Unsupervised outlier detection in high-dimensional text data: a comparative analysis

Zuleaizal Sidek[1,2], Sharifah Sakinah Syed Ahmad[1], Noor Hasimah Ibrahim Teo[3]
[1]Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia
[2]Institut Tun Perak, Melaka, Malaysia
[3]School of Computing, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (Melaka), Melaka, Malaysia

Article Info

Article history:

Received Nov 18, 2024
Revised Mar 11, 2025
Accepted May 27, 2025

Keywords:

Anomalies
Isolation forest
Latent dirichlet allocation
Local outlier factor
Outlier detection
User reviews

ABSTRACT

Outlier detection in user reviews is a critical task for identifying anomalous and potentially valuable insights within large datasets. This study presents a comparative analysis of three different algorithms for outlier detection in user reviews: isolation forest, local outlier factor (LOF), and latent dirichlet allocation (LDA). The performance of each algorithm was evaluated using accuracy and silhouette score for outlier detection and clustering quality. LDA performed best with 0.98 accuracy and a silhouette score of 0.13. Isolation forest followed with 0.90 accuracy and a score of 0.11. LOF had lower results with 0.42 accuracy and a score of -0.05 due to its sensitivity to neighbors. The study contributes by systematically exploring the impact of parameter variations on algorithm performance, providing valuable insights for high-dimensional text data analysis. Despite the promising results, limitations include the dependence on preprocessing and specific parameter settings. Future work will explore hybrid approaches and broader datasets to enhance scalability and adaptability.

Corresponding Author:

Zuleaizal Sidek
Department of Intelligent Computing and Analytics
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia
Email: zulsidek@gmail.com

1. INTRODUCTION

Outlier detection involves identifying data points that significantly deviate from the majority of a dataset. These outliers, often referred to as anomalies or abnormalities, can provide critical insights into underlying processes and inform decision-making in various domains, including machine monitoring, financial markets, environmental modeling, and social network analysis [1]. While traditional outlier detection techniques have shown success with numerical data, textual data presents unique challenges due to its high dimensionality, sparsity, and contextual dependencies. This paper addresses these challenges within the context of user review datasets, a growing area of interest fueled by the rise of e-commerce and online review platforms.

User review datasets are crucial for understanding customer experiences, monitoring product or service quality, and detecting fraud or biased content. Anomalies in such datasets can reveal unique consumer sentiments or fraudulent reviews, but detecting these anomalies remains challenging. High-dimensional textual data exacerbates issues of sparsity [2], [3], while contextual ambiguities, such as polysemy (e.g., the word "Apple" referring to either a fruit or a company), hinder robust similarity measures [4], [5]. These difficulties necessitate alternative techniques to improving the accuracy and robustness of outlier detection in textual datasets [6].

Traditional outlier detection techniques such as statistical, distance-based, and clustering-based methods have been extensively applied to numerical data. Statistical approaches for outlier detection include Z-score and Mahala Nobis distance, which identify anomalies based on deviations from the dataset's mean or covariance structure [7]. Additionally, advanced techniques such as centroid embeddings combined with minimum covariance determinant (MCD) offer robust covariance estimation for high-dimensional text data, reducing misclassification of novel inputs [8]. Rare document frequency and ranking methods further enhance statistical outlier detection by addressing data sparsity and distance concentration challenges [9]. Distance-based methods, including K-nearest neighbors (K-NN), rely on distances between data points [10], while clustering techniques like density-based spatial clustering of applications with noise (DBSCAN) and K-means identify outliers as points that do not conform to cluster structures [11]. However, their application to textual data is limited by dimensionality and sparsity issues, requiring significant preprocessing [12].

Efforts to adapt these methods for text data have included similar measures like cosine similarity and Jaccard index. Despite their utility, these methods are sensitive to noise, such as misspellings and slang, which reduces robustness [13], [14]. Machine learning approaches, including latent dirichlet allocation (LDA) and local outlier factor (LOF), have also been explored. LDA models text topics and can highlight unusual documents within a corpus [15], while LOF detects density-based anomalies [16]. While effective, these techniques require extensive tuning to handle the intricacies of text data [17]. Recent advancements in natural language processing (NLP) offer new possibilities. Embedding-based methods, such as those utilizing bidirectional encoder representations from transformers (BERT), provide contextualized representations of text that capture semantic nuances [18]. Combined with clustering algorithms, BERT embeddings can improve outlier detection in user reviews [19]. However, challenges like the high dimensionality of embedding and computational inefficiencies remain, necessitating further research into optimizing these methods.

