Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 3, March 2026, pp. 1017~1024
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i3.pp1017-1024

Exploring word embeddings and clustering algorithms for user reviews

Zuleaizal Sidek[1,2], Sharifah Sakinah Syed Ahmad[1]
[1]Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia
[2]Institut Tun Perak, Melaka, Malaysia

Article Info

Article history:

Received Apr 18, 2024
Revised Jan 15, 2026
Accepted Feb 26, 2026

Keywords:

Clustering algorithms
Silhouette score
Text analysis
User reviews
Word embedding

ABSTRACT

The rapid advancement of information technology has led to a significant surge in the volume of unstructured textual data. This has posed a major problem in terms of analyzing, organizing, and automatically clustering text for research purposes, which is crucial for extracting valuable insights. The process of manually clustering the unstructured data, such as customer reviews on the Internet, which capture the opinions of customers regarding products, services, and social events, requires significant financial resources, manpower, and time. Most of the studies are directed towards the analysis of sentiment in user reviews. In order to address the issues effectively, automated text clustering could assist in categorizing reviews into various themes, thereby simplifying the analysis process. Therefore, in this paper, we present and compare the result of experiment the combination of five text clustering techniques, namely K-means, fuzzy C-mean (FCM), non-negative matrix factorization (NMF), latent dirichlet allocation (LDA), and latent semantic analysis (LSA) with different embedding techniques, namely term frequency–inverse document frequency (TF-IDF), Word2Vec, and global vectors (GloVe). The experiments revealed that LDA is a reliable algorithm as it consistently produces good results across three-word embeddings. The highest Silhouette score recorded in the experiments was 0.66 using LDA and Word2Vec as word embedding. Simultaneously, the application of LSA in conjunction with Word2Vec yields superior outcomes, as evidenced by a Silhouette score of 0.65.

Corresponding Author:

Sharifah Sakinah Syed Ahmad
Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka (UTeM)
76100 Melaka, Malaysia
Email: sakinah@utem.edu.my

1. INTRODUCTION

The emergence of the age of big data has resulted in an enormous amount of data being spread over all aspects of our lives. The ever-growing amount of textual information has created challenges to customers in finding the desired content. Before this, textual data were manually classified and clustered which is time-consuming, difficult, and costly. In today's world, it is evident that relying solely on manual text classification and clustering is insufficient to address the requirements. The emergence of automatic text classification and clustering significantly contributes to the state of the art of natural language processing (NLP).

Approximately 80% of the material now accessible on the Internet exists in an unstructured, unlabeled text-based. The exponential growth of unstructured text data can be attributed to the widespread availability of digital information, including emails, text messages, blogs, social media posts, and product

evaluations. For people using e-commerce platforms, the process of sifting through numerous reviews prior to making a purchase can present a formidable challenge. Unsupervised learning enables the exploration of difficulties that lack a labeled dataset and do not have prior knowledge regarding the outcome.

Clustering is an analytical method employed for discovering the structural relationships between variables. Unsupervised machine learning methodologies can be employed to autonomously assess pre-processed data derived from these websites, with the aim of enhancing the user experience for consumers prior to making a purchase. In addition to assisting potential consumers in making an informed purchase decision, reviews written by current customers allow product teams to fine-tune their offerings. These days, businesses rely on consumer reviews and other forms of feedback to gain valuable insights in this digital, diverse range of reviews, which often consist of a substantial volume [1]. Various methodologies, including topic modeling and text clustering, offered several benefits such as identifying common themes, gaining insight into the strengths and weaknesses of the business, feedback prioritization, market research, competitive analysis and customer engagement.

Text clustering can only be performed following the completion of the vectorization procedure, which is referred to as word embedding. Word embeddings convert textual data, which is incomprehensible to machine learning algorithms, into a numerical format that they can capture the contextual basic form of words, both in terms of their semantic and grammatical similarities, as well as their relationship with other words. For the purpose of detecting relevant and useful terms within user reviews, various techniques have been utilized, including frequency-based word embedding methods like term frequency-inverse document frequency (TF-IDF). This study primarily concentrates on the clustering of an unlabeled dataset through the utilization of several word embedding strategies, namely Word2Vec [2], GloVe [3], and bidirectional encoder representations from transformers (BERT) [4]. In this paper, five unsupervised techniques have been utilized for the purpose of clustering user reviews.

