ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

IMPROVING SENTIMENT ANALYSIS AND TOPIC EXTRACTION IN INDONESIAN TRAVEL APP REVIEWS THROUGH BERT FINE-TUNING

Oky Ade Irmawan[1], Indra Budi[2], Aris Budi Santoso[3], Prabu Kresna Putra[4]

[1,2,3]Faculty of Computer Science, University of Indonesia, Jakarta, Indonesia
[4]National Research and Innovation Agency, Jakarta, Indonesia

email: oky.ade@ui.ac.id[1], indra@cs.ui.ac.id[2], aris.budi@ui.ac.id[3], prab003@brin.go.id[4]

Abstract

The increasing use of the internet in Indonesia has an influence on the presence of Online Travel Agents (OTA). Through the OTA application, users can book transportation and accommodation tickets more easily and quickly. The increasingly rigorous competition is causing companies like PT XYZ to be able to provide solutions to the needs and problems of their customers in the field of online ticket booking. Many customers submit reviews of the use of the PT XYZ application through Playstore and Appstore, and it needs a technique to group thousands of reviews and detect the topics discussed by customers automatically. In this study, we classified reviews from Android and iOS applications using BERT that had been adjusted through fine-tuning with IndoBERT, as well as modeling topics using LDA to evaluate the coherence score of each sentiment. The result of the comparison of hyperparameter models for the most optimal classification is epoch 4 with a learning rate of 5e-5. The accuracy obtained is 0.91, with an f1-score of 0.74. In addition, testing was carried out to compare BERT with other traditional machine learning. The best performing algorithm was Logistic Regression using TF-IDF word embeddings, achieving an accuracy of 0.890 and an F1-score of 0.865. Therefore, it can be inferred that the accuracy achieved by the fine-tuned classification model of IndoBert is sufficiently high for application in the PT XYZ review classification. Using a coherence score, we found 29 positive topics, 6 neutral topics, and 3 negative topics that were considered the most optimal. This finding can be used as evaluation material for PT XYZ to provide the best service to customers.

Keywords : Online Travel Agent, Sentiment Analysis, Topic Modeling, BERT, LDA

Received: 02-04-2024 | **Revised:** 07-07-2024 | **Accepted:** 10-07-2024
DOI: https://doi.org/10.23887/janapati.v13i2.77028

INTRODUCTION

The increase in internet and smartphone users in Indonesia is changing the way people search for transportation tickets and accommodation. This is one of the drivers of the emergence of various Online Travel Agent (OTA) businesses [1]. PT XYZ is one of the OTAs in Indonesia with its products, such as hotel reservations, airplane tickets, trains, and buses. At the end of 2019, COVID-19 began to spread in Indonesia [2]. The travel industry is a business that has been quite affected during the COVID-19 period, including PT XYZ [3]. Based on data from Google's destinations insights, in 2023 travel demand has begun to bounce back compared to 2022 and 2020 [4], but the number of sales at PT XYZ has not returned to pre-pandemic times. This is still PT XYZ's focus to return the number of transactions at least close to the number of transactions before the pandemic, namely by looking for things that need to be improved.

Some factors that affect customer satisfaction to repurchase are product quality, system functionality, customer service, and return policy [5]. One of the best ways to provide good service is to respond to questions, input, or complaints quickly and professionally [6]. PT XYZ also has Android and iOS applications that get reviews from customers. However, from the large enough review data, there is no automatic management of customer review data. Currently, to find what problems often arise is still using the manual method. The customer review data is reported to the customer service team. By using machine learning techniques, PT XYZ can extract important aspects of opinions expressed from customer reviews which can later increase customer satisfaction and business productivity [7].

Several things have been done in previous studies on sentiment analysis to extract information from application reviews, such as investment applications [8], marketplaces [9], and COVID-19 contact tracing applications [10]. All is trying to find information to improve the app to be better. Sentiment analysis, a discipline within Natural Language Processing (NLP), is employed to recognize and classify opinions, feelings, and

attitudes conveyed through textual content [11]. Both traditional techniques and deep learning can be used to perform sentiment analysis. But nowadays, deep learning generates better performance compared to traditional techniques [12].

Deep learning, a machine learning component, has transformed text classification by autonomously acquiring data representations. Different deep learning techniques have been applied in text classification, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and transformer models [13] [14] [15]. These techniques have proven effective in capturing textual information, learning relationships between words, and handling multiple segments of a series simultaneously. In his study, Chinnalagu conducted an evaluation of various deep learning approaches, including the utilization of Bidirectional Encoder Representations from Transformers (BERT) as a type of transformer model. The findings indicated that this model exhibited superior accuracy and performance metrics [16].

This research aims to facilitate the management of customer reviews in PT XYZ that have not been effective using NLP techniques. Specifically, this research focuses on sentiment analysis to distinguish user sentiment that is expressed in reviews on the Google Play Store and Apple App Store. The importance of this research is in its potential to unlock valuable insights into customer experience and satisfaction and help the company in increasing sales transactions.

In this study, BERT was used as an advanced deep learning model to refine sentiment analysis tasks. Bert's two-way approach allows more stable understanding of the context of user-expressed sentiment, thereby it will increase the accuracy of sentiment classification [17]. In addition, recognizing that customer feedback varies, our study is not only limited to sentiment analysis, but also involves modeling topics using Latent Dirichlet Allocation (LDA). Using LDA, our goal is to discover the underlying topics and themes that commonly appear in reviews, thus providing a thorough understanding of the issues and preferences expressed by users.

Through this research, we aim to revolutionize the process of sentiment analysis and topic modeling in the context of online travel agent application reviews in Indonesia using PT XYZ as a case study. By leveraging BERT and LDA, we wish to automate and improve the extraction of valuable information from customer reviews, thus enabling companies to make more informed decisions. At the end, users will increase their level of satisfaction and experience of using PT XYZ application.

