

Improving the term weighting log entropy of latent dirichlet allocation

Muhammad Muhajir^{1,2}, Dedi Rosadi¹, Danardono¹

¹Department of Mathematics, Faculty Mathematics and Natural Science, Universitas Gadjah Mada, Yogyakarta, Indonesia

²Department of Statistics, Faculty Mathematics and Natural Science, Universitas Islam Indonesia, Yogyakarta, Indonesia

Article Info

Article history:

Received Nov 6, 2023

Revised Jan 6, 2024

Accepted Jan 14, 2024

Keywords:

Global weight

Indonesian language

Latent dirichlet allocation

Local weight

Term weighting log entropy

TF-IDF

Topic modelling

ABSTRACT

The process of analyzing textual data involves the utilization of topic modeling techniques to uncover latent subjects within documents. The presence of numerous short texts in the Indonesian language poses additional challenges in the field of topic modeling. This study presents a substantial enhancement to the term weighting log entropy (TWLE) approach within the latent dirichlet allocation (LDA) framework, specifically tailored for topic modeling of Indonesian short texts. This work places significant emphasis on the utilization of LDA for word weighting. The research endeavor aimed to enhance the coherence and interpretability of an Indonesian topic model through the integration of local and global weights. Local Weight focuses on the distinct characteristics of each document, whereas global weight examines the broader perspective of the entire corpus of documents. The objective was to enhance the effectiveness of LDA themes by this amalgamation. The TWLE model of LDA was found to be more informative and effective than the TF-IDF LDA when compared with short Indonesian text. This work improves topic modeling in brief Indonesian compositions. Transfer learning for NLP and Indonesian language adaptation helps improve subject analysis knowledge and precision, this could boost NLP and topic modeling in Indonesian.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Dedi Rosadi

Department of Mathematics, Faculty Mathematics and Natural Science, Universitas Gadjah Mada

Sekip Utara Bulaksumur, Yogyakarta, Indonesia

Email: dedirosadi@ugm.ac.id

1. INTRODUCTION

Natural language processing (NLP) is a machine-learning technology that allows computers to interpret, manipulate, and understand human language. Topic modeling is one of the techniques in NLP that aims to identify and extract topics hidden in collections of text documents, especially short texts such as tweets, product reviews, and public opinion surveys, that are dense with information and public opinion [1]. Topic modeling for short texts presents challenges such as lack of context, extensive configuration requirements, and potential bias in the model due to the brevity of the text. The shortcomings and limitations of such texts demand topic modeling algorithms that effectively extract the themes and meanings contained [2].

The Indonesian language has a unique structure and linguistic richness makes topic modeling more complex, especially for short texts, increasing the need for customized methods to overcome these challenges. latent dirichlet allocation (LDA) assumes the role of the most widely adopted method for this purpose, demonstrating its efficacy across diverse languages [3]–[5]. However, concerning the Indonesian language, additional investigation is essential to refine techniques for generating more accurate and important topic models [6].

Hidayati and Parlina [7] on comparison of topic modeling algorithm performance on Indonesian short texts which compares the topic extraction performance of LDA, non-negative matrix factorization (NMF), and gibbs sampling dirichlet multinomial mixture (GSDMM) algorithms from short Indonesian texts. This study found that LDA outperformed NMF and GSDMM regarding topic coherence scores, but human judgment showed that the word clusters generated by NMF and GSDMM were easier to infer. The present article explores an enhanced approach to evaluating word significance within LDA, specifically emphasizing the term weighting log entropy (TWLE) method for modeling topics in short Indonesian texts.

TWLE method is a topic modeling technique for calculating word significance based on a document's word entropy [8]–[10]. Entropy quantifies the diversity or uncertainty of word occurrences within a document, clarifying the information included by these words. However, traditional log entropy weighting takes an insufficient approach and fails to account for local and global attributes of words present within a corpus [11]–[13]. To address these limitations of log entropy weighting approaches such as log entropy weighting in this research, we developed TWLE as an improved weighting alternative that considers both local and global attributes of words [14], [15]. Local weights help capture context significance through evaluation of document frequency while global weights indicate importance within topic formation by considering corpus wide frequency [16], [17].