While there have been significant advancements in outlier detection, a critical gap remains in understanding how variations in parameters impact the performance of detection methods. Addressing this gap is vital for improving the practical application of these techniques. Understanding the effects of parameter variations, such as contamination levels in isolation forest, the number of neighbors in LOF, and topic counts in LDA, is crucial for optimizing detection methods. The experiments conducted in this study focus on LDA, LOF, and isolation forest methods. Specifically, the results highlight the following: different contamination levels for isolation forest, varying numbers of neighbors for LOF, and different topic counts for LDA. These experiments address the gaps in current methodologies by exploring the effects of parameter variations on outlier detection performance.

The remainder of this paper is organized as follows: section 2 provides an overview of the proposed methodology, detailing the preprocessing steps, modeling techniques, and evaluation metrics. Section 3 presents experimental results and section 4 presents the discussion of the findings. Finally, section 5 concludes with a summary of contributions and potential directions for further work.

2. METHOD

In this section we discuss both the datasets used for the initial experiment, in which we select three models to be used for finding outliers in e-commerce user reviews data. We discussed the process through which we arrived at the choice of an outlier detection model. The model contains three main phases. The first phase is dataset pre-processing, where raw text data is cleaned, representation of text data where the preprocessed text is converted into a numerical format that can be used as input to machine learning models. The second phase involves data modelling for three outliers' detection algorithms and the final phase is conducting performance evaluation on the models.

2.1. Dataset preparation

The dataset preparation phase encompasses several essential preprocessing steps to ensure the data is clean, consistent, and suitable for analysis. These steps include the removal of null values, stopwords (commonly used words that do not contribute significant meaning, such as "the," "and," or "is"), punctuation, special characters, and duplicate entries. Additionally, all text is converted to lowercase to maintain uniformity and reduce redundancy caused by case sensitivity.

The dataset used in this experiment was extracted from user reviews on the Shopee e-commerce platform [20] with 10,000 reviews. These reviews represent a valuable source of textual data, offering insights into user experiences and sentiments. After completing the preprocessing steps, the cleaned dataset was transformed into a structured format using the bag-of-words (BoW) model. The BoW approach represents the text data by counting the frequency of individual words or n-grams (sequences of adjacent words) in the dataset. This method converts the raw text into a fixed-length vector, where each dimension corresponds to a unique word or n-gram in the corpus. This process, commonly referred to as feature engineering or feature extraction, is critical for converting unstructured text data into a numerical format that can be processed by machine learning algorithms.

2.2. Outliers detection model

The experiments explore the performance of three outlier detection methods LOF, isolation forest, and LDA on a dataset of user reviews extracted from the Shopee e-commerce platform. The primary objective is to assess how variations in key parameters affect the models' ability to identify anomalies, leveraging the receiver operating characteristic (ROC) metric to evaluate their performance.

The LOF is a density-based algorithm that detects anomalies by measuring the local deviation of a data point’s density relative to its neighbors. LOF computes a local reachability density for each data point and compares it to that of its K-NN. The LOF score is defined as (1):

$$\text{LOF}(p) = \frac{\sum_{i=1}^{k} \frac{1rd_{min(i)}}{1rd_{min(p)}}}{k} \tag{1}$$

where $lrd_{min}(p)$ represents the local reachability density of a point p. The parameter k, the number of neighbors, was varied in this study to understand its impact on anomaly detection performance.