Apps Review

App reviews coming from Google Play Store and Apple App Store platforms have an important role to play in sentiment analysis research [18]. Both platforms provide a forum for users to convey their experiences and opinions about mobile apps. Data from app reviews is a valuable resource that can provide insight into user sentiment towards an app. Both app review platforms provide a platform for users to rate and write reviews related to performance, functionality, and user satisfaction to the application. The review may contain various aspects, such as bug reports, user experiences, and other features [19].

Sentiment Analysis

Sentiment analysis, alternatively called opinion mining, is a subfield of NLP dedicated to identifying, categorizing, and extracting insights into opinions and emotions expressed in textual content. Sentiment analysis theory encompasses several concepts and techniques used to analyze human opinions reflected in texts, including positive, negative, and neutral judgments [20]. There are various methods used in sentiment analysis, such as traditional methods and deep learning [21]. This study applies a deep learning approach utilizing the BERT model, chosen for its outstanding performance in sentiment analysis applications [22] [23].

Bidirectional Encoder Representations from Transformers (BERT)

BERT, an architecture based on transformers, was created by Google's research team in 2018, drawing on the success of transformer architecture in tasks related to Natural Language Processing (NLP) [24]. A standout feature of BERT is its ability to understand text contextually from both directions [17]. In previous models, the context of the text was understood only from left to right (sequentially). However, BERT can process text simultaneously from both directions, which allows BERT to better understanding about word relationships in text [25]. IndoBERT is a BERT variant explicitly trained on an extensive collection of Indonesian texts [26]. It is better designed to capture the nuances and context of Indonesian, making it very effective for NLP tasks involving Indonesian text [27].

Latent Dirichlet Allocation (LDA)

LDA is used to discover hidden topic patterns and themes in data. LDA is a probabilistic model used to model topics in a corpus of text. The model was developed by Davil Ble, Andrew Ng, and Michael Jordan in 2003 [28]. LDA aims to identify topics or themes contained in a collection of documents. The way LDA works in general is model initialization, learning process, allocation of words to topics, parameter estimation, and topic selection [29].

METHOD

The approach employed in this research is illustrated in Figure 1. It begins with the collection of application review data, followed by the assignment of sentiment labels (positive, negative, or neutral). After labeling, the next phase is conducting sentiment analysis with BERT which will be continued to modeling topics with LDA to find out what topics most often arise from review data.

Data Collection

Review data is obtained through PT XYZ's appfollow dashboard. Reviews can also be directly obtained by downloading reviews from the Google Play Store and Apple App Store that formed as csv. The total review data obtained between January 2019 and September 2023 is 27,584 data. From the review data, only reviews that contain text and have 2 or more words are used, and only reviews that use Indonesia language. From this process, review data that can be used in research amounted to 8,144 reviews.

Data Labeling

After the data collection process, 2,000 review data were randomly selected for manual labeling. The labeling was carried out by two annotators, the one who is studying bachelor degree of Informatics Engineering and other worked as Quality Assurance in one of the startup companies in Indonesia. From the labeling results, it was found that there was about 8.3% of the labeling difference between the two annotators, so we tried to adjust the review. The adjustment of 8.3% of the data resulted in a fixed review classification so that it was not biased. The final dataset consists of 2,000 samples categorized into different sentiment classes: 1,359 positive, 116 neutral, and 525 negative. Table 1 provides the number of records in each class along with several examples of data from each sentiment class.

Preprocessing

Preprocessing plays an important role in sentiment analysis, as it can have a significant impact on classification performance [30]. There are several stages that can be done, including text cleaning. Text data needs to be cleaned to remove special characters, punctuation, and other irrelevant elements [8]. Case folding can also be done in preprocessing to convert all letters into a similar form, i.e. all letters are converted into lowercase letters [31]. In text data, some of them contain non-standard words. Therefore, it is necessary to normalize data into standard words in accordance with the Big Indonesian Dictionary (KBBI) [32]. In addition, there is a stemming process to convert words into their basic form which is useful for reducing the variety of words that have the same meaning. However, for some text classification techniques will eliminate this process because it affects accuracy, for example in studies using BERT [9]. Therefore, this study is skipped stemming in preprocessing.

Pre-Train IndoBERT

IndoBERT is a pretrained language model using a large amount of Indonesian data [26]. During the pre-training phase, IndoBERT experiences a comprehensive learning process that enables it to grasp and generate more accurate representations of Indonesian text [34]. The pre-training process involves presenting the model with text in Indonesian from various sources, such as news articles, books, and websites, so that the model can learn complex language patterns [35][36]. Consequently, IndoBERT demonstrates utility across a range of NLP assignments, including hate speech detection [37], question answering [38], and sentiment analysis [8]. By leveraging pre-trained models such as IndoBERT, users can significantly reduce the time and resources needed to train models from scratch while achieving higher-quality representations of Indonesian text [39]. For our specific task, we used the pre-trained IndoBERT model available from Hugging Face (indobenchmark/indobert-base-p1). We fine-tuned this model with our own dataset to adapt it for sentiment analysis.