First, K-means clustering is an algorithm that defines clusters as partitions of data [5] but [6] mentioned it fails to leverage prior knowledge about the distribution of hidden class labels obtained from limited labeled data. Despite its limitations, the K-means clustering algorithm is credited with flexibility, efficiency, and ease of implementation. The simplicity and low computational complexity have given the K-means clustering algorithm a wide acceptance in many domains for solving clustering problems. A study done by [7] showed k-means clustering achieved the best result with a Silhouette Score of 0.6.

Second, in the domain of information retrieval, document modeling, and clustering, latent dirichlet allocation (LDA) is a technique that is used to identify the latent topic structure of textual data and it is an unsupervised machine learning technique [8]. The major benefit of LDA is that it can deal efficiently with the variation of both words and documents [9]. LDA could be a useful tool to automatically discover underlying topics within the complaint's dataset [10]. While [11] mentioned LDA is easy to implement, understand and use but it renders poor results when the number of training images is large.

Third, typically, the non-negative matrix factorization (NMF) can be interpreted as a dimensionality reduction method, which factorizes the instance matrix into non-negative low-rank approximations. The literature analyzes the equivalence of NMF and spectral clustering, indicating that NMF can be also applied to clustering [12]. NMF methods that have been used so far are mostly unsupervised, which means they do not pay much attention to any supervised information that could be hidden in the data [13], [14] also mentioned NMF can reduces the sparsity from 86% to 44.9%.

Forth, fuzzy C-means (FCMs) methods determine which data belong to clusters, they often provide better results than definite methods. The FCM cluster is the most popular fuzzy method. Its simplicity is one of the positive features of the FCM method [15], [16] says the bigger the fuzzification constant implies that each textual data may have more topics. While [17] mentioned FCM algorithm can overcome inconsistency and ambiguity in data and can produce more complex groups. Study done by [18] observed that the best results in grouping the unlabeled data, unsupervised learning technique were obtained with the K - Means and FCMs algorithms.

Fifth, latent semantic analysis (LSA) is a widely used NLP technique that aims to examine topics underlying the corpus. LSA relies on the following main idea: words with similar meanings will occur in similar pieces of text [19]. However, a major drawback of LSA is its expensive computational cost [9]. The LSA technique is classified as an unsupervised learning method, as it lacks a ground truth. The presence or absence of latent concept is uncertain. Throughout the 1990s and into the 2000s, LSA was demonstrated to be able to model various cognitive functions, including the learning and understanding of word meaning [20].

The results of the experiments have been examined and compared with the state-of-the-art word embedding techniques available. The paper is structured as follows: In section 2, explains the methodology on how the experiment is carrier out. The details about the analysis of the findings are highlighted in section 3. In conclusion, the paper is summarized in section 4.

2. METHOD

To be able to conduct an extensive study into the word embeddings in the text clustering model, the following Figure 1 provides an overview of our methodology.

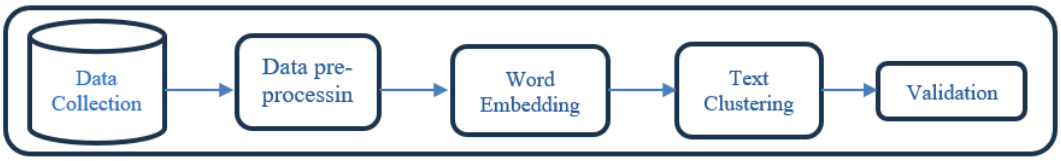

Figure 1. Research methodology

2.1. Data collection

The data collection process involves retrieving or gathering the necessary data for analysis. The data can be acquired from a variety of sources, such as social media platforms, online review platforms, and e-commerce websites selling multiple products and services. For this paper, the reviews that were gathered consisted of textual reviews that were acquired from Kaggle and crawled from its website.