Isolation forest, on the other hand, is a tree-based algorithm that isolates anomalies by randomly selecting a feature and then randomly selecting a split value between the feature's minimum and maximum values. The isolation depth, or the number of splits required to isolate a data point, is smaller for anomalies. The algorithm's effectiveness is determined by the contamination parameter, which represents the proportion of outliers in the dataset. The anomaly score for a data point x is given by (2):

$$\text{Score } (x) = 2 - \frac{E(h(x))}{c(n)} \tag{2}$$

where $E(h(x))$ is the average path length to isolate x, and $c(n)$ is the average path length of a binary search tree built on nnn data points.

LDA, a topic modeling technique, was used to identify anomalous reviews by examining their topic distributions. LDA assumes that each document is a mixture of latent topics, and each topic is characterized by a distribution over words. The generative process in LDA involves the following:

- For each document, a topic distribution θ is drawn from a dirichlet distribution $Dir(\alpha)$.
- For each word in the document, a topic is sampled from θ, and the word is generated from the topic’s word distribution ϕ, also drawn from a dirichlet distribution $Dir(\beta)$.

The number of topics was varied to analyze how granularity affects anomaly detection. Anomalies were identified as reviews with topic distributions that significantly deviated from the dataset's overall patterns.

2.3. Model evaluation

The performance of the models was evaluated using the ROC curve and the silhouette score. The ROC curve measures the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different thresholds, providing insights into the model's ability to detect anomalies relative to normal data points. The silhouette score was used to assess the quality of clustering achieved by the models. This metric evaluates how well each data point is matched to its assigned cluster and how distinct that cluster is from others. The score ranges from -1 to 1, with values closer to 1 indicating better-defined clusters and values closer to -1 suggesting incorrect clustering.

3. RESULTS AND DISCUSSION

This section presents the results and discussions from the comparative analysis of three unsupervised outlier detection algorithms applied to a high-dimensional text dataset of user reviews. The goal is to evaluate the effectiveness of each algorithm in identifying anomalies in the data, considering various performance metrics and their practical implications. We examined the performance of the three algorithms. The evaluation encompasses accuracy, silhouette scores, and ROC curve analysis to provide a comprehensive assessment of each method's strengths and limitations. The discussion is organized into two subsections which are performance analysis and ROC curve analysis.

3.1. Performance analysis result

This evaluation provides a detailed assessment of each algorithm's performance metrics, including accuracy and silhouette scores, to understand how well they detect outliers in text data.

3.1.1. Isolation forest

We evaluated the performance of the isolation forest algorithm for detecting outliers in the Shopee review dataset. The algorithm was tested with various levels of contamination to understand its impact on accuracy and clustering quality, as measured by the silhouette score. The results are summarized in Table 1.

Table 1. The performance of isolation forest algorithm at four contamination levels

Contamination	Accuracy	Silhouette score
0.01	0.98	0.14
0.05	0.94	0.13
0.1	0.89	0.11
0.2	0.79	0.07

Overall, the isolation forest algorithm achieved an accuracy of 0.90 across all contamination levels, with an overall silhouette score of 0.11. With a contamination level of 0.01, the isolation forest algorithm achieved the highest accuracy of 0.98. The silhouette score, which measures the quality of clustering, was 0.14. This indicates that the algorithm was highly effective at identifying outliers with minimal misclassification [21]. The isolation forest algorithm works on the principle that outliers are more susceptible to isolation than normal data points. However, the relatively low silhouette score suggests that while the outliers were correctly identified, the cohesion within the clusters could be improved. At a contamination level of 0.05, the accuracy slightly decreased to 0.94, and the silhouette score was 0.13. This slight drop in performance indicates that increasing the contamination level introduces more noise into the model, making it slightly harder to accurately distinguish outliers from the normal data points [22].

With a contamination level of 0.1, the accuracy further declined to 0.89, and the silhouette score dropped to 0.11. This trend suggests that as more data points are considered as potential outliers, the algorithm's ability to accurately identify true outliers diminishes [23]. The clustering quality, as indicated by the silhouette score, also worsens, reflecting increased difficulty in maintaining clear cluster boundaries. At the highest tested contamination level of 0.2, the accuracy dropped significantly to 0.79, and the silhouette score was 0.07. This substantial decline in both accuracy and silhouette score indicates that a higher contamination level leads to considerable noise, making it challenging for the algorithm to maintain high precision in outlier detection and clear cluster separation.