Hyperparameter Model

Hyperparameters in a BERT model, as in many other machine learning models, are settings that affect how the model learns and operates. Utilizing hyperparameters in BERT models is crucial as it can impact both the effectiveness and efficiency of the model [40]. Some of the important hyperparameters in the BERT model are as follows:

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

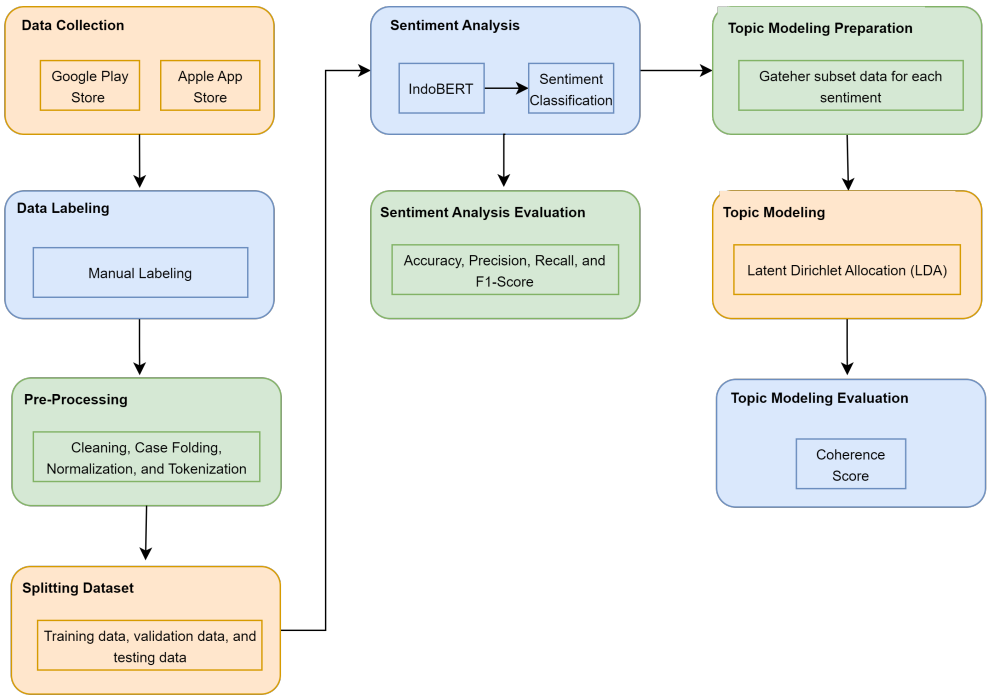

Figure 1. Research Methodology

Table 1. Dataset Specification and Examples

Sentiment	Number of Records	Example Review
Positive	1,359	Great, it helps with booking plane tickets, and the prices are also cheaper
Neutral	116	Just about to try it, hopefully the app is great, so I can keep using it regularly
Negative	525	After selecting the seat and making the payment, it didn't work, such a mess!!!

1. Batch Size
   It refers to the number of data samples provided to the model to be processed at once. Larger batch sizes tend to result in faster learning because they use more data on each iteration. However, batch sizes that are too large require more GPUs and can cause unstable training. Batch sizes that are too small can also result in slow and inefficient training [41]. The batch size used in this study is 32.

2. Epoch
   The quantity of epochs within the hyperparameters dictates the frequency with which the model will observe the entire dataset throughout the training phase. The small number of epochs can lead to insufficiently trained models, while too many epochs can lead to overfitting of training data [42]. We tried to use epochs from 1 to 5 to compare which one was more optimal.

3. Learning Rate
   The learning rate serves as a parameter controlling the quantity of learning iterations conducted by the optimization algorithm when training the model. It dictates the pace at which the model assimilates information from the data and approaches the optimal solution. Learning rates that are too high can lead to unstable learning, while learning rates that are too low can slow convergence. Commonly used learning rate values are 5e-5, 4e-5, 3e-5, and 2e-5 [8].

Confusion Matrix

To assess the effectiveness of the sentiment classification model, a confusion matrix is employed, offering comprehensive insights into

the model's predictive capabilities and ensuring precise outcomes [43]. In this investigation, we will determine the confusion matrix using matrix accuracy, f1 score, precision, and recall, using the following formula:

1. Accuracy
   Accuracy assesses the proportion of accurate predictions relative to the entire dataset, reflecting the model's proficiency in correctly predicting sentiment [43].

$$Accuracy = \frac{TP+TN}{\mathrm{TP+FP+TN+FN}} \quad (1)$$

2. F1-score
   F1-score is a matrix that combines values from precision and recall. It gives a better representation of model performance [43].

$$f1 - score = \frac{2\ x\ precision\ x\ recall}{precision + \mathrm{recall}} \quad (2)$$

3. Precision
   Precision assesses the proportion of accurate predictions among all positive predictions made [43].

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4. Recall
   Recall evaluates the proportion of correct predictions among all data instances that are actually positive [43].

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Coherence Score

Coherence Score evaluation on Latent Dirichlet Allocation (LDA) is one of the methods used to measure the quality of topics generated by the model [44]. LDA is an unguided learning algorithm used to find hidden topics in document collections. The Coherence Score evaluation helps in determining the optimal number of topics to use in the LDA model [45].

Each topic is scored using the Coherence Score metric, which involves calculating how often a pair of words present in the topic appears together in the document. The Coherence Score used in this study is frequency-based. This metric calculates how often related words in a topic appear together in a document, and then measures coherence or alignment between those words [46].

RESULT AND DISCUSSION

In this study, we classify and model the topics based on user reviews of the PT. XYZ from Android and iOS platforms. We conduct various trials to find the best techniques in topic classification and modeling. Here are the results of the research.

Comparison of BERT Hyperparameter Model

In this study, we utilized IndoBERT, a pretrained transformer model specifically designed for Indonesian text. We performed a train-validation split on our dataset, with 70% used for training and 30% for validation. Our experimentation with the Adam optimizer involved testing learning rates of 2e-5, 3e-5, 4e-5, and 5e-5 to determine the most effective setting for optimizing model parameters during training. The training process spanned 5 epochs, during which the model iteratively processed training data batches, calculated loss, and updated weights through backpropagation. Following each epoch, we assessed the model's performance on the validation set to monitor key metrics such as loss, accuracy, and F1-score. This approach ensured that the fine-tuned IndoBERT model was robustly trained and validated for accurate sentiment analysis on Indonesian text.

