Improving Topic Modeling Performance with Term Frequency for Feature Optimization on Final Project Dataset

Putu Manik Prihatini[1], I Ketut Gede Sudiartha[2],
Sri Andriati Asri[3], Elina Rudiastari[4],
I Nyoman Eddy Indrayana[5], and Putu Indah Ciptayani[6]

[1,2,3,4,5,6] Information Technology Department, Politeknik Negeri Bali, Bali, Indonesia
manikprihatini@pnb.ac.id

Abstract. Unstructured data is generated in huge volumes every day from sources such as emails, social media, and business documents. This data, although extremely useful for decision-making, requires extensive preprocessing and feature extraction to be useful. Text mining extracts meaningful patterns from large text datasets, revealing new insights through feature extraction, which identifies distinctive data attributes. Topic modeling, especially using Latent Dirichlet Allocation, plays an important role in this process by uncovering semantic structures hidden within large text collections. Latent Dirichlet Allocation operates with the Bag of Words approach but can be improved by integrating Term Frequency-Inverse Document Frequency, which filters out less important words and improves topic accuracy and processing speed. This research compares the performance of Bag of Words and Term Frequency-Inverse Document Frequency methods on Latent Dirichlet Allocation to extract topics from Indonesian final project abstracts. Through testing with the coherence score metric, it is shown that Latent Dirichlet Allocation combined with Term Frequency-Inverse Document Frequency can be used to extract the hidden topics better than Bag of Words.

Keywords: Bag of Words, Final Project, Latent Dirichlet Allocation, Term Frequency-Inverse Document Frequency

1 Introduction

Today, huge volumes of data are generated daily from various sources and are driven by technological trends such as the Internet of Things, the development of Cloud Computing, and the widespread use of smart devices (Oussous et al., 2018). Most of the data circulating today is unstructured, such as email messages, customer reviews, social media posts, news articles, and business documents. This unstructured data holds valuable information that can be used to support decision-making. However, unstructured data usually has a non-standardized format, requiring extensive preprocessing and feature extraction to be used (Sedlakova et al., 2023). The

A. A. N. G. Sapteka et al. (eds.), *Proceedings of the International Conference on Sustainable Green Tourism Applied Science - Engineering Applied Science 2024 (ICoSTAS-EAS 2024)*, Advances in Engineering Research 249,
https://doi.org/10.2991/978-94-6463-587-4_7

development of text mining and natural language processing (NLP) has spurred the development and advancement of algorithms capable of handling unstructured data (Dwivedi et al., 2023).

Text mining is the process of extracting interesting and meaningful patterns from large volumes of text documents (Talib et al., 2016). Text mining can discover new information that is not immediately apparent from a collection of documents (Vidhya, 2021). This new information is discovered by extracting features from the document collection. Feature Extraction involves the extraction of pertinent details from raw data to identify and highlight the most distinctive attributes within a dataset, such as images, text, or voice, thereby providing a comprehensive representation and description of the data (Salau & Jain, 2019). Text feature extraction plays a crucial role in text mining by enabling the identification and extraction of relevant information and patterns from textual data, facilitating tasks such as sentiment analysis, topic modeling, and information retrieval.

Topic modeling is a revolutionary technique in text mining, used to uncover the hidden semantic structure within large collections of documents (Kherwa & Bansal, 2018). Latent Dirichlet Allocation (LDA), introduced by Blei, is a topic modeling technique that operates on words, documents, and the entire corpus, allowing it to capture dynamic changes in it (Blei et al., 2003). Various reasoning methods have been applied to LDA in different studies, but Gibbs sampling is a widely used approach. Incorporating fuzzy logic concepts into the Gibbs Sampling LDA inference mechanism during the Indonesian text feature extraction phase resulted in faster convergence and improved performance compared to using Gibbs Sampling (Prihatini et al., 2017). Three new topic models based on Gibbs Sampling LDA called pack are developed for sentiment analysis at the packet level, where these new models have better performance than the basic models because the additional parameters can precisely influence the process of generating words in reviews (Osmani et al., 2020). PAN-LDA, an LDA-based topic model incorporating COVID-19 case data and news articles, uses collapsed Gibbs sampling for parameter inference and enhances machine learning algorithms by generating more identifiable topics and improving time series data forecasting (Gupta & Katarya, 2021). Indonesian Sentiment Lexicon was added to LDA algorithm used for sentiment analysis and topic modeling on Twitter crawling data (Dikiyanti et al., 2021). LDA with Collapsed Gibbs sampling was used to identify distinct content clusters in war-related news (Khairova et al., 2024).

