Ingénierie des Systèmes d'Information
Vol. 31, No. 4, April, 2026, pp. 1333-1342
Journal homepage: http://iieta.org/journals/isi

Does Topic-Based Aspect Labeling Improve Sentiment Analysis? A Comparative Study of LDA-Guided Segmentation with BERT and TF-IDF

Kathleen Felicia Annabel, Hapnes Toba*, Swat Lie Liliawati

Faculty of Smart Technology and Engineering, Maranatha Christian University, Bandung 40164, Indonesia

Corresponding Author Email: hapnestoba@it.maranatha.edu

https://doi.org/10.18280/isi.310427

Received: 27 August 2025
Revised: 1 December 2025
Accepted: 20 March 2026
Available online: 30 April 2026

Keywords:
sentiment analysis, automatic aspect labeling, term frequency-inverse document frequency weighting, bidirectional encoder representations from transformers weighting, topic modeling, Latent Dirichlet Allocation

ABSTRACT

Aspect-based sentiment analysis (ABSA) enables fine-grained opinion understanding by associating sentiments with specific text aspects. However, manual aspect annotation remains a bottleneck for large-scale applications. This study investigates whether automatic aspect labeling using Latent Dirichlet Allocation (LDA) topic modeling can effectively support downstream sentiment classification without sacrificing accuracy. We used a product review dataset containing 26,873 labeled instances. Aspect labels were assigned automatically by selecting the dominant topic for each review based on LDA topic probabilities. We then trained and evaluated two sentiment classifiers, bidirectional encoder representations from transformers (BERT) and term frequency-inverse document frequency (TF-IDF) with logistic regression, under two conditions. The first was a baseline setting without topic segmentation. The second involved training separate models on topic-specific subsets created from the automatic aspect labels. Automatic annotation achieved 66% agreement with manual labeling. This result suggests moderate reliability but also reveals substantial ambiguity in the aspect assignments. Contrary to expectations, topic-based segmentation did not improve classification performance. The baseline BERT model achieved the highest performance, with an F1-score of 0.83 and an accuracy of 0.94. It outperformed all topic-specific BERT models, which reached an average F1-score of 0.79, as well as all TF-IDF-based variants. Student's t-tests revealed no statistically significant differences between segmented and non-segmented models ($p > 0.05$ for all comparisons). Error analysis identified three main sources of misclassification: keyword ambiguity, mixed-sentiment expressions, and topic overlap. We conclude that LDA offers a viable path toward automatic aspect labeling. However, topic-based segmentation does not automatically translate into performance gains in sentiment classification. Future work should explore more precise annotation methods and domain-adaptive topic modeling techniques.

1. INTRODUCTION

Sentiment analysis is a process for examining opinions, thoughts, and feelings expressed by individuals. It can be applied to a wide range of topics, products, subjects, and services. It helps gather information and support decision-making based on these opinions. Using this technique, various data sources, such as marketplace reviews, can be analyzed to determine whether sentiment is positive or negative. These results help researchers understand consumers' overall perceptions of the reviewed products.

Several studies have been conducted to develop methods for sentiment analysis. Recent research has mainly focused on sentiment analysis across different languages and approaches [1, 2]. Jurek et al. [3] proposed a lexicon-based sentiment analysis algorithm that utilizes sentiment normalization and evidence-based combination functions. Manek et al. [4] introduced a feature-extraction technique based on the Gini index and a support vector machine (SVM) for classification. Meanwhile, Chen et al. [5] proposed an LSTM-based model for more detailed, emotion-aware sentiment analysis of product reviews in Mandarin.

In traditional sentiment analysis, the polarity of each term in a document is determined independently of the document's domain [6]. In the context of product reviews, sentiment analysis can be trained to identify and analyze relevant information based on input keywords, which helps determine consumer opinions about a product [7]. However, several challenges remain in sentiment analysis. These include the difficulty of developing robust techniques for heterogeneous big data, the frequent uncertainty or incompleteness of available data, and the need to analyze semantic relationships across multiple data sources. Further exploration is often necessary to understand additional factors that may interact with sentiment. Accounting for these factors enables a more accurate interpretation of public opinion [8]. The limitations of traditional approaches lie in two areas: a lack of understanding of emotions in the presence of negation and a heavy dependence on surface-level features [9].

To address these limitations, aspect-based sentiment

analysis (ABSA) has gained growing interest. According to Nazir et al. [10], ABSA provides a clearer understanding of the challenges in sentiment analysis than traditional methods. It focuses directly on sentiment instead of relying primarily on language structures. In ABSA, each aspect is connected to an entity, and the concept of aspects extends beyond judgments to encompass thoughts, perspectives, underlying themes, or social influences related to a specific event. Therefore, ABSA offers a promising approach for analyzing sentiment over time across various types of media content.

This study adopts an ABSA approach. Unlike typical sentiment analysis, this study's approach evaluates texts by pinpointing sentiments associated with specific aspects. For example, in the sentence "The quality of this shirt is good, but the price is expensive," ABSA would analyze two aspects: the shirt's quality and its price. Sentiment toward shirt quality is positive, whereas sentiment toward price is negative [11].

To perform ABSA, aspect labels need to be assigned to each document. One way to identify aspect labels in a dataset is to use topic modeling with Latent Dirichlet Allocation (LDA). Previous studies have shown that topic modeling, especially LDA-based techniques, is effective for extracting aspects and grouping them into coherent clusters during sentiment analysis tasks [12].

There are several methods for labeling datasets: manual, semi-automatic, and fully automatic. Based on existing studies, manual annotation remains widely used because of its perceived accuracy. However, when performed manually and on a per-document basis, annotation can be time-consuming, especially for large datasets. Therefore, this study utilizes automatic aspect labeling to facilitate more efficient aspect-based sentiment classification.

The ABSA approach in our study employs a transformer-based model, specifically the bidirectional