2.2. Data pre-processing

Preprocessing refers to the activity of preparing data in order to enhance its accuracy, ensuring that the results accurately reflect the data [21] and are significant to the research objective. We do data preprocessing primarily to transform raw data into a machine-readable format because it may have the tendency to contain a significant quantity of noise as well as textual information that is not useful [22]. Besides, we are also removing special characters, stopwords, punctuations, duplications, URLs and tokenization are crucial steps in this process that ensure the data is clean before the vectorization process. Importantly, we do not perform lemmatization and stemming as some research warns against hastily using stemming or lemmatization because it might change the results [23]. To ensure the user reviews clustering model learns the relationships between words in the corpus, these procedures must be applied. The quality of the model's results depends on thorough data preprocessing, which can be a time-consuming procedure.

2.3. Word embedding

Word Vectorization or Word Embedding refers to the process of transforming words into a vector. This vector is then used to make predictions about words and identify similarities or semantics between words as word embedding has been demonstrated to be useful in a number of natural language processing tasks [24]. In this paper, there are 3 word embedding techniques are used: 1) TF-IDF (using sklearn TfidfVectorizer); 2) Word2Vec (using python genism.model Word2Vec); and 3) GloVe (using keras tensorflow). To extract valuable information from natural language text or sentences using machine learning and deep learning techniques, it is necessary to convert the text into a vector [25].

2.4. Text clustering

Text clustering refers to the systematic procedure of grouping documents that have similar content into the same group. Text clustering enables the discovery of complex and significant relations between words in a corpus that would otherwise be challenging to cluster. K-Means, FCMs, LSA, LDA, and NMF were selected to cluster the user reviews that has been through word embedding.

2.5. Validation

When the embedded text data has been clustered using a selected clustering technique, the subsequent step involves analyzing the results by comparing the silhouette score of the clustered data and visualizing the findings. The Silhouette Coefficient or silhouette score is a quantitative measure used to evaluate the effectiveness of a clustering technique. The range of the value of the variable spans between -1 to 1 following the silhouette score (1).

$$SC = \frac{1}{n} X \sum_{i}^{n} \frac{b-a}{\max(a,b)} \quad (1)$$

a - distance between data point within cluster
b - distance between cluster

3. RESULTS AND DISCUSSION

For this experiment, three datasets were selected: Shopee customer reviews (e-commerce) [26], Malaysia restaurant reviews (Food & Beverage) [27], and Malaysia telecommunication reviews (Service) [28]. The file sizes of both datasets were 46MB, 75MB, and 5MB, respectively. Shopee customer reviews consist of reviews made by customers who have made purchases on the Shopee electronic commerce platform, with a special emphasis on the English language. Both the Malaysia restaurant reviews and Malaysia telecommunication reviews databases consist of reviews sourced from TripAdvisor and Google, respectively. It is interesting to note that both datasets include reviews written in both English and Malay languages. We used a total of five different algorithms to analyze the dataset. Evaluation was conducted utilizing silhouette score measurement metrics in association with three distinct word embedding techniques TF-IDF initially, Word2Vec, and GloVe. We performed on three selected datasets mentioned before. A separate table is used to present the results acquired for each dataset.

3.1. Shopee customer reviews

Prior to conducting the word embedding and clustering procedures, the Elbow Method is employed to determination of the optimum K value is based on the presence of an elbow curve on the graph, that is 4. Results for experiment using dataset Shopee customer reviews as shown in Table 1. Table 1 presents interesting findings, as LDA algorithm consistently produces similar results across various word embeddings. LDA is a highly efficient algorithm for effectively handling substantial volumes of structured and semi-structured textual data [29]. This finding illustrates that LDA offers its ability to employ any word embedding techniques. Despite the consistent results produced by LDA, LSA has outperform performance compared to other algorithms, as shown by its Silhouette score of 0.65. On the other hand, except for LDA, 4 out of 5 algorithms demonstrate very low performance when utilizing TF-IDF. Word2Vec is a technique that can be considered reliable for word embeddings as it has significant efficacy compared to GloVe, especially with regard to TF-IDF. Furthermore, [23] provides further evidence that TF-IDF is not as effective as the other techniques when it comes to extracting relevant text documents. According to [30], FCMs will not produce better results as it demonstrates superior clustering performance for large image datasets compared to text, with a lower error rate. The silhouette score for FCMs has been demonstrated to be 0.001, 0.31, and 0.009 when employing TF-IDF, Word2Vec, and GloVe, respectively. Based on the results of the experiment, it can be assumed that the use of W0rd2Vec and GloVe as word embeddings provides better results for both LDA and LSA.