LDA is a generative probabilistic model for a corpus that generates documents without considering word order, relying solely on the Bag of Words (BoW) approach (Kherwa & Bansal, 2018). In addition to BoW, the use of the Term Frequency-Inverse Document Frequency (TF-IDF) method has also been used in several studies. The integration of the TF-IDF algorithm under Spark and CountVectorizer with the LDA method for clustering news topics has shown that the processing speed of LDA topic model clustering is enhanced for large data samples based on Spark (Zhou et al., 2020). TF-IDF is integrated with the LDA method to filter out less important words, allowing LDA to generate more accurate topics (Nugroho et al., 2022). Count Vectorizer and TF-IDF are incorporated into the LDA feature extraction method to uncover topics from COVID-19 tweets (Sofi & Selamat, 2023). In addition, comparing the performance of

TF-IDF and LDA has also been done in several studies (Prihatini et al., 2018; Rani & Bidhan, 2021).

From the studies mentioned above, so far there has been no research that compares how the BoW and TF-IDF methods affect the performance of LDA topic model, especially on Indonesian documents. Therefore, in this research, the LDA topic model is used to extract topics for Indonesian final project abstracts, where the performance of the BoW and TF-IDF methods on the LDA topic model was compared using coherence metrics. Through the results of this comparison, it is expected to be known which method provides better performance for the LDA topic model to extract Indonesian language documents.

2 Methodology

The methodology used in this study follows the flowchart shown in Figure 1.

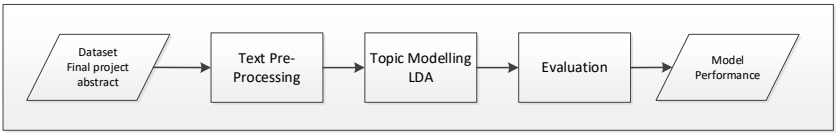

Figure 1. Research flowchart

2.1 Dataset Final Project Abstract

The final project text was obtained by downloading the student final project report file on the repository system. The downloaded files were stored in pdf file format. The amount of final project data stored in the repository is 76 data. Only the abstract part of the file was used in the dataset and copied into a text file.

2.2 Text-Preprocessing

The tokenization stage is a process of breaking down the document text into the smallest forms such as paragraphs, sentences, or words. The case folding stage is a process of converting all letters in the tokenized document into lowercase letters. In addition, the case folding process also removes numbers, punctuation marks, and spaces at the beginning and end of sentences, and removes tokens with a length of less than two letters. The filtering stage is a process to eliminate words that have no meaning to the content of the text such as prepositions, conjunctions, and the like. The stemming stage is a process to obtain the root word of each word in the document text, by removing the affixes on the word.

2.3 Topic Modelling LDA

The initialization stage involves assigning initial values and topics to text units. This process begins by calculating a random value for each word and assigning a topic to

each word using a multinomial distribution based on this random value, as shown in Equation (1). *K* represents the total number of topics, *n* denotes the total number of words, and *p* is the probability (Heinrich, 2005).

$$p(\vec{n}|\vec{p}) = \prod_{k=1}^{K} p_k^{n^{(k)}} = Mult(\vec{n}|\vec{p}, 1) \tag{1}$$

The topic sampling stage involves defining a new topic for each word in each text file. This process starts by decrementing the count in the word-topic and document-topic matrices. Then, the probability value for each word is calculated, as shown in Equation (2), where *nkt* is the number of times word *n* is assigned to topic *k*, β is a constant parameter for the topic, *nkm* is the number of times topic *k* appears in the document, α is a constant parameter for the word, *W* is the total number of words in the corpus, *V* is the total number of unique words in the corpus, and *K* is the total number of topics (Heinrich, 2005). The new topic for each word is then determined using a multinomial distribution based on the calculated probability value. After that, the counts in the word-topic and document-topic matrices are incremented according to the new topic assignment. This entire process is repeated until convergence is reached, as described in Equation (3), where *N* is the total number of unique words in the corpus and *zi* is the probability value for each word *n* (Prihatini et al., 2017).