The purpose of this research is to improve the log entropy method via TWLE for modeling Indonesian text topics. The complexity of Indonesian short texts is morphologically and semantically complex, and a topic model that can incorporate local and global weights will increase accuracy and the interpretability [18]–[20]. LDA is combined with different topics in order to compare their effects on the modeling results. We hope to gain a thorough understanding of how the improved TWLE technique can produce more precise and accurate topic models [21]. To verify our methodology the study was carried out using Indonesian tweets corpus and Google reviews corpus to contrast enhanced TWLE to popular methods for weighting words such as TF IDF when evaluating performance evaluation metrics like coherence scores and perplexity scores [22].

This research provides two significant contributions. The first is to develop and introduce an improved TWLE algorithm designed for Indonesian language specifically. It integrates local and global weights, while its design takes account of local weighting issues as well [23]. Moreover, extensive experiments were performed on this method as evidence for its efficacy against existing word weighting methods. Based on the structure of this paper, section 2 describes the concept of the term weighted log entropy and the evaluation metrics for LDA modeling. In section 3, the results of the study are discussed, and the last section discusses the conclusion.

2. METHOD

2.1. Weight function

A weight function is used to transform the X matrix, enabling an inference regarding the information present at the intersection of row i and column j . This function provides a more precise approximation of the document-term correlation. In this context, the column corresponds to the document, while the row signifies the term, which may be a keyword or a phrase. The product of the local weight and the global function is an expression of this transformation in (1) [24].

$$a(i, j) = L(i, j) * G(i) \quad (1)$$

The weight assigned to term i in document j is represented by the local weight function $L(i, j)$. $G(i)$ provides an expression of weight of term i within all documents in a set.

2.2. Local weighted

Local weight function that are considered trivial are proportional to the frequency of term i in document j , and are denoted $tf(i, j)$ [14]. A logarithmic scale can also be applied for term frequency to mitigate its negative impacts from exceedingly large numerical values [25]. The following equations define local weighting schemes in (2) [11].

$$L(i, j) = \log (tf(i, j) + 1) \quad (2)$$

2.3. Global weighted

Global weighting encompasses all of the training document collection. The goal of global term weighting is to assign discriminative values to individual terms with an eye toward those that exhibit more discriminations [25]. In (3), it can be shown that its function is widely known, and similar to global entropy [11].

$$G(i) = 1 + \frac{\{\sum_j p(i,j) \log p(i,j)\}}{\log (ndocs+1)} \tag{3}$$

Other global weighted functions can be defined using symbols like $gf(i)$, representing the global frequency of term i . $df(i)$, representing documents that contain term i , and $ndocs$ as representing total documents or text fragments under consideration. Furthermore, conditional probability $p(i,j) = tf(i,j)/gf(i)$ for document j given the presence of term i is multiplied by $df(i)$.

2.4. TWLE

TWLE weighting is one of the word weighting methods utilized by LDA. TWLE involves assigning high weight to words which contain important term in document [21]. TWLE assigns heavy weight to words with low entropy, reflecting its contribution in defining subjects and topics within documents [16], [26]–[28]. TWLE weighting is expressed in the (4).

$$TWLE(i, j) = L(i, j) * G(i) \tag{4}$$

Documents will be more clear if you give importance to key words. This will also lead to better accuracy and interpretation of LDA topic model results [8], [29].