The LSA clustering visualization depicted in Figure 2 that splits the Shopee customer reviews into four distinct clusters which are 0, 1, 2, and 3 that consist 9982, 6, 7, and 5 samples, respectively. The vast majority of samples are located in cluster 0. The remaining clusters hold insignificant quantities of samples due to their distance from the centroid.

Table 1. Results for Shopee reviews dataset

Algorithm	Results (Silhouette Score)		
	TF-IDF	Word2Vec	GloVe
K-means	0.007	0.31	0.06
Fuzzy C-means	0.001	0.31	0.009
Non-negative matrix factorization (NMF)	0.01	0.19	0.14
Latent dirichlet allocation (LDA)	0.59	0.59	0.59
Latent semantics analysis (LSA)	0.07	0.65	0.57

3.2. Malaysia restaurant reviews

For the second experiment, Malaysia restaurant reviews from TripAdvisor.com website are used and the results are recorded in Table 2. In Table 2, LSA continues to outperform the other algorithms, but slightly lower than the first experiment, where it achieved a score of 0.63 using Word2Vec as word embeddings. In both the first and second experiments, the use of Word2Vec as word embeddings leads to better outcomes for LSA. In this experiment, the gap in Silhouette scores between LSA and LDA is about 0.01. According to [30], LDA gets a better result than LSA. However, our results show the opposite for both experiments. This is rather interesting. It should also be highlighted that LSA produces a fairly acceptable result of 0.42 when using GloVe as word embedding, as compared to TF-IDF. While, The LDA algorithm consistently produces similar results when applied to the three different word embedding techniques. It confirms [30] that both LDA and LSA are able to generate relevant topic on dataset. Similarly to the previous experiment, the use of TF-IDF and GloVe to K-Means, FCMs, and NMF gives unsatisfactory outcomes. The word embedding techniques used in the three clustering algorithms outlined earlier fail to achieve a Silhouette Score of 0.1 or

higher. Table 2 further demonstrates the extreme poor results that the TF-IDF generates, with Silhouette score values as low as 0.005 and even lower, 0.001, when compared to the first experiment. This demonstrates that applying frequency-based word embedding to text clustering semantically is inappropriate.

Figure 3 depicts the visualization of the LSA clustering using Word2Vec for the second experiment. It seems nearly identical to the first experiment, however it is different as the Silhouette score slightly lower compared to the first experiment. The distinction is also seen in the visualization edge, which appears more disperse compared to the previous, which exhibits a darker coloration.

Table 2. Results for Malaysia restaurant reviews dataset

Algorithm	Results (Silhouette Score)		
	TF-IDF	Word2Vec	GloVe
K-means	0.009	0.36	0.07
Fuzzy C-means	0.005	0.32	0.07
Non-negative matrix factorization (NMF)	0.01	0.31	0.07
Latent dirichlet allocation (LDA)	0.62	0.62	0.62
Latent semantics analysis (LSA)	0.11	0.63	0.42