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + W\,\beta} \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K} n_{m,\neg i}^{k} + K\,\alpha} \tag{2}$$

$$\left\| \sum_{n=1}^{N} z_i - \sum_{n=1}^{N} z_{i-1} \right\| \leq \frac{\alpha}{\beta} \tag{3}$$

In this research, the corpus weight is obtained using two methods, namely BoW and TF-IDF. In the BoW approach, each distinct word is treated as a separate dimension (or axis) in the vector space. For a text set containing *n* unique words, the resulting vector space has *n* dimensions, with each dimension representing a unique word from the text set. Each text document is then represented as a point in this vector space. The position of a point along a specific dimension is determined by the frequency of that dimension's word in the corresponding document. TF-IDF is formulated to emphasize words that appear frequently in a specific document while being relatively uncommon across the entire corpus, thereby aiding in identifying terms that are particularly important for that document. For the TF-IDF method, it is done using Equation (4). *TermFreq(w,d)* refers to the frequency of occurrence of word *w* in document *d*, *N* refers to the number of documents *d*, and *DocFreq(w)* refers to the number of documents *d* containing word *w*.

$$TF - IDF_{(w,d)} = TermFreq_{(w,d)} * \log \frac{N}{DocFreq(w)} \tag{4}$$

2.4 Evaluation

Coherence measures have been introduced by the NLP community to assess the quality of topics generated by various topic models. The coherence measure relies on the co-occurrence of word pairs. For a given ordered list of words $T = \{w_1, w_2, \ldots, w_n\}$, the *UMass* coherence is defined as shown in Equation (5) (Rosner et al., 2014). A Boolean document model is used to estimate word probabilities p, such that $p(w_m, w_l)$ represents the ratio of the number of documents containing both words w_m and w_l to the total number of documents in the corpus D. To prevent the calculation of the logarithm of zero, a smoothing count of $1/D$ is added.

$$C_{UMass}(T) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{p(wm, wl) + \frac{1}{D}}{p(wl)} \tag{5}$$

3 Result and Discussion

3.1 Result

During the text pre-processing, the dataset was generated with the number of terms as presented in Table 1. The terms produced during the text processing were extracted using the LDA method to derive the feature value for each term. In this research, the number of LDA topics was tested starting from 1 to 76. The coherence value of the LDA model generation process with the BoW method can be seen in Table 2 and visualized in Figure 2. The coherence value of the LDA model generation process with the TF-IDF method can be seen in Table 3 and visualized in Figure 3.

Table 1. The results of text preprocessing

Stage	Number of terms
Tokenization	11,290
Case Folding	9,788
Filtering and Stemming	6,502

Table 2. Summary of coherence value for LDA-BoW

Number of topics	Coherence value	Number of topics	Coherence value
10	0.3662	46	0.4077
16	0.3592	50	0.4142
20	0.3945	56	0.4171
26	0.3901	60	0.4110
30	0.3888	**66**	**0.4351**
36	0.4197	70	0.4256
40	0.3835	76	0.4298

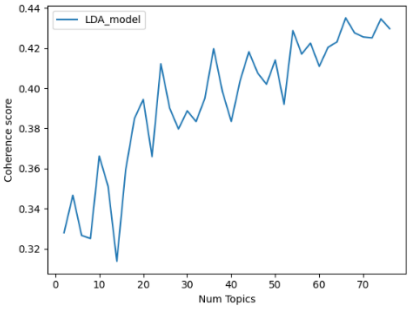

Figure 2. Coherence score for LDA-BoW

Table 3. Summary of coherence value for LDA-TFIDF

Number of topics	Coherence value	Number of topics	Coherence value
10	0.3356	46	0.4107
16	0.3711	50	0.4155
20	0.3955	56	0.4298
26	0.3759	60	0.4450
30	0.3635	66	0.4363
36	0.3752	70	0.4555
40	0.3881	**76**	**0.4610**

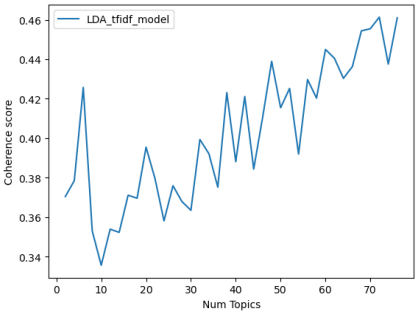

Figure 3. Coherence score for LDA-TFIDF

The final results of the feature values for the first three topics obtained from the LDA model with the BoW method can be seen in Table 4, and with the TF-IDF method can be seen in Table 5.