2.5. Topic coherence

Coherence value (C_v) is an assessment measure to measure the quality of topics within a set of words, measuring how closely related words and concepts representing topics are semantically. Coherence value also serves to evaluate optimal numbers of topics for LDA models often denoted by its symbolic symbol of C_v based on sliding windows, segmented lists of top words, and an indirect confirmation metric such as normalized unidirectional mutual information (NPMI). Calculations can be completed using this formula (5) [30]–[32].

$$C_v = \frac{2}{K.(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K NPMI (w_i, w_j) \tag{5}$$

Were,

$$NPMI (w_i, w_j) = \left(\frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i).P(w_j)}}{-\log (P(w_i, w_j) + \epsilon)} \right)$$

2.6. Perplexity

Evaluation of probabilistic models often utilizes log likelihood analysis on test sets that have been collected as samples for training or testing purposes. Datasets typically consist of one portion for training purposes and another intended solely for testing purposes. In LDA, a test set comprises unseen documents denoted as w_d , and the model is defined by the topic matrix Φ and the hyperparameter α for document topic distribution. LDA parameters θ are omitted since they represent the topic-distributions for the documents in the training set and can be ignored when calculating the likelihood of unseen documents. As a result, the assessment focuses on the log-likelihood [33].

$$\mathcal{L}(w) = \log P(w|\Phi, \alpha) = \sum_d \log P(w_d|\Phi, \alpha) \tag{6}$$

To assess models, the likelihood of unseen documents w_d is computed, considering the topics Φ and the hyperparameter α for the topic distribution $\theta_d w_d$ of documents. This likelihood serves for model comparison, where a higher likelihood indicates a superior model. The perplexity of held-out texts often serves as a measure for topic models and is defined in (7).

$$Perplexity (W) = \exp \left\{ - \frac{\mathcal{L}(w)}{\text{count of tokens}} \right\} \tag{7}$$

Perplexity is a decreasing function of the unseen documents w_d and log-likelihood $\mathcal{L}(w)$. Smaller perplexity values signify better models but due to the intractable nature of the likelihood $P(w_d|\Phi, \alpha)$ for a single document, evaluating $\mathcal{L}(w)$ hence perplexity is also intractable. Several sampling approaches have been developed to estimate this probability [34]–[36].

2.7. Research stage

This research advances the use of TWLE in the process of latent dirichlet allocation. The study makes use of TWLE to enhance LDA model accuracy as well as efficiency. Figure 1 depicts the process of research, which starts with the incorporation into TWLE to LDA and then concludes by analyzing and evaluating the results to evaluate the improvement.