Based on the results comparing epoch and learning rate, the results are mentioned in Table 2. From the comparisons epoch values from 1 to 5 and learning rates from 2e-5 to 5e-5, it was found that the combination of epoch values 4, learning rate 5e-5, and batch size 32 give the most optimal level of accuracy and f1-score. The accuracy value obtained from the comparison results is between 0.85 to 0.91. This shows that in this case, epoch 4 tuning and 5e-5 learning rate provide the right balance between convergence speed and model accuracy. Thus, the selection of those values can be considered as best practices in the process of training models for a given case.

Comparison with Traditional Machine Learning

To compare the performance of the BERT model with traditional machine learning techniques, experiments were conducted with several classification algorithms, namely Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). Both Bag of Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF) were used for word embedding. Evaluation of performance is conducted utilizing 10-fold cross-validation. The evaluation results obtained the average

value of each matrix which can be seen in Table 3. Logistic Regression with TF-IDF word embedding has the best accuracy value, reaching 0.890. Meanwhile, the highest Bag of Word for word embedding is 0.885 using the Naive Bayes model. From the entire comparison algorithm, the resulting accuracy value is still below the BERT model. But the f1-score obtained is higher than the BERT model. This suggests that the BERT model tends to make more precise predictions overall. But traditional techniques are better at handling the balance between precision and recall, especially if the dataset has unbalanced classes.

Coherence Score Topic Modeling

In this study, coherence values were used as a metric to determine the best number of topics in the LDA (Latent Dirichlet Allocation) model for positive, neutral, and negative sentiments, respectively. Topic coherence measures the degree of alignment between words in a topic, which helps us to understand the interpretability and representation of the topics generated by the model. By analyzing coherence values for different numbers of topics, we can identify points where the increasing number of topics no longer provides a significant increase in coherence. The results show that positive sentiment has the optimal number of topics of 29 with a coherence value of 0.58, where the coherence value had reached its peak. Furthermore, for neutral sentiment, the best number of topics is 6 with a coherence value of 0.58. For negative sentiment, the optimal number of topics is 3 with a coherence value of 0.44. Figure 2 illustrates the coherence scores for all sentiments, with specific charts showing (a) positive sentiment, (b) neutral sentiment, and (c) negative sentiment. These findings provide valuable guidance in setting up LDA models for sentiment analysis and ensuring that the resulting topics optimally reflect the characteristics and complexity of each observed sentiment.

Topic Modeling Visualization

After determining the optimal number of topics based on the coherence score for each sentiment, the next step is to visualize them to identify topics formed from the linkages between keywords. The circles represent the overall frequency of each topic in the corpus; The larger the circle, the more documents in the corpus are related to the topic. Meanwhile, quadrants in the visualization show relationships between various topics. Interrelated topics will be grouped together in one quadrant, while different topics will be spread in different quadrants.

A visualization of the results of topic modeling on positive sentiment can be found in Figure 3, where there is a total of several interrelated topics seen with circles stacked on top of each other. These related topics relate to words that reflect user satisfaction with the app, as well as expectations to maintain its quality in the future. Examples of descriptive analysis results of positive sentiment reviews can be seen in Table 3. Some of the topics of discussion that emerged include lower prices than competitors, ease of use of the application, and various promos and discount vouchers that make prices more competitive. The findings of this positive sentiment analysis can be the basis for PT XYZ to evaluate what has been going well and needs to be maintained.

Meanwhile, for neutral sentiment, there are six topics visualized in Figure 4. Two of the six topics are unrelated to other topics, while the other two topics are related to each other. The words that appear most often in neutral sentiment are 'star', 'multiply', 'promo', 'discount', and 'upgrade'. To explore what topics emerge from the results of this visualization, see Table 4. Some of the topics that came up included discussions about requests to add other payment methods and installments, requests to increase deals and discounts, and requests to reduce taxes on hotel bookings. The findings of these topics in neutral sentiment can be a consideration of whether the wishes of customers can be accommodated by the company, in accordance with the focus to be carried out.

On negative sentiment, there are only three topics illustrated in Figure 5. Some of the most common words are 'refund', 'price', 'promo', 'hotel', 'discount', and 'fund'. In the visualization of topic modeling for negative sentiment, the three topics have no relationship with each other because they are in different quadrants. The themes identified through the outcomes of descriptive analysis are presented in Table 5. Topics that are widely discussed by customers include the refund process that takes a long time, the hotel booking experience that was initially successful but was later informed that the room was full, and problems related to prices that were initially affordable but became expensive during the payment page. These things need to be the focus of improvement by PT XYZ to maintain customer trust and ensure they remain willing to make transactions in the PT XYZ application.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

Table 2. Comparison of Bert Hyperparameter Model

Epoch	Learning Rate	Accuracy	F1-Score	Recall	Precision
1	2e-5	0.89	0.69	0.66	0.86
	3e-5	0.88	0.68	0.68	0.68
	4e-5	0.88	0.70	0.69	0.71
	5e-5	0.90	0.61	0.63	0.59
2	2e-5	0.89	0.68	0.67	0.73
	3e-5	0.89	0.67	0.66	0.73
	4e-5	0.90	0.68	0.66	0.77
	5e-5	0.90	0.67	0.65	0.73
3	2e-5	0.89	0.70	0.67	0.76
	3e-5	0.88	0.71	0.69	0.74
	4e-5	0.85	0.66	0.67	0.66
	5e-5	0.90	0.69	0.67	0.79
4	2e-5	0.89	0.70	0.67	0.76
	3e-5	0.89	0.69	0.67	0.75
	4e-5	0.90	0.63	0.64	0.91
	5e-5	**0.91**	**0.74**	**0.72**	**0.77**
5	2e-5	0.89	0.68	0.66	0.74
	3e-5	0.88	0.71	0.69	0.74
	4e-5	0.90	0.68	0.66	0.75
	5e-5	0.90	0.71	0.69	0.79