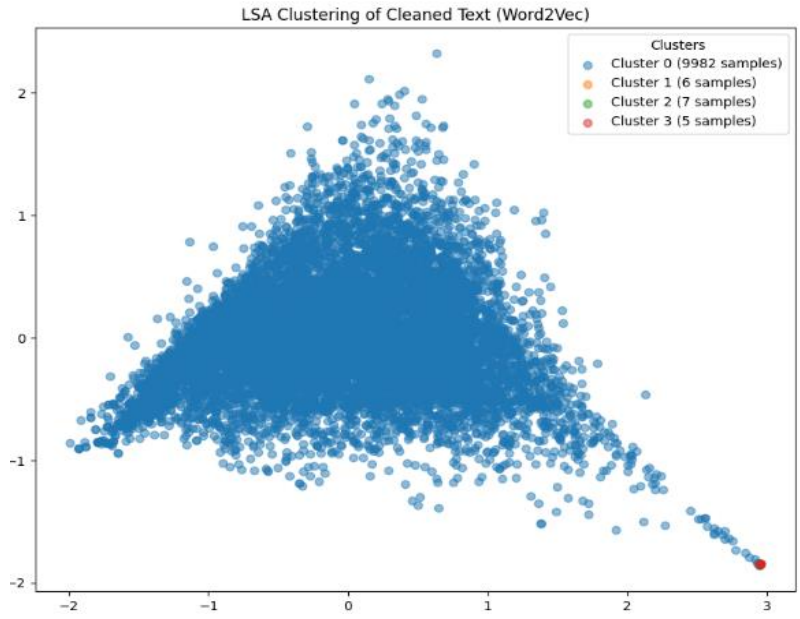

Figure 2. The LSA clustering (Shopee customer reviews)

3.3. Malaysia telecommunication reviews

The third experiment utilized the Malaysia telecommunication reviews dataset, which contains reviews written in both English and Malays by native Malaysian. This dataset was selected due to its combination of mix languages and its small size, at only 9MB. This experiment aims to investigate a small set of data and compare it with the findings of the second and third experiments. The results are presented in Table 3. Remarkably, LDA outperforms the results compared to the other algorithms. Additionally, LDA outperforms the results of the two preceding experiments with 0.66 Silhouette score. This validates the findings of [30] which indicate that LDA gives better outcomes compared to LSA. This is also demonstrates the efficacy of LDA in handling small datasets, specifically user reviews as mentioned by [31] that the LDA algorithm produced better results, suggesting that it is possible to generate more coherent clusters of topics. The LSA algorithm produced statistically significant results, with a difference of 0.04 Silhouette score, compared to the LDA algorithm when utilizing Word2Vec. In the third experiment, the LDA consistently produced identical results and this finding demonstrates that the LDA algorithm can be used for semantic text clustering regardless of word embedding techniques used.

The experiment demonstrates that the LDA algorithm utilizing Word2Vec achieves a more balance distribution of samples clustering (Figure 4). Cluster 0, 1, 2, and 3 contain 2,987, 2,660, 1,320, and 3,032 samples, respectively. Additionally, the visualization clusters the samples effectively, as very little points that represent samples reside on top of one another and confirm by [32] that LDA as an unsupervised machine

learning algorithm is effective for topic modeling and not only produced good results, but also demonstrated an outstanding degree of accuracy in sentiment analysis at both the documents and words dataset.

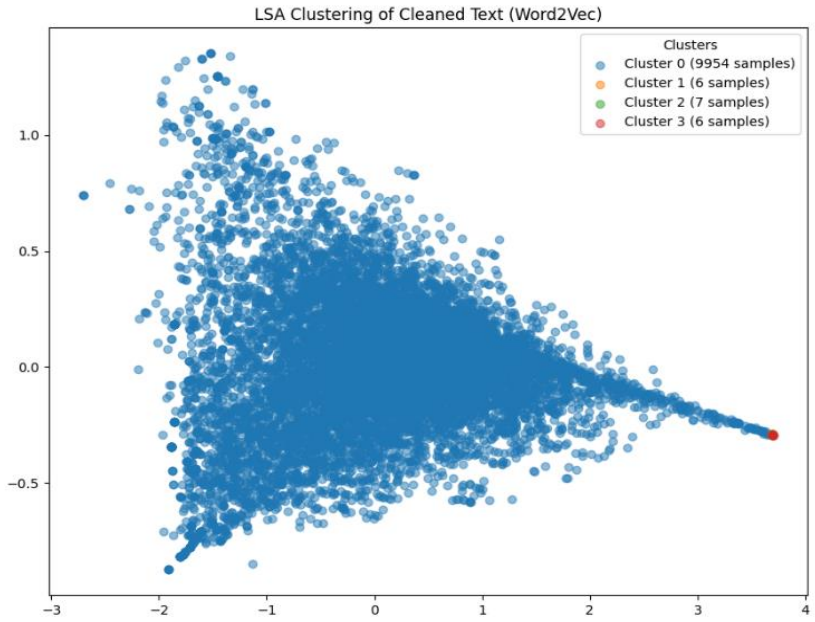

Figure 3. The LSA clustering (Malaysian restaurant reviews)