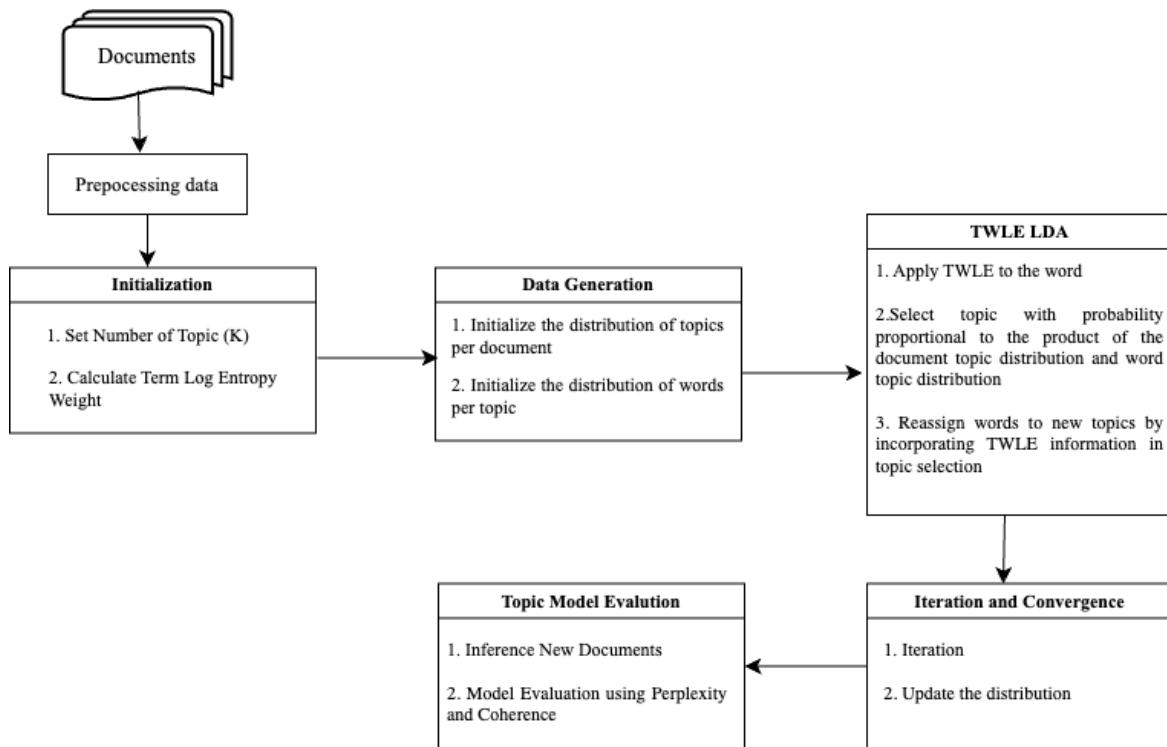


Figure 1. Research flowchart

3. RESULTS AND DISCUSSION

3.1. Experiment

The study applied a data collection technique utilizing a Python application's text crawling feature to analyze Indonesian users' reviews and comments on the Twitter platform regarding Hacker Bjorka. The data was gathered under the hashtags "Bjorkanism" and "Hacker Bjorka" resulting in a total of 3,061 comments, as well as a Google Play Store review for the XL app. Data taken was 1,000 data containing feedback on the performance and functionality of the XL application available for analysis.

A preprocessing step comprising various techniques was subsequently carried out. The techniques included case folding, elimination of account names and hashtags, removal of URLs, deletion of punctuation, normalization of words, elimination of stopwords using the Sastrawi package, and exclusion of specific words. These steps aimed to minimize cumulative discrepancies in verb tenses. To maintain clarity in the outcome, word stemming and complex terminology pairs were deliberately avoided, as they could result in the selection of one tense compared to others.

3.2. Improved the weight of LDA

After reviewing the literature on text based information retrieval, TWLE weighting scheme [27] or the TF-IDF scheme [28] were recommended due to their superior retrieval performance compared to IDF and raw term frequency. The objective of the two forthcoming experiments was to investigate the efficacy of the two predominant weighting schemes, namely TF-IDF, and TWLE, in the context of LDA based information retrieval system, specifically on extensive datasets. Table 1 listed the weights of TWLE and TF-IDF with perplexity and coherence evaluation metrics for the number of distinct topics used in this study.

The Table 1 showed that the TWLE model had a lower perplexity value, indicating better extraction of topics from the data. However, for a higher number of topics, the perplexity value increased, making it more challenging to efficiently describe the data. The difference in perplexity values between TWLE and LDA

models using TF-IDF was insignificant, denoting similar performance. Perplexity values for various topics tended to fluctuate without a clear trend. In general, both TWLE and TF-IDF methods yield similar performance in terms of perplexity. So, to determine the superiority of the model can be seen from the coherence value.

The TWLE coherence score model has the highest coherence score (0.233) for topic 3, but there are variations between different topics. The TF-IDF model has a slightly lower consistency score for each number of subjects, but the difference is insignificant. Similar scores of coherences were observed also for Topics 3 and 4. The difference in coherence scores is due to the difference in the way the data is modeled and represented using TWLE and TF-IDF. Overall, the TWLE and LDA models using TF-IDF performed similarly in terms of consistency scores, but the TWLE model tended to have slightly higher consistency scores than the LDA model using TF-IDF. This means it is likely that the TWLE model is more effective when evaluated as a perplexity and coherence scoring measurement.