Table 4. Top three topics of LDA-BoW

Topic	Feature
0	0.061*"pasien" + 0.061*"kendala" + 0.061*"sehat" + 0.061*"salah" + 0.043*"uji" + 0.038*"antarmuka" + 0.037*"jalan" + 0.031*"ambil" + 0.031*"online" + 0.031*"basis"
1	0.147*"langgan" + 0.117*"pesan" + 0.092*"catat" + 0.069*"manual" + 0.060*"sesuai" + 0.057*"hasil" + 0.051*"nota" + 0.051*"rekap" + 0.050*"modeling" + 0.034*"kece"
2	0.005*"muat" + 0.005*"berkas" + 0.005*"denpasar" + 0.005*"beli" + 0.005*"alamat" + 0.005*"jasa" + 0.005*"fitur" + 0.005*"dinas" + 0.005*"dokumen" + 0.005*"jual"

Table 5. Top three topics of LDA-TFIDF

Topic	Feature
0	0.248*"ajar" + 0.073*"orang" + 0.059*"guru" + 0.052*"dasar" + 0.032*"dart" + 0.032*"flutter" + 0.029*"hubung" + 0.027*"mobile" + 0.004*"harap" + 0.002*"hasil"
1	0.005*"muat" + 0.005*"berkas" + 0.005*"denpasar" + 0.005*"beli" + 0.005*"alamat" + 0.005*"jasa" + 0.005*"fitur" + 0.005*"dinas" + 0.005*"dokumen" + 0.005*"jual"
2	0.084*"siswa" + 0.081*"uji" + 0.066*"nilai" + 0.058*"camat" + 0.056*"kantor" + 0.048*"laksana" + 0.045*"efektif" + 0.036*"pegawai" + 0.031*"liput" + 0.030*"database"

3.2 Discussion

The data in Table 1 shows that the original dataset had been parsed into 11,290 terms at the tokenization stage. These tokenized terms were processed at the case folding stage which produced 9,788 terms. In the case folding process, 1,502 terms (± 14% of the total number of tokenized terms) were removed because they contained numbers, punctuation, spaces at the beginning and end of sentences, and the term length was less than two letters. The case folding terms were processed again at the filtering stage which produced 6,502 terms. In the filtering process, 3,286 terms (± 34% of the total number of case folding terms) were removed because they were included in the stop words list. The filtered terms were processed again at the steaming stage. The number of terms produced by the stemming process is the same as the filtering results because there is no term deletion in the stemming process, only the process of removing affixes attached to the terms. From the results of text processing, it can be concluded that, the tokenization of 76 abstract files of Indonesian final project into 11,290 words, at the end of text processing only left 6,502 words, of which 4,788 words were discarded at this stage. In other words, about 42% of the words in the abstract file are meaningless words, and only about 58% of the words are declared meaningful for use in the feature extraction stage. Word selection at this stage plays an important role in the feature

extraction process because it is one of the crucial factors for the success of the LDA topic model to determine hidden topics in the abstract file dataset.

Table 2 indicates that the highest coherence value for LDA results with the BoW method was achieved at the number of topics = 66 with a value of 0.4351. Meanwhile, Table 3 indicates that the highest coherence value for LDA results with the TF-IDF method was achieved at the number of topics = 76 with a value of 0.4610. The coherence metric measurement for LDA with TF-IDF method showed higher results than the coherence value obtained from LDA with BoW method. In addition, the number of topics with the highest coherence value obtained by the TF-IDF method is similar to the number of abstract files used as datasets. This shows that the LDA topic model with TF-IDF method is able to generate topics hidden in the dataset with better quality than the LDA topic model with BoW method. However, the difference in values produced by the two methods is not much different. This could be due to the limited number of datasets used, as well as the limited number of meaningful word features generated at the text processing stage. This condition is in accordance with the results of text processing in the previous stage which showed that only half of the terms in the original dataset were declared meaningful for use in the feature extraction stage so that it was less able to represent the overall meaning contained in the original dataset. Therefore, in order to increase the coherence value, in addition to adding more datasets, text processing algorithms must also be improved in order to produce higher quality features for the feature extraction stage.

The final results of the feature values for the first three topics obtained through the LDA topic model with the BoW method in Table 4 and the LDA topic model with the TF-IDF method in Table 5, were selected to represent all the successfully generated topics, where the results showed that the feature values on all generated topics were successfully extracted by the LDA topic model. These results show that the use of BoW and TF-IDF methods can provide good performance for the LDA topic model in extracting features from the abstract file dataset. This is in accordance with the coherence value produced by the two LDA topic models, which shows the quality of the LDA topic model built is quite good. The resulting feature value is an important component for the next stage of the text mining process. This result can be a reference that both BoW and TF-IDF methods can be applied to the LDA topic model and have good performance on Indonesian documents. The TF-IDF method has better performance than the BoW method but requires performance improvement in order to provide more optimal results.