3.1.2. Local outlier factor

We evaluated the performance of the LOF algorithm for detecting outliers in the Shopee review dataset. The algorithm was tested with various numbers of neighbors to understand its impact on accuracy and clustering quality, as measured by the silhouette score. The results are summarized in Table 2.

Table 2. Performance of LOF with varying number of neighbors

Number of neighbors	Accuracy	Silhouette score
5	0.72	0.02
10	0.53	0.03
20	0.29	-0.05

Overall, the LOF algorithm achieved an accuracy of 0.42 across all numbers of neighbors, with an overall silhouette score of -0.05. With 5 neighbors, the LOF algorithm achieved an accuracy of 0.72 and a silhouette score of 0.02. This indicates moderate success in identifying outliers, though the silhouette score suggests poor clustering quality, implying that the detected outliers do not form well-separated clusters. With 10 neighbors, the accuracy dropped to 0.53, and the silhouette score was 0.03. This decrease indicates that increasing the number of neighbors makes it more challenging to accurately identify outliers and maintain cluster cohesion. At the highest tested number of 20 neighbors, the accuracy dropped significantly to 0.29, and the silhouette score was -0.05. This indicates poor performance in both detecting outliers and clustering quality, highlighting the negative impact of using too many neighbors for this dataset.

3.1.3. Latent dirichlet allocation

We evaluated the performance of the LDA algorithm for detecting outliers in the Shopee review dataset. The algorithm was tested with various numbers of topics to understand its impact on accuracy and clustering quality, as measured by the silhouette score. The results are summarized in Table 3. Overall, the LDA algorithm achieved an accuracy of 0.98 across all numbers of topics, with an overall silhouette score of 0.13. With 5 topics, the LDA algorithm achieved an accuracy of 0.98 and a silhouette score of 0.17. This indicates that the algorithm is highly effective in identifying outliers, with the detected outliers forming well-separated clusters.

Table 3. Performance of latent dirichlet allocation

Number of topics	Accuracy	Silhouette score
5	0.98	0.17
10	0.98	0.14
15	0.98	0.11
20	0.98	0.10

The high silhouette score suggests good clustering quality. At 10 topics, the LDA algorithm maintained its high accuracy of 0.98, with a silhouette score of 0.14. While the clustering quality slightly decreased compared to 5 topics, the algorithm still performed exceptionally well in terms of outlier detection. With 15 topics, the accuracy remained at 0.98, and the silhouette score dropped to 0.11. This further decrease in clustering quality indicates that while the algorithm continues to identify outliers accurately, the separation between clusters becomes less distinct. At the highest tested number of 20 topics, the accuracy was still 0.98, but the silhouette score decreased to 0.10. This trend suggests that increasing the number of topics slightly reduces clustering quality but does not affect the accuracy of outlier detection.

3.2. Receiver operating characteristic curve analysis

In this section, we present the ROC curve analysis for the isolation forest algorithms as depicted in Figure 1, highlighting the true positive and false positive rates across different thresholds. This analysis helps evaluate the trade-offs between detecting outliers and the risk of false positives. We analyze the performance of the isolation forest algorithm using the ROC curve, which provides a graphical representation of the algorithm's diagnostic ability.