Table 1. Perplexity and coherence measurements for TWLE and TF-IDF

Data	Evaluation matrix	Number of topics (n)	TWLE	TF-IDF		
Twitter	Perplexity	2	-8.358	-8.334		
		3	-8.489	-8.473		
		4	-8.572	-8.560		
		5	-8.665	-8.642		
		2	0.280	0.326		
	Coherence	3	0.335	0.335		
		4	0.278	0.305		
		5	0.267	0.306		
		Google review	Perplexity	2	-8.358	-8.334
				3	-8.489	-8.473
4	-8.572			-8.560		
5	-8.665			-8.642		
2	0.280			0.326		
Coherence	3	0.335	0.335			
	4	0.278	0.305			
	5	0.267	0.306			

Through the division of this data into two suitable subsets, it's possible to effectively train the LDA model based on the training data, while analyzing the model's performance and determining the most optimal parameters in light of the never ever before seen test data. This will allow an accurate evaluation of the efficiency of the LDA approach by using TWLE as well as the TF-IDF algorithm to enhance the generation of topics. Table 2 displays the scores of coherences and perplexity for both the tests and training data using TWLE and TF-IDF weights on all subjects.

Table 2. Perplexity and coherence scores of TWLE and TF-IDF weights for testing and training data

Data	Evaluation matrix	Number of topics (n)	TWLE model		TF-IDF model			
			Training	Testing	Training	Testing		
Twitter	Perplexity	2	-7.944	-8.044	-7.950	-8.050		
		3	-8.020	-8.138	-8.014	-8.140		
		4	-8.073	-8.211	-8.074	-8.189		
		5	-8.117	-8.267	-8.121	-8.288		
		2	0.234	0.395	0.234	0.292		
	Coherence	3	0.258	0.358	0.268	0.334		
		4	0.244	0.391	0.244	0.377		
		5	0.263	0.387	0.251	0.398		
		Google review	Perplexity	2	-7.944	-8.044	-7.950	-8.050
				3	-8.020	-8.138	-8.014	-8.140
4	-8.073			-8.211	-8.074	-8.189		
5	-8.117			-8.267	-8.121	-8.288		
2	0.234			0.395	0.234	0.292		
Coherence	3	0.258	0.358	0.268	0.334			
	4	0.244	0.391	0.244	0.377			
	5	0.263	0.387	0.251	0.398			

Table 2 illustrates the usage of various weighting models for the TF-IDF and TWLE models for the topic model. The TWLE model has a lower complexity, and the perplexity point has a variety in the testing and training data sets with the same pattern. In particular, the complexity of the testing data is lower than that

of the training data, indicating the model's effectiveness. This demonstrates the superiority of TWLE, as evidenced by the consistently lower perplexity scores compared to TF-IDF.

Based on the coherence values, the TWLE model outperforms the TF-IDF model, where the TF-IDF model achieves slightly higher coherence values for some topics. The difference in coherence value between the training data and the test data remains the same. For example, TWLE achieved the highest test score of 0.395 in two topics. On the other hand, TF-IDF recorded slightly lower coherence values, but with a similar trend for different topics. Both TWLE and TF-IDF show similar performance in terms of coherence scores, but there are slight differences between them. A comparison of the execution time of TWLE and TF-IDF weights in LDA is conducted to identify the differences in execution speed and computational efficiency and optimize efficiency. Figure 1 depicts a runtime analysis that can determine which method has shorter execution time or requires fewer resources, leading to a more efficient and faster solution.

Based on Figure 2, we can see how the execution time for the TWLE and TF-IDF models changes depending on the number of topics used in the topic analysis for the Twitter and Google Review datasets. For the Twitter dataset, the TWLE model starts at around 48 ms and shows a steady decrease to 6.42 ms as the number of topics is increased from 2 to 10. The TF-IDF model for Twitter shows a similar pattern of decrease up to 6 topics, after which the execution time stabilizes at around 24 ms, without any further decrease. On the Google Review dataset, the TWLE model shows a consistent decrease from 87 ms to 18.6 ms as the number of topics increases, indicating an increase in computational efficiency with an increase in model complexity. On the other hand, the TF-IDF model for Google Review starts with a higher time at 2 topics and decreases to about 35.7 ms at 10 topics, but experiences an increase in execution time at 8 topics before dropping again, indicating instability in execution time as the number of topics increases.