Table 3. Results for Malaysia telecommunication reviews dataset

Algorithm	Results (Silhouette Score)		
	TF-IDF	Word2Vec	GloVe
K-means	0.01	0.28	0.11
Fuzzy C-means	0.04	0.26	0.05
Non-negative matrix factorization (NMF)	0.01	0.25	0.07
Latent dirichlet allocation (LDA)	0.66	0.66	0.66
Latent semantics analysis (LSA)	0.18	0.62	0.41

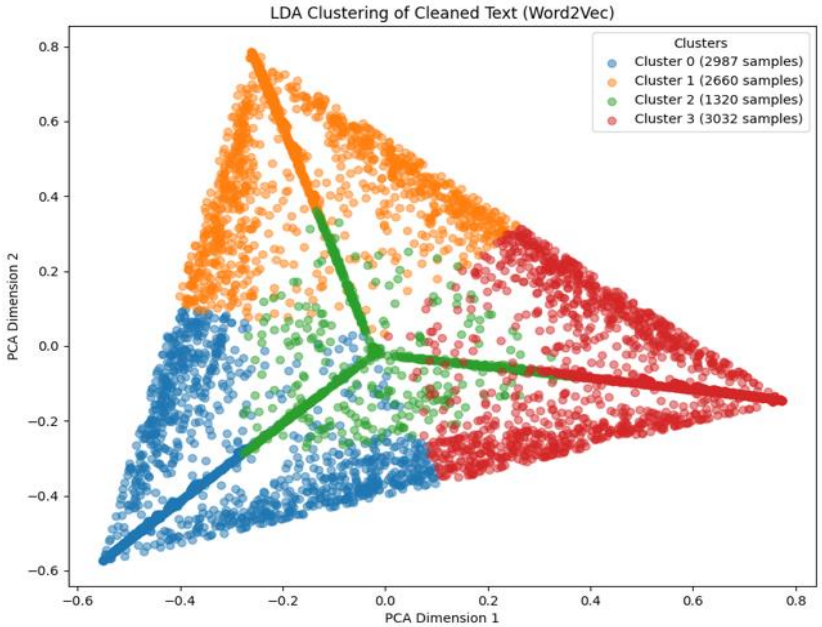

Figure 4. LDA clustering (Malaysia Telecommunication Reviews)

4. CONCLUSION

Nowadays, there is a significant spike in data volume, demanding rapid data processing. This study investigates the techniques of word embedding such as TF-IDF, Word2Vec and GloVe and demonstrates the outcomes of their application in text clustering algorithms. We use K-Means, FCMs, NMF, LDA and LSA to cluster the user reviews. Three distinct user review datasets are used for the experiments: 1) Shopee customer reviews; 2) Malaysia restaurant reviews; and 3) Malaysia telecommunication reviews. In conclusion, with the use of clustering algorithms on the embedded textual data, it is possible to cluster distinct classes or centroids that represent clusters including related texts. The implementation of LDA and LSA algorithms has demonstrated significant efficacy in the clustering of short textual dataset, and there is potential for being used in the context of large textual datasets. from the three experiments, the LSA and LDA algorithms are highly compatible with Word2Vec as an embedding technique for semantic text clustering.

As for future works, it should be considered to use BERT as a word embedding technique as it has its ability to accurately represent both the preceding and succeeding context of a word. This will enable BERT to more effectively capture the intended meaning of a sentence.