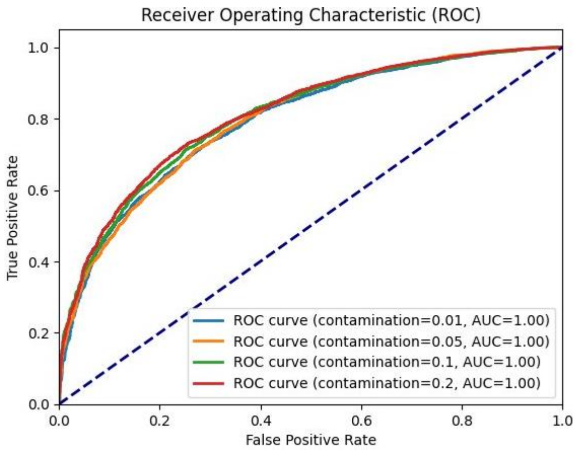

Figure 1. The ROC curve for isolation forest algorithm

The ROC curve plots the TPR against the FPR at various threshold settings. Figure 1 shows the ROC curve for the isolation forest algorithm. Initial performance (Threshold 1): at the lowest threshold, the model does not identify any outliers (TPR=0.0) and does not incorrectly classify any normal points as outliers (FPR=0.0). This indicates a very conservative approach initially. Middle performance (Thresholds 2 to 5): as the threshold increases, the TPR increases, showing that more true outliers are being detected. However, this comes with a corresponding increase in the FPR, indicating more false positives. For instance, at Threshold 2, the TPR is 0.2 while the FPR is 0.6. By Threshold 5, the TPR reaches 0.8 but the

FPR is also high at 0.95. Maximal performance (Threshold 6): at the highest threshold, the model detects all true outliers (TPR=1.0), but nearly all normal points are also classified as outliers (FPR=0.98). This reflects a very aggressive approach, where almost everything is flagged as an outlier.

4. DISCUSSION

The ROC analysis reveals that while the isolation forest algorithm is capable of detecting all true outliers at the highest threshold, it does so at the cost of a very high false positive rate. This indicates that while the algorithm is effective in identifying outliers, it requires careful tuning of the threshold to balance between true positive rate and false positive rate [24]. Choosing an optimal threshold that provides a reasonable balance between TPR and FPR is crucial. For instance, a moderate threshold that offers a TPR of around 0.6 to 0.8 might be preferable, as it allows for a higher detection rate with a relatively lower, though still significant, false positive rate. When comparing the two algorithms, LDA and isolation forest both demonstrated strong performance in detecting outliers in text data, with LDA showing more consistent results across different configurations. Isolation forest, while effective at lower contamination levels, showed decreased performance as contamination increased, indicating the need for careful parameter tuning.

LDA's robustness and consistent performance can be attributed to its ability to model the underlying topic structure of the text data. By identifying the distribution of topics within the documents, LDA can effectively highlight documents that do not conform to the typical topic distributions, making it highly suitable for detecting various forms of outliers, such as reviews with abnormal language or irrelevant content. The stability of LDA's performance across different numbers of topics (5, 10, 15, and 20) with an overall accuracy of 0.98 and a reasonable silhouette score of 0.13 indicates that it can handle a wide range of topic granularity without significant loss of detection capability. Isolation forest's approach to partitioning data points based on random sub-sampling and tree-based isolation is inherently advantageous for identifying outliers in high-dimensional spaces. The high accuracy at lower contamination levels (0.98 with 1% contamination) demonstrates its strength in environments where outliers are sparse. However, as the contamination level increases, the accuracy declines (0.79 at 20% contamination), and the silhouette score also drops, reflecting poorer clustering quality. This suggests that while isolation forest can effectively isolate and identify outliers when they are rare, its performance diminishes as outliers become more prevalent, necessitating careful adjustment of the contamination parameter to maintain a balance between true positives and false positives.

LOF, however, struggled with high-dimensional text data, showing lower accuracy and silhouette scores. Its performance was highly sensitive to the number of neighbors, which poses a challenge for its application in text outlier detection [25]. LOF's local density-based approach can be effective in lower-dimensional spaces where density variations are more pronounced. However, in high-dimensional text data, the concept of local density becomes less meaningful due to the curse of dimensionality, leading to less reliable outlier detection. The significant drop in performance with an increasing number of neighbors (accuracy dropping to 0.29 and a negative silhouette score at 20 neighbors) indicates that LOF is not well-suited for high-dimensional text data, where finding an optimal number of neighbors is inherently challenging.

LDA emerged as the most robust and reliable method for this task, providing high accuracy and reasonable clustering quality across different topic settings. This is likely due to its ability to capture the semantic structure of the text data through topic modeling, allowing for effective identification of reviews that deviate from the normal topic distributions [26]. Its consistent performance makes it a preferable choice for applications involving text outlier detection, such as identifying spam reviews or reviews with off-topic content.