These graphs generally reflect that both models become more time efficient when they are faced with more complex topic modeling tasks, but the TWLE model shows a more consistent decrease in execution time compared to the TF-IDF model. This means that as the number of topics in the topic analysis increases, the model using the TWLE approach experiences a continuous and predictable decrease in the time required to complete its task (runtime) from one topic count to the next. On the other hand, the model using the TF-IDF approach does not show the same pattern of decrease; either the execution time is stable or it does not decrease regularly with the increase in the number of topics. In this context, 'consistent' indicates that there is a regular and reliable decreasing pattern in the time performance of the TWLE model which is not as obvious as that of the TF-IDF model.

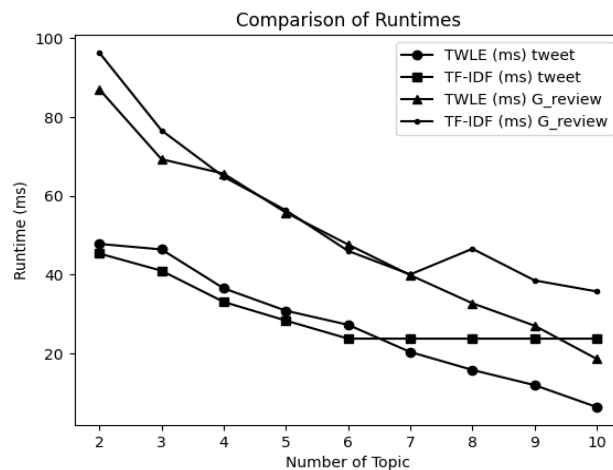


Figure 2. Comparison of TWLE and TF-IDF weight execution times

4. CONCLUSION

The comparative study concludes that TWLE LDA is superior to the TF-IDF LDA. The TWLE method, which incorporates both local and global weights, performs better than the TF-IDF approach in terms of various evaluation metrics. These include scores for coherence and perplexity. The TWLE method of weighting is better at identifying coherent and informative topics within short Indonesian text, particularly in experiments. TWLE produces a more constant runtime compared with the TF-IDF model. This indicates that the increase in topics will lead to a decrease in runtime.