Table 3. Comparison BERT with Traditional Machine Learning

Models	Word Embedding	Accuracy	F1-Score	Recall	Precision
BERT		0.910	0.740	0.720	0.770
Random Forest	BoW	0.861	0.848	0.861	0.845
	TF-IDF	0.858	0.838	0.858	0.846
SVM	BoW	0.861	0.836	0.861	0.834
	TF-IDF	0.887	0.863	0.887	0.859
Naïve Bayes	BoW	0.885	0.865	0.885	0.863
	TF-IDF	0.849	0.818	0.849	0.811
Logistic Regression	BoW	0.887	0.872	0.887	0.874
	TF-IDF	**0.890**	**0.865**	**0.890**	**0.857**

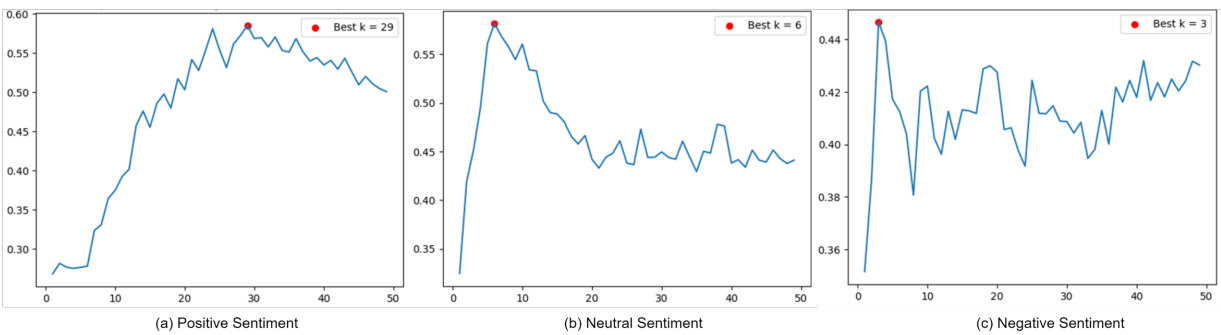

Figure 2. All Sentiment Coherence Score

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

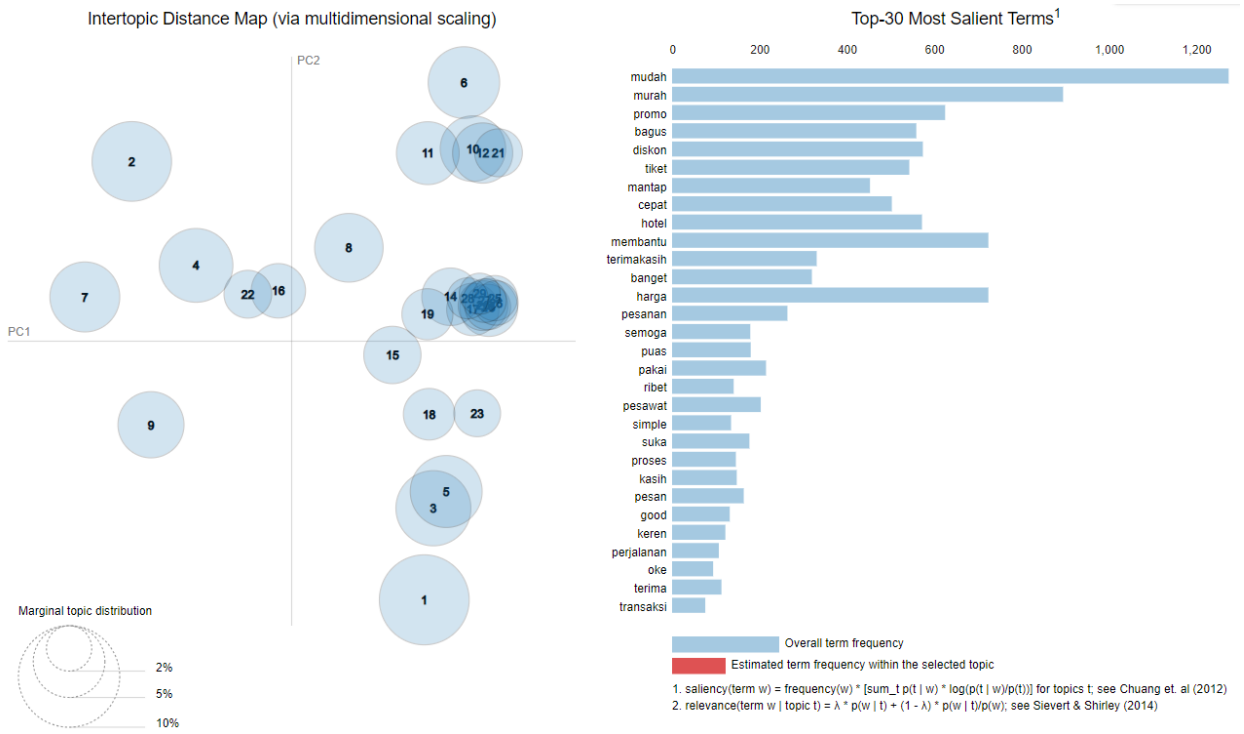