ACKNOWLEDGEMENTS

The authors express their sincere appreciation to the reviewers for their valuable feedback, which helped enhance the quality of this article. The funding for this research was provided by 'Skim Zamalah UTeM'of Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.

REFERENCES

[1] N. A. N. M. Idros, H. Mohamed, and R. Jenal, "The use of expert review in component development for customer satisfaction towards E-hailing," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, pp. 347–356, Jan. 2019, doi: 10.11591/ijeecs.v17.i1.pp347-356.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," *Advances in Neural Information Processing Systems*, Oct. 2013, [Online]. Available: http://arxiv.org/abs/1310.4546.

[3] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[5] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00564-9.

[6] K. L. Kouadio, J. Liu, R. Liu, Y. Wang, and W. Liu, "K-Means Featurizer: A booster for intricate datasets," *Earth Science Informatics*, vol. 17, no. 2, pp. 1203–1228, 2024, doi: 10.1007/s12145-024-01236-3.

[7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[8] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent dirichlet allocation: Extracting topics from software engineering data," in *The Art and Science of Analyzing Software Data*, Elsevier, 2015, pp. 139–159.

[9] A. Meddeb and L. Ben Romdhane, "Using topic modeling and word embedding for topic extraction in Twitter," *Procedia Computer Science*, vol. 207, pp. 790–799, 2022, doi: 10.1016/j.procs.2022.09.134.

[10] K. Sofoklis, "Comparing unsupervised learning approaches for topic classification of bank complaints: An NLP study," Utrecht University, 2023.

[11] S. Ounacer, D. Mhamdi, S. Ardchir, A. Daif, and M. Azzouazi, "Customer sentiment analysis in hotel reviews through natural language processing techniques," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, pp. 569–579, 2023, doi: 10.14569/IJACSA.2023.0140162.

[12] X. Li, Y. Guan, B. Fu, and Z. Luo, "Anomaly-aware symmetric non-negative matrix factorization for short text clustering," *Knowledge and Information Systems*, vol. 67, no. 2, pp. 1481–1506, Apr. 2025, doi: 10.1007/s10115-024-02226-z.

[13] J. Chavoshinejad, S. A. Seyedi, F. Akhlaghian Tab, and N. Salahian, "Self-supervised semi-supervised nonnegative matrix factorization for data clustering," *Pattern Recognition*, vol. 137, p. 109282, May 2023, doi: 10.1016/j.patcog.2022.109282.

[14] Ş. Ö. Birim, "Product insights from customer-generated data using topic modeling with BERTopic and sentiment analysis with XLM-T: An experiment on Turkish reviews," *PREPRINT (Version 1) available at Research Square*, pp. 1–30, Feb. 2024, doi: 10.21203/rs.3.rs-3981153/v1.

[15] S. E. Hashemi, F. Gholian-Jouybari, and M. Hajiaghaei-Keshteli, "A fuzzy C-means algorithm for optimizing data clustering," *Expert Systems with Applications*, vol. 227, p. 120377, Oct. 2023, doi: 10.1016/j.eswa.2023.120377.

[16] H. Murfi, Y. J. Agung, S. Nurrohmah, Y. Satria, C. Za'in, and D. Rahayu, "Eigenspace-based fuzzy C-means with large language model BERT for topic detection," Jan. 2023, doi: 10.21203/rs.3.rs-3637575/v1.

[17] R. Astuti, N. Rahaningsih, U. Hayati, C. L. Rohmat, and N. Suarna, "Implementation of fuzzy C-means algorithm with optimized parameter grid for clustering electronic product sales," *East Asian Journal of Multidisciplinary Research*, vol. 2, no. 4, pp. 1647–1660, Apr. 2023, doi: 10.55927/eajmr.v2i4.3929.

[18] N. S. Ayyildiz, A. Akcay, B. Yalcuva, A. Sayar, S. Ertugrul, and T. Cakar, "Segmentation for factoring customers: using unsupervised machine learning algorithms," in *2023 Innovations in Intelligent Systems and Applications Conference, ASYU 2023*, Oct. 2023, pp. 1–7, doi: 10.1109/ASYU58738.2023.10296639.