REFERENCES

- [1] C. D. P. Laureate, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14223–14255, Dec. 2023, doi: 10.1007/s10462-023-10471-x.
- [2] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5133–5260, Jun. 2023, doi: 10.1007/s10462-022-10254-w.
- [3] A. Srivastav and S. Singh, "Proposed model for context topic identification of English and Hindi news article through LDA approach with NLP technique," *Journal of The Institution of Engineers (India): Series B*, vol. 103, no. 2, pp. 591–597, Apr. 2022, doi: 10.1007/s40031-021-00655-w.
- [4] M. Inoue, H. Fukahori, M. Matsubara, N. Yoshinaga, and H. Tohira, "Latent dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing," *Japan Journal of Nursing Science*, vol. 20, no. 2, Apr. 2023, doi: 10.1111/jjns.12520.
- [5] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, "Effective comparison of LDA with LSA for topic modelling," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 1245–1250. doi: 10.1109/ICICCS48265.2020.9120888.
- [6] A. Djuraidah, B. Sartono, and Y. Putranto, "Topic modelling and hotel rating prediction based on customer review in Indonesia," *International Journal of Management and Decision Making*, vol. 20, no. 1, 2021, doi: 10.1504/IJMDM.2021.10036033.
- [7] N. N. Hidayati and A. Parlina, "Performance comparison of topic modeling algorithms on Indonesian short texts," in *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, Nov. 2022, pp. 117–120. doi: 10.1145/3575882.3575905.
- [8] H. Núñez *et al.*, "Using entropy-based local weighting to improve similarity assessment," *Departament de Ciències de la Computació*, 2002.
- [9] M. Ghosh and A. Dey, "Fractional-weighted entropy-based fuzzy G-2DLDA algorithm: a new facial feature extraction method," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2689–2707, Jan. 2023, doi: 10.1007/s11042-022-13328-7.
- [10] J. Xu, "A weighted linear discriminant analysis framework for multi-label feature extraction," *Neurocomputing*, vol. 275, pp. 107–120, Jan. 2018, doi: 10.1016/j.neucom.2017.05.008.
- [11] P. Nakov, A. Popova, and P. Mateev, "Weight functions impact on LSA performance," in *Proceedings EuroConference Recent Advance Natural Language Processing RANLP 2001*, 2001, pp. 187–193.
- [12] S. Koltcov, V. Ignatenko, and O. Koltsova, "Estimating topic modeling performance with sharma–mittal entropy," *Entropy*, vol. 21, no. 7, Jul. 2019, doi: 10.3390/e21070660.
- [13] J. Risch and R. Krestel, "My approach = your apparatus?," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, May 2018, pp. 283–292. doi: 10.1145/3197026.3197038.
- [14] A. N. K. Zaman and C. G. Brown, "Latent semantic indexing and large dataset: study of term-weighting schemes," in *2010 Fifth International Conference on Digital Information Management (ICDIM)*, Jul. 2010, pp. 1–4. doi: 10.1109/ICDIM.2010.5664669.
- [15] R. N. Rathi and A. Mustafi, "The importance of term weighting in semantic understanding of text: A review of techniques," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 9761–9783, Mar. 2023, doi: 10.1007/s11042-022-12538-3.
- [16] X. Li, A. Zhang, C. Li, J. Ouyang, and Y. Cai, "Exploring coherent topics by topic modeling with term weighting," *Information Processing & Management*, vol. 54, no. 6, pp. 1345–1358, Nov. 2018, doi: 10.1016/j.ipm.2018.05.009.
- [17] Y. Zou, J. Ouyang, and X. Li, "Supervised topic models with weighted words: multi-label document classification," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 4, pp. 513–523, Apr. 2018, doi: 10.1631/FITEE.1601668.
- [18] R. K. Dey and A. K. Das, "Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32967–32990, Sep. 2023, doi: 10.1007/s11042-023-14653-1.
- [19] Indra, E. Winarko, and R. Pulungan, "Trending topics detection of Indonesian tweets using BN-grams and Doc-p," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 2, pp. 266–274, Apr. 2019, doi: 10.1016/j.jksuci.2018.01.005.
- [20] E. Aytac and M. Khayet, "A Topic modeling approach to discover the global and local subjects in membrane distillation separation process," *Separations*, vol. 10, no. 9, Sep. 2023, doi: 10.3390/separations10090482.
- [21] Z. Tang, W. Li, and Y. Li, "An improved supervised term weighting scheme for text representation and classification," *Expert Systems with Applications*, vol. 189, Mar. 2022, doi: 10.1016/j.eswa.2021.115985.
- [22] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–30, Mar. 2021, doi: 10.1155/2021/6619088.
- [23] S. S. Samant, N. L. B. Murthy, and A. Malapati, "Improving term weighting schemes for short text classification in vector space model," *IEEE Access*, vol. 7, pp. 166578–166592, 2019, doi: 10.1109/ACCESS.2019.2953918.
- [24] D. I. Witter, "Downdating the latent semantic indexing model for conceptual information retrieval," *The Computer Journal*, vol. 41, no. 8, pp. 589–601, Aug. 1998, doi: 10.1093/comjnl/41.8.589.
- [25] M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, "Short text clustering algorithms, application and challenges: a survey," *Applied Sciences*, vol. 13, no. 1, Dec. 2022, doi: 10.3390/app13010342.
- [26] I. Niño-Adan, D. Manjarres, I. Landa-Torres, and E. Portillo, "Feature weighting methods: a review," *Expert Systems with Applications*, vol. 184, Dec. 2021, doi: 10.1016/j.eswa.2021.115424.
- [27] S. T. Dumais, "Improving the retrieval of information from external sources," *Behavior Research Methods, Instruments, & Computers*, vol. 23, no. 2, pp. 229–236, Jun. 1991, doi: 10.3758/BF03203370.
- [28] C. D. Manning, P. Raghavan, and H. Schütze, *An introduction to information retrieval*. Cambridge University Press, 2008.
- [29] M. D. Lee, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text document similarity," in *XXVII Annual Conference of the Cognitive Science Society*, 2005, pp. 1254–1259.
- [30] S. Lafia, W. Kuhn, K. Caylor, and L. Hemphill, "Mapping research topics at multiple levels of detail," *Patterns*, vol. 2, no. 3, Mar. 2021, doi: 10.1016/j.patter.2021.100210.
- [31] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 830–836. doi: 10.18653/v1/D18-1096.
- [32] S. H. Mohammed and S. Al-Augby, "LSA & LDA topic modeling classification: comparison study on e-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.
- [33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [34] L. Huang, J. Ma, and C. Chen, "Topic detection from microblogs using T-LDA and perplexity," in *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, Dec. 2017, pp. 71–77. doi: 10.1109/APSECW.2017.11.