Figure 3. Topic Modeling Visualization of Positive Sentiment

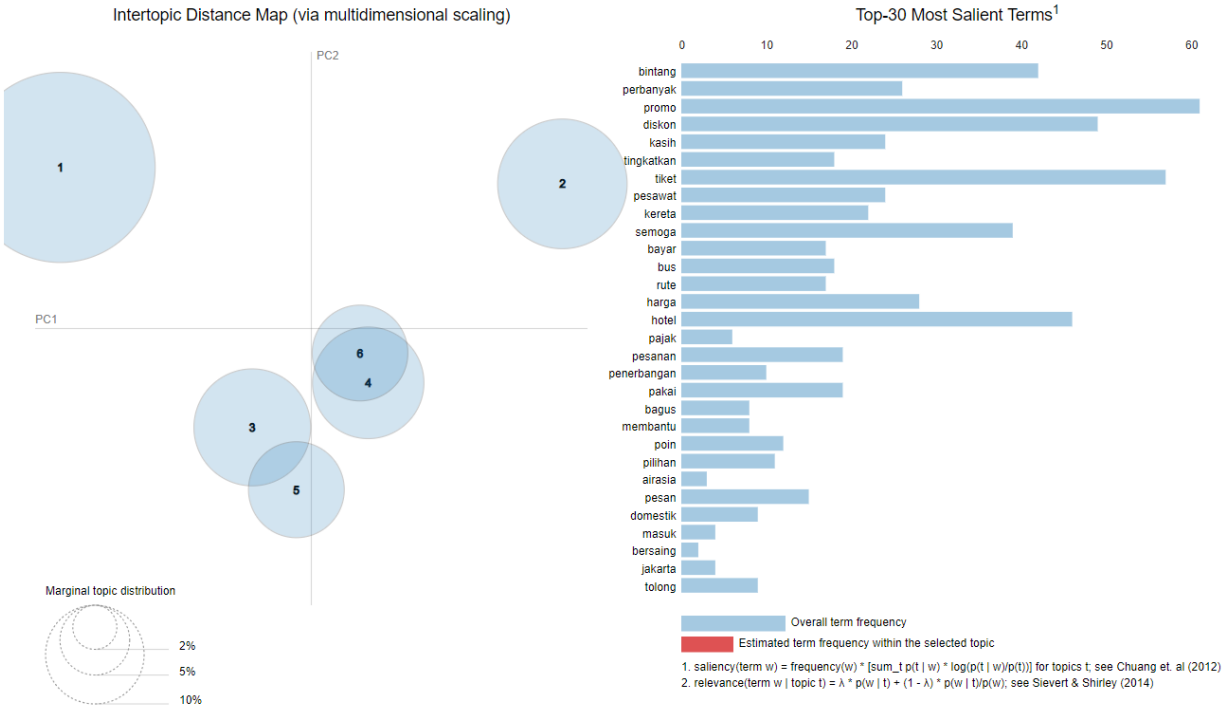

Figure 4. Topic Modeling Visualization of Neutral Sentiment

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

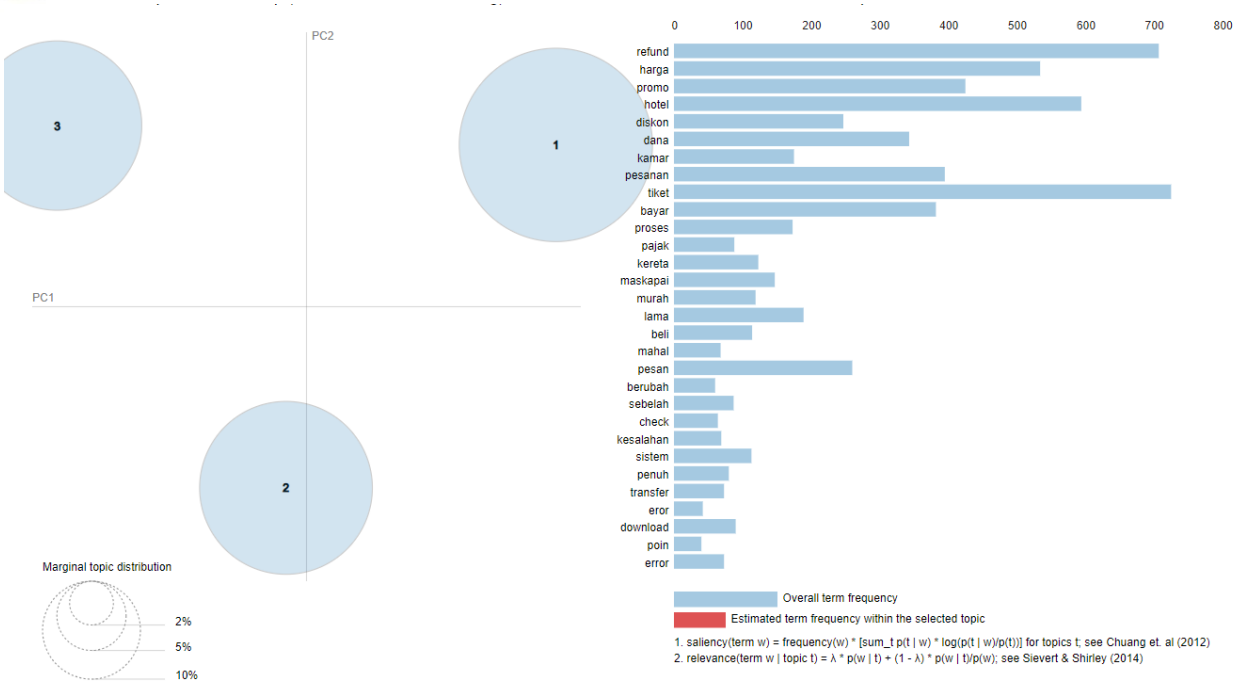

Figure 5. Topic Modeling of Negative Sentiment

Table 4. Descriptive Analysis of Positive Sentiment Topic Modeling

Term	Main Topic
cheap, price, best, compared, displayed	prices shown are cheaper than others
order, easy, hotel, ticket, help, search, train, flight	easy and helpful in booking hotel, train, and flight tickets
discount, price, competitive, recommendation, voucher	promos and voucher discounts make prices competitive