Isolation forest also performed well but requires careful handling of contamination levels to balance true and false positives. Its tree-based approach to isolating data points works well in identifying sparse outliers but becomes less effective as the proportion of outliers increases. This trade-off between true positive rate and false positive rate, highlighted by the ROC analysis, suggests that while isolation forest is powerful, it requires fine-tuning and domain knowledge to optimize its parameters for effective outlier detection in text data.

LOF, while useful in certain contexts, may not be the best choice for high-dimensional text data due to its sensitivity to parameter settings and lower overall performance. The negative silhouette score and declining accuracy with an increasing number of neighbors reflect its difficulty in distinguishing outliers in complex text datasets. This sensitivity to the number of neighbors makes LOF less practical for text outlier detection, where high dimensionality and sparse data points are common [27].

In summary, LDA stands out as the most effective and reliable algorithm for detecting outliers in user reviews, providing consistent high accuracy and meaningful clustering. Isolation forest, while effective in certain scenarios, requires careful parameter tuning to avoid high false positive rates. LOF, due to its sensitivity to parameter settings and lower performance in high-dimensional spaces, is less suited for this task. These findings suggest that for applications involving text data, such as identifying anomalous user reviews, LDA should be the preferred choice, with isolation forest as a secondary option when contamination levels are carefully managed.

5. CONCLUSION

This study compared three unsupervised algorithms—isolation forest, LOF, and LDA—for detecting outliers in high-dimentional text data from Shopee user reviews, focusing on accuracy and silhouette scores. LDA performed best, achieving 0.98 accuracy and a silhouette score of 0.13, demonstrating strong anomaly detection and clustering consistency. Isolation forest followed with 0.90 accuracy and a silhouette score of 0.11 but required careful parameter tuning. LOF showed potential but struggled with parameter sensitivity, achieving 0.42 accuracy and a silhouette score of -0.05. The primary contribution of this research lies in evaluating the impact of parameter variations across these algorithms and providing insights into their effectiveness in high-dimensional textual data. However, the study is limited by its focus on a single dataset and the challenges of parameter optimization, which could impact generalizability. Future research should explore these algorithms on diverse datasets, incorporate advanced text representations like transformer-based embeddings, and consider ensemble methods to enhance performance. Automating parameter tuning and extending the analysis to broader contexts would further refine outlier detection in high-dimensional textual data.

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.

FUNDING INFORMATION

Skim Zamalah UTeM of Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Zuleaizal Sidek	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Sharifah Sakinah Syed Ahmad		✓		✓		✓			✓	✓		✓		
Noor Hasimah Ibrahim Teo	✓				✓					✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

Not applicable.

ETHICAL APPROVAL

Not applicable.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/shymammoth/shopee-reviews.

REFERENCES

[1] A. Anderberg *et al.*, "Dimensionality-aware outlier detection: theoretical and experimental analysis," *arXiv*, 2024, doi: 10.48550/arXiv.2401.05453.

[2] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park, "Outlier detection for text data," in *Proceedings of the 17th SIAM International Conference on Data Mining, SDM 2017*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017, pp. 489–497, doi: 10.1137/1.9781611974973.55.

[3] L. Ruff *et al.*, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, May. 2021, doi: 10.1109/JPROC.2021.3052449.

[4] S. Mukherjee, S. Dutta, and G. Weikum, "Credible review detection with limited information using consistency features," *arXiv*, 2017, doi: 10.48550/arXiv.1705.02668.

[5] M. Chen and A. Prabakaran, "Credibility analysis for online product reviews," *International Journal of Multimedia Data Engineering and Management*, vol. 9, no. 3, pp. 37–54, Jul. 2018, doi: 10.4018/ijmdem.2018070103.

[6] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 38, Nov. 2020, doi: 10.1016/j.cosrev.2020.100306.

[7] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean Journal of Anesthesiology*, vol. 70, no. 4, pp. 407–411, 2017, doi: 10.4097/kjae.2017.70.4.407.

[8] X. Wang, "EAD: effortless anomalies detection, a deep learning-based approach for detecting outliers in English textual data," *PeerJ Computer Science*, vol. 10, Nov. 2024, doi: 10.7717/peerj-cs.2479.