- [35] P. Tijare and P. J. Rani, "Exploring popular topic models," *Journal of Physics: Conference Series*, vol. 1706, no. 1, Dec. 2020, doi: 10.1088/1742-6596/1706/1/012171.
- [36] B.-X. Du and G.-Y. Liu, "Topic analysis in LDA based on keywords selection," *Journal of Computers*, vol. 32, no. 4, pp. 1–12, Aug. 2021, doi: 10.53106/199115992021083204001.

BIOGRAPHIES OF AUTHORS






Muhammad Muhajir    is currently doctoral students Mathematics at Universitas Gadjah Mada. He graduated from Bachelor of Statistics Program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia in 2011. He earned his Master of mathematics degree from the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University, and graduated in august 2014. Then, he continued to work in the statistics study program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia until now. The subjects of his specialization include: data mining, machine learning, multivariate, data science, and text mining. He can be contacted at email: mmuhajir@uii.ac.id.



Dedi Rosadi    currently works as the (full) professor at the research group Statistical Computing, the Department of Mathematics, Universitas Gadjah Mada. He graduated from the Statistics Study Program at Gadjah Mada University in February 1996 and started to work as a lecturer at UGM after this graduation. In August 1997, he continued his study to obtain a master of science degree in Stochastic Modeling (Applied Statistics) at the University of Twente, The Netherlands, and graduated in June 1999. From September 2001 to September 2004, he took my doctoral study (in Econometrics) at the Institute of Econometrics and Operation Research (EOS 119), the Vienna University of Technology (TU Wien), Austria. After finishing his doctoral study, he came back to Yogyakarta and continued work at the Universitas Gadjah Mada until now. In September 2013, he received the full professor title. The subjects of his specialization include biostatistics, data science, (statistical) machine learning cluster-statistics, computational statistics, statistical finance, and time series analysis cluster-statistics. He can be contacted at email: dedirosadi@ugm.ac.id.



Danardono    is an Associate Professor at Department of Mathematics Universitas Gadjah Mada, Yogyakarta. He graduated from the Statistics Bachelor Program, Faculty of Mathematics and Natural Sciences, Gadjah Mada University (UGM) in 1992. Since 1994 he has worked in the Faculty of Mathematics and Natural Sciences, UGM; and is also affiliated with the Clinical Epidemiology Unit, Faculty of Medicine, UGM. He got a Master of Public Health in Biostatistics (MPH) from the Department of Biostatistics and Demography, Faculty of Public Health, Khon Kaen University, Thailand. The field of epidemiology and medicine motivated him to do methodological research in this area as his doctoral study, in 2005, he got a Ph.D. in Statistics from Umeå University, Sweden. His main research interests are in survival data (time-to-event data) analysis and longitudinal data analysis. His research focuses on developing and applying survival and longitudinal data analysis for epidemiological, medical, and actuarial problems. He can be contacted at email: danardono@ugm.ac.id.