Table 5. Descriptive Analysis of Neutral Sentiment Topic Modeling

Term	Main Topic
upgrade, pay, method, bank, installment	ask if there is no installment payment method, and ask for an additional payment method with another bank
add more, discount, promo, hopefully, please, price	request more promo dan discount
promo, hotel, tax, discount, price, suggestion	increase promo and reduce tax for hotel bookings

Table 6. Descriptive Analysis of Negative Sentiment Topic Modeling

Term	Main Topic
refund, money, ticket, process, long, disappointed, please, confirm, wait, cancellation	the ticket cancellation refund process takes a long time and makes customers disappointed
hotel, room, hours, full, please, response, night, service	the customer was disappointed because after making a hotel reservation, he was informed that the room was full
price, promo, discount, pay, hotel, cheap, tax, expensive, change	the initial price is low but when the payment turns expensive

CONCLUSION

This research focuses on the implementation of the BERT model that has been adjusted through fine-tuning techniques using IndoBERT. In this study, several experiments with some hyperparameters are also carried out, with the aim to optimize model performance. So, this research reveals optimal performance using epoch 4, learning rate 5e-5, and batch size 32. This configuration allows the model to achieve an accuracy level of 0.91 and an f1 score of 0.74. Based on comparisons with traditional machine learning, BERT still has higher accuracy, but a smaller f1-score. Because of the unbalance of the dataset used in this study.

In addition, this study also explores the value of coherence using Latent Dirichlet Allocation (LDA) for each sentiment studied. The results are classified into positive, negative, and neutral sentiment. For positive sentiment, there are 29 cohesive topics. For neutral sentiment, it has 6 topics, and for the negative sentiment, it has 3 topics. These findings provide additional insight to understand the structure and content of each sentiment found in user reviews. Thus, this study not only deepens the understanding of the performance of BERT model in sentiment analysis, but also reveals new insights into the use of LDA to the hidden sentiments in a data set.

For further research, we suggest comparing other deep learning techniques, such as LSTM, RNN and CNN. Comparison with other deep learning techniques aims to better understand the most efficient sentiment analysis modeling techniques. In addition, considering the comparison of preprocessing processes using stemming techniques or without stemming can also be useful research to do in the future research.

REFERENCES

[1] M. I. Rosyidi, “Indonesian Online Travel Agencies: Profiling the services, employment, and users,” vol. 259, no. Isot 2018, pp. 211–216, 2019, doi: 10.2991/isot-18.2019.47.

[2] A. P. Kirana and A. Bhawiyuga, “Coronavirus (COVID-19) Pandemic in Indonesia: Cases Overview and Daily Data Time Series using Naïve Forecast Method,” *Indones. J. Electron. Electromed. Eng. Med. informatics*, vol. 3, no. 1, pp. 1–8, 2021, doi: 10.35882/ijeeemi.v3i1.1.

[3] A. K. Yudha, J. Tang, and N. Leelawat, “COVID-19 Impact on Tourism Business Continuity in Indonesia: A Preliminary Systematic Review,” *J. Disaster Res.*, vol. 17, no. 6, pp. 913–922, 2022, doi: 10.20965/jdr.2022.p0913.

[4] Google, “Country-specific travel demand,” *Destination Insights with Google*, 2023. .

[5] E. Spurer and L. Legentil, “How does customer satisfaction impact the performance of an e-commerce company?,” *Tampere Univ. Appl. Sci.*, 2023, [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/804118/Legentil_Leonie_Spurer_Elisa.pdf?sequence=3.

[6] E. Saprudin and H. Albanna, “The Effect of Service Quality, Personal Selling, and Complaint Handling on Customer Retention of Sharia Bank Customers with Customer Satisfaction as Intervening Variable,” *Bull. Islam. Econ.*, vol. 1, no. 2, pp. 19–33, 2023, doi: 10.14421/bie.2022.012-03.

[7] M. Syamala and N. J. Nalini, “A deep analysis on aspect based sentiment text classification approaches,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 1795–1801, 2019, doi: 10.30534/ijatcse/2019/01852019.

[8] S. Setyani, S. S. Prasetiyowati, and Y. Sibaroni, “Multi Aspect Sentiment Analysis of Mutual Funds Investment App Bibit Using BERT Method,” vol. 9, no. 1, pp. 44–56, 2023.

[9] Syaiful Imron, E. I. Setiawan, Joan Santoso, and Mauridhi Hery Purnomo, “Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 586–591, 2023, doi: 10.29207/resti.v7i3.4751.

[10] K. Ahmad, F. Alam, J. Qadir, B. Qolomany, I. Khan, and T. Khan, “Sentiment Analysis of Users ’ Reviews on COVID-19 Contact Tracing Apps with a Benchmark Dataset,” pp. 1–11.

[11] S. Assem and S. Alansary, “Sentiment Analysis From Subjectivity to (Im) Politeness Detection : Hate Speech From a Socio-Pragmatic Perspective,” *2022 20th Int. Conf. Lang. Eng.*, vol. 20, no. Im, pp. 19–23, 2022, doi: 10.1109/ESOLEC54569.2022.10009298.

[12] P. P. A., “Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis,” *J. Soft Comput. Paradig.*, vol. 3, no. 2, pp. 123–134, 2021, doi: 10.36548/jscp.2021.2.006.

[13] Y. Zhou, “A Review of Text Classification Based on Deep Learning,” *ACM Int. Conf.*

Proceeding Ser., pp. 132–136, 2020, doi: 10.1145/3397056.3397082.