[19] Ioana, "Latent semantic analysis: Intuition, math, implementation," 2024, [Online]. Available: https://medium.com/data-science/latent-semantic-analysis-intuition-math-implementation-a194aff870f8.
[20] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2–3, pp. 259–284, Jan. 1998, doi: 10.1080/01638539809545028.
[21] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*, vol. 6, no. 1, Oct. 2019, doi: 10.1186/s40537-019-0255-7.
[22] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[23] G. C. Banks, H. M. Woznyj, R. S. Wesslen, and R. L. Ross, "A review of best practice recommendations for text analysis in R (and a user-friendly app)," *Journal of Business and Psychology*, vol. 33, no. 4, pp. 445–459, Jan. 2018, doi: 10.1007/s10869-017-9528-3.
[24] A. Moreo, A. Esuli, and F. Sebastiani, "Word-class embeddings for multiclass text classification," *Data Mining and Knowledge Discovery*, vol. 35, no. 3, pp. 911–963, Feb. 2021, doi: 10.1007/s10618-020-00735-3.
[25] F. Torregrossa, R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier, "A survey on training and evaluation of word embeddings," *International Journal of Data Science and Analytics*, vol. 11, no. 2, pp. 85–103, Feb. 2021, doi: 10.1007/s41060-021-00242-8.
[26] T. Ng, "Shopee text reviews," *Kaggle*, 2020. https://www.kaggle.com/datasets/shymammoth/shopee-reviews (accessed Jan. 08, 2024).
[27] C. K. Ng, "Malaysia restaurant review datasets," *Kaggle*, 2022. https://www.kaggle.com/datasets/choonkhonng/malaysia-restaurant-review-datasets (accessed Jan. 10, 2024).
[28] A. H. A. Mufleh, "Malaysian telecommunication Google Play reviews," *Kaggle*, 2023. https://www.kaggle.com/datasets/ammarhezamahmed/malaysian-telecommunication-google-play-reviews (accessed Jan. 15, 2024).
[29] O. Iparraguirre-Villanueva *et al.*, "Search and classify topics in a corpus of text using the latent dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 246–256, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp246-256.
[30] Muhamad Gatot Supiadin and Arif Dwi Laksito, "Evaluating LDA and LSA for topic modeling in the Indonesian natural disaster," *Indonesian Journal of Computer Science*, vol. 12, no. 6, Dec. 2023, doi: 10.33022/ijcs.v12i6.3478.
[31] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artificial Intelligence in Medicine*, vol. 117, p. 102096, Jul. 2021, doi: 10.1016/j.artmed.2021.102096.
[32] A. Farkhod, A. Abdusalomov, F. Makhmudov, and Y. I. Cho, "Lda-based topic modeling sentiment analysis using topic/document/sentence (Tds) model," *Applied Sciences (Switzerland)*, vol. 11, no. 23, p. 11091, Nov. 2021, doi: 10.3390/app112311091.

BIOGRAPHIES OF AUTHORS

Zuleaizal bin Sidek is the founder and managing director of Kencana Niaga, an IT start-up and independent provider of data analytics for interdisciplinary analysis. He is also a postgraduate student in Philosophy Doctor in Data Science at the Universiti Teknikal Malaysia (UTeM). He is currently working as Data Scientist at the Institut Tun Perak (Melaka state owned company). He has multiple experiences in IT for more than 10 years at the Universiti Teknologi MARA as an IT Manager. His first paper on e-government services: a value of E-service in locbnal government: a fuzzy approach evaluation. His areas of interest are big data management and analytics, machine learning, and blockchain technology. He can be contacted at email: zulsidek@gmail.com.

Assoc. Prof. Dr Sharifah Sakinah Syed Ahmad is currently an associate professor in the Department of Intelligent Computing and Analytics (ICA), Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka (UTeM). She received her bachelor's and master's degrees in applied mathematics from the School of Mathematics at the University of Science, Malaysia. Following this, she received her Ph.D. from the University of Alberta, Canada in 2012 in intelligent systems. She can be contacted at email: sakinah@utem.edu.my.