
5	0.377	0.480	0.364	0.482
6	0.384	0.472	0.375	**0.524**
7	0.414	0.454	0.382	0.470
8	0.399	0.443	0.395	0.475
9	0.413	0.488	0.409	0.456
10	0.392	0.455	0.409	0.452

Obtained based on the experimental results of the best scenario test which can be seen in Table 2, namely by applying the stemming process and implementing the stopwords crime process with a coherence score of 0.524, the interesting thing here is when the stemming process is carried out but the stopwords destruction process is not implemented by the LSA model in getting the score. coherence, this makes stopwords as a test variable that plays an important role in increasing the coherence score in each scenario. stemming combined with stopwords cream gives the best results because the process of words that have affixes will be returned to the form to form dreams and words that have no meaning will eliminate the process of text data used to model topics, so that the correlation between topics becomes better and increase the coherence score.

After obtaining the best scenario from the experiments carried out, namely the scenario of implementing the stemming process and implementing the stopwords removal process with a total of six topics, then we will use this scenario in implementing the LSA model to see what topics are news related to Telkom University.

Table 3. Topics and the top 10 keywords for the topic

Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*	*Topic 6*
telkom	indonesia	Indonesia	bumn	bandung	mahasiswi
mahasiswa	universitas	telkomsel	beasiswa	Park	telkom
university	versi	gelar	mahasiswi	versi	bandung
universitas	webometrics	rider	buka	podomoro	senior
indonesia	daftar	jalan	telkom	mahasiswa	webometrics
versi	swasta	luncur	telkomsel	gelar	kasus
kampus	kampus	teknologi	kampus	webometric	korban

daftar	turut	bandung	bandung	rider	versi
mahasiswi	qs	digital	erick	ekstrem	park
webometrics	wur	aksi	daftar	tinggal	ekstrem

From Table 3, information can be obtained that topic 1 is news that discusses the ranking of Telkom universities through the **webometric** version, topic 2 is news that discusses Telkom university rankings through the **qs wur** version, topic 3 is news that discusses Telkom universities based on technology campuses, it is marked by the keywords **technology** and **digital**, topic 4 is news that discusses scholarships at Telkom universities and Telkom universities which still have links to BUMN which are marked by the keywords **bumn** and **scholarships**, topic 5 is news that discusses the life around Telkom university is marked with the keywords **bandung** and **stay**, and topic 6 is news that discusses social events related to Telkom university marked with the keywords **senior, victim and case**.

4. CONCLUSION

Based on the results and discussion, research has been carried out on modeling Telkom university news topics on online news portals using the LSA method, two conclusions are obtained, namely, the first use of stopwords and stemming has an influence on the evaluation results using the topic coherence score, this is obtained from the best scenario in modeling topics using the LSA method for data in the form of news headlines related to Telkom university, namely the text that is carried out by the application of the stop words removal process and stemming process with the optimal number of topics, namely six with a coherence score of 0.524, this is because stemming eliminates words affixes and stopwords removal that remove words that appear frequently but have no meaning. The second conclusion is that there are six topics that are reported on online news portals about Telkom university, where topic one and topic two are both themed Telkom university rankings, topic three is themed Telkom university technology-based campus, the fourth topic is Telkom university which has links to SOEs and information on scholarships, topics five and six have almost the same theme, namely about life around and social events that occur in the Telkom university environment.

REFERENCES

[1] M. Tanikawa, “What Is News? What Is the Newspaper? The Physical, Functional, and Stylistic Transformation of Print Newspapers, 1988-2013 MIKI TANIKAWA,” 2017. [Online]. Available: http://ijoc.org.

[2] G. Costa and R. Ortale, “Document clustering and topic modeling: A unified bayesian probabilistic perspective,” in Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, Nov. 2019, vol. 2019-November, pp. 278–285. doi: 10.1109/ICTAI.2019.00047.

[3] T. Iwata, T. Hirao, and N. Ueda, “Topic Models for Unsupervised Cluster Matching,” IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 4, pp. 786–795, Apr. 2018, doi: 10.1109/TKDE.2017.2778720.

[4] A. Moodley and V. Marivate, “Topic modelling of news articles for two consecutive elections in South Africa,” in 2019 6th International Conference on Soft Computing and Machine Intelligence, ISCMI 2019, Nov. 2019, pp. 131–136. doi: 10.1109/ISCMI47871.2019.9004342.

[5] Y. Kalepalli, T. Shaik, D. Pasupuleti, and S. Manne, “Effective Comparison of LDA with for Topic Modelling,” International Confrence on Intelligent Computing Control System (ICICCS)), pp. 1245–1250, 2020.

[6] K. Rajendra Prasad, M. Mohammed, and R. M. Noorullah, “Visual topic models for healthcare data clustering,” Evolutionary Intelligence, vol. 14, no. 2, pp. 545–562, Jun. 2021, doi: 10.1007/s12065-019-00300-y.

[7] D. Sarkar, Text Analytics with Python. Apress, 2016. doi: 10.1007/978-1-4842-2388-8.

[8] P. Kherwa and P. Bansal, “Latent Semantic Analysis: An Approach to Undestand Semantic of Text,” International Conference on Current Trends in Computer, Electrical, Electronics and Communication, pp. 870–874, 2017.

[9] P. P. G. Neogi, A. K. Das, S. Goswami, and J. Mustafi, “Topic Modeling for Text Classification,” in Advances in Intelligent Systems and Computing, 2020, vol. 937, pp. 395–407. doi: 10.1007/978-981-13-7403-6_36.

[10] H. A. Fathan, P. E. Cergas, W. Kurniawan, G. Akbar, and P. Ridwan, “Twitter Topic Modeling on Football News,” International Conference on Computer and Communication Systems, pp. 467–471, 2018.

[11] S. Syed and M. Spruit, “Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation,” in Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017, Jul. 2017, vol. 2018-January, pp. 165–174. doi: 10.1109/DSAA.2017.61.

[12] Shelly Maysar, “13 Portal Berita Online Terbaik di Indonesia,” Akudigital.com, Dec. 04, 2021.

[13] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[14] K. Al-Sabahi CVTE, K. Al-Sabahi, Z. Zuping, and Y. Kang, “Latent Semantic Analysis Approach for Document Summarization Based on Word Embeddings,” 2018. [Online]. Available: https://www.researchgate.net/publication/326290389

[15] F. Yi, B. Jiang, and J. Wu, “Topic Modeling for Short Texts via Word Embedding and Document Correlation,” IEEE Access, vol. 8, pp. 30692–30705, 2020, doi: 10.1109/ACCESS.2020.2973207.