[14] R. Kora and A. Mohammed, "A Comprehensive Review on Transformers Models For Text Classification," *3rd Int. Mobile, Intelligent, Ubiquitous Comput. Conf. MIUCC 2023*, pp. 60–66, 2023, doi: 10.1109/MIUCC58832.2023.10278387.

[15] C. Zhang, "Text Classification Using Deep Learning Methods," in *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, 2022, pp. 1327–1332.

[16] A. Chinnalagu and A. K. Durairaj, "Comparative Analysis of BERT-base Transformers and Deep Learning Sentiment Prediction Models," *Proc. 2022 11th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2022*, pp. 874–879, 2022, doi: 10.1109/SMART55829.2022.10047651.

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[18] S. Gunathilaka and N. De Silva, "Aspect-based Sentiment Analysis on Mobile Application Reviews," *22nd Int. Conf. Adv. ICT Emerg. Reg. ICTer 2022*, pp. 183–188, 2022, doi: 10.1109/ICTer58063.2022.10024070.

[19] E. Noei and K. Lyons, "A survey of utilizing user-reviews posted on google play store," *CASCON 2019 Proc. - Conf. Cent. Adv. Stud. Collab. Res. - Proc. 29th Annu. Int. Conf. Comput. Sci. Softw. Eng.*, no. November, pp. 54–63, 2020.

[20] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.

[21] D. Kansara and V. Sawant, "Comparison of traditional machine learning and deep learning approaches for sentiment analysis," in *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*, 2020, pp. 365–377.

[22] A. Areshey and H. Mathkour, "Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model," *Sensors*, vol. 23, no. 11, 2023, doi: 10.3390/s23115232.

[23] M. Y. A. Salmony and A. R. Faridi, "Bert Distillation to Enhance the Performance of Machine Learning Models for Sentiment Analysis on Movie Review Data," *Proc. 2022 9th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2022*, pp. 400–405, 2022, doi: 10.23919/INDIACom54597.2022.9763262.

[24] A. Veltman, D. W. J. Pulle, and R. W. De Doncker, "The Transformer," *Power Syst.*, no. Nips, pp. 47–82, 2016, doi: 10.1007/978-3-319-29409-4_3.

[25] D. Fimoza, A. Amalia, and T. Henny Febriana Harumy, "Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT," *2021 Int. Conf. Data Sci. Artif. Intell. Bus. Anal. DATABIA 2021 - Proc.*, pp. 27–34, 2021, doi: 10.1109/DATABIA53375.2021.9650096.

[26] E. Fernandez, Anderies, M. G. Winata, F. H. Fasya, and A. A. S. Gunawan, "Improving IndoBERT for Sentiment Analysis on Indonesian Stock Trader Slang Language," *Proc. 2022 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2022*, pp. 240–244, 2022, doi: 10.1109/IoTaIS56727.2022.9975975.

[27] D. Sebastian, H. D. Purnomo, and I. Sembiring, "BERT for Natural Language Processing in Bahasa Indonesia," *2022 2nd Int. Conf. Intell. Cybern. Technol. Appl. ICICyTA 2022*, pp. 204–209, 2022, doi: 10.1109/ICICyTA57421.2022.10038230.

[28] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," *Art Sci. Anal. Softw. Data*, vol. 3, pp. 139–159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, doi: 10.1016/B978-0-12-411519-4.00006-9.

[30] M. Işik and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1405–1421, 2020, doi: 10.3906/elk-1907-46.

[31] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874,

no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[32] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.

[33] B. T. Hung, "Domain-specific versus general-purpose word representations in sentiment analysis for deep learning models," in *Frontiers in Intelligent Computing: Theory and Applications: Proceedings of the 7th International Conference on FICTA (2018), Volume 1*, 2020, pp. 252–264.

[34] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[35] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020, [Online]. Available: http://arxiv.org/abs/2009.05387.

[36] S. M. Isa, G. Nico, and M. Permana, "Indobert for Indonesian Fake News Detection," *ICIC Express Lett.*, vol. 16, no. 3, pp. 289–297, 2022, doi: 10.24507/icicel.16.03.289.

[37] A. Marpaung, R. Rismala, and H. Nurrahmi, "Hate Speech Detection in Indonesian Twitter Texts using Bidirectional Gated Recurrent Unit," *KST 2021 - 2021 13th Int. Conf. Knowl. Smart Technol.*, pp. 186–190, 2021, doi: 10.1109/KST51265.2021.9415760.

[38] B. K. Jha, C. M. V. Srinivas Akana, and R. Anand, "Question Answering System with Indic multilingual-BERT," *Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021*, no. Iccmc, pp. 1631–1638, 2021, doi: 10.1109/ICCMC51019.2021.9418387.

[39] Y. Pan *et al.*, "Reusing Pretrained Models by Multi-linear Operators for Efficient Training," no. NeurIPS, 2023, [Online]. Available: http://arxiv.org/abs/2310.10699.

[40] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[41] Devansh, "How does Batch Size impact your model learning," *Medium*, 2022. https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa (accessed Mar. 28, 2024).

[42] A. Komatsuzaki, "One epoch is all you need," *arXiv Prepr. arXiv1906.06669*, 2019.

[43] M. Fahmy Amin, "Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial," *J. Eng. Res.*, vol. 6, no. 5, pp. 0–0, 2022, doi: 10.21608/erjeng.2022.274526.

[44] S. Syed and M. Spruit, "Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation," *Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017*, vol. 2018-Janua, pp. 165–174, 2017, doi: 10.1109/DSAA.2017.61.

[45] W. Zhao *et al.*, "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC Bioinformatics*, vol. 16, no. 13, p. S8, 2015, doi: 10.1186/1471-2105-16-S13-S8.

[46] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," *WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min.*, pp. 399–408, 2015, doi: 10.1145/2684822.2685324.