4 Conclusion

LDA is used to extract document topics from Indonesian final project abstracts. LDA performance is compared using BOW and TF-IDF methods. This research involved extensive text preprocessing, so only about 58% of the words were found to be meaningful for use in the feature extraction stage. LDA is applied to extract feature values and find hidden topics, with the coherence score metric used to evaluate the results. This research shows that LDA, when combined with TF-IDF, has a coherence

value of 0.4610, higher than using BOW. The visualization and analysis results show that LDA feature extraction and topic modeling perform well to reveal meaningful patterns from unstructured Indonesian document text data.

Acknowledgment

This paper was supported by grants of DIPA Politeknik Negeri Bali based on Contract of Research No. 01907/PL8/AL.04/2024.

References

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent dirichlet allocation Michael I. Jordan. *Journal of Machine Learning Research, 3*.

Dikiyanti, T. D., Rukmi, A. M., & Irawan, M. I. (2021). Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm. *Journal of Physics: Conference Series*, *1821*(1), 12054. https://doi.org/10.1088/1742-6596/1821/1/012054.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Gupta, A., & Katarya, R. (2021). PAN-LDA: A latent dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. *Computers in Biology and Medicine*, *138*, 104920. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104920.

Heinrich, G. (2005). *Parameter Estimation for Text Analysis*.

Khairova, N., Holyk, Y., Sytnikov, D., Mishcheriakov, Y., & Shanidze, N. (2024). Topic modelling of ukraine war-related news using Latent Dirichlet Allocation with collapsed Gibbs sampling. *ISW-CoLInS 2024. Intelligent Systems Workshop at CoLInS 2024 : Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume III: Intelligent Systems Workshop*, *3688*, 1–15. https://ceurws.org/Vol-3688/.

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A comprehensive review. *ICST Transactions on Scalable Information Systems*, *7*, 159623. https://doi.org/10.4108/eai.13-7-2018.159623.

Sofi, S. M., & Selamat, A. (2023). Aspect based sentiment analysis: Feature extraction using Latent Dirichlet Allocation (LDA) and Term Frequency - Inverse Document Frequency (TF-IDF) in Machine Learning (ML). *Malaysian Journal of Information and Communication Technology (MyJICT)*, 169–179. https://doi.org/10.53840/myjict8-2-102.

Nugroho, S. A., Bachtiar, F. A., & Wihandika, R. C. (2022). Aspect extraction in e-commerce using Latent Dirichlet Allocation (LDA) with term Frequency-Inverse Document

Frequency (TF-IDF). *Jurnal Ilmiah Kursor*, *11*(2), 53. https://doi.org/10.21107/kursor.v11i2.247.

Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched Latent Dirichlet Allocation for Sentiment Analysis. *Expert Systems*, *37*(4). https://doi.org/10.1111/exsy.12527.

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/https://doi.org/10.1016/j.jksuci.2017.06.001.

Prihatini, P. M., Putra, I. K. G. D., Giriantari, I. A. D., & Sudarma, M. (2017). Fuzzy-Gibbs latent Dirichlet Allocation model for feature extraction on Indonesian documents. *Contemporary Engineering Sciences*, *10*, 403–421. https://doi.org/10.12988/ces.2017.7325.

Prihatini, P. M., Suryawan, I. K., & Mandia, I. N. (2018). Feature extraction for document text using Latent Dirichlet Allocation. *Journal of Physics: Conference Series*, *953*(1). https://doi.org/10.1088/1742-6596/953/1/012047.

Rani, U., & Bidhan, K. (2021). Comparative assessment of extractive summarization: TextRank, TF-IDF and LDA. *Journal of Scientific Research*, *65*(01), 304–311. https://doi.org/10.37398/JSR.2021.650140.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *ArXiv Preprint ArXiv:1403.6397*.

Salau, A., & Jain, S. (2019). *Feature extraction: A Survey of the Types, Techniques, Applications*. 158–164. https://doi.org/10.1109/ICSC45622.2019.8938371.

Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Alois Ettlin, D., Grübner, O., Rinaldi, F., & von Wyl, V. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, *2*(10), e0000347. https://doi.org/10.1371/journal.pdig.0000347.

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*. https://doi.org/10.14569/IJACSA.2016.071153.

Vidhya, K. A. (2021). Text mining process, techniques and tools : an overview. *International Journal of Information Technology and Management*.

Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., & N. Xiong, N. (2020). News text topic clustering optimized method based on TF-IDF Algorithm on Spark. *Computers, Materials & Continua*, *62*(1), 217–231. https://doi.org/10.32604/cmc.2020.06431.