[9] W. A. Mohotti and R. Nayak, "Efficient outlier detection in text corpus using rare frequency and ranking," *ACM Transactions on Knowledge Discovery from Data*, vol. 14, no. 6, pp. 1–30, Dec. 2020, doi: 10.1145/3399712.

[10] A. Rehman and S. B. Belhaouari, "Unsupervised outlier detection in multidimensional data," *Journal of Big Data*, vol. 8, no. 1, p. 80, Dec. 2021, doi: 10.1186/s40537-021-00469-z.

[11] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, Oct. 2004, doi: 10.1023/B:AIRE.0000045502.10941.a9.

[12] K. Bhade, V. Gulalkari, N. Harwani, and S. N. Dhage, "A systematic approach to customer segmentation and buyer targeting for profit maximization," in *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, Jul. 2018, pp. 1–6, doi: 10.1109/ICCCNT.2018.8494019.

[13] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9, pp. 1–17, Aug. 2020, doi: 10.3390/info11090421.

[14] K. A. Sharou, Z. Li, and L. Specia, "Towards a Better Understanding of Noise in Natural Language Processing," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021, pp. 53–62, doi: 10.26615/978-954-452-072-4_007.

[15] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.

[16] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, pp. 1–24, Dec. 2021, doi: 10.3390/bdcc5010001.

[17] R. Bansal, N. Gaur, and S. N. Singh, "Outlier Detection: Applications and techniques in Data Mining," in *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, Jan. 2016, pp. 373–377, doi: 10.1109/CONFLUENCE.2016.7508146.

[18] M. V. Koroteev, "BERT: A Review of Applications in Natural Language Processing and Understanding," *arXiv*, 2021, doi: 10.48550/arXiv.2103.11943.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2018, doi: 10.48550/arXiv.1810.04805.

[20] T. Ng, "Shopee Text Reviews," Kaggle, 2020, [Online]. Available: https://www.kaggle.com/datasets/shymammoth/shopee-reviews.

[21] H. Xiang *et al.*, "OptIForest: optimal isolation forest for anomaly detection," *arXiv*, 2023, doi: 10.48550/arXiv.2306.12703.

[22] D. Cortes-Polo, "Isolation forests: looking beyond tree depth," *arXiv*, 2021, doi: 10.48550/arXiv.2111.11639.

[23] L. Utkin, A. Ageev, and A. Konstantinov, "Improved anomaly detection by using the attention-based isolation forest," *arXiv*, 2022, doi: 10.48550/arXiv.2210.02558.

[24] R. Gao, T. Zhang, S. Sun, and Z. Liu, "Research and improvement of isolation forest in detection of local anomaly points," *Journal of Physics: Conference Series*, vol. 1237, no. 5, pp. 1-7, Jun. 2019, doi: 10.1088/1742-6596/1237/5/052023.

[25] R. C. Ripan *et al.*, "An isolation forest learning based outlier detection approach for effectively classifying cyber anomalies," *arXiv*, 2020, doi: 10.48550/arXiv.2101.03141.

[26] V. Bystrov, V. Naboka, A. S. Bystrova, and P. Winker, "Choosing the number of topics in LDA models-a monte carlo comparison of selection criteria," *arXiv*, 2022, doi: 10.48550/arXiv.2212.14074.

[27] A. Agarwal and N. Gupta, "Comparison of outlier detection techniques for structured data," *arXiv*, 2021, doi: 10.48550/arXiv.2106.08779.

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 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 Information and Communication Technology, 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.

Dr. Noor Hasimah Ibrahim Teo is a senior lecturer at the Universiti Teknologi Teknologi MARA Melaka, Malaysia. She received her Ph.D. in Computer Science from the University of Warwick, United Kingdom in the year 2019 and an M.Sc. in Computer Science from Universiti Teknologi MARA, Malaysia. She received Ernst Mach Grant ASEA-UNINET, OeAD Austria for a postdoctoral visit to the Technological University of Vienna in 2022. Her research interests are concentrated in the field of ontology, text processing, machine learning, and question answering system. She can be contacted at email: shimateo@uitm.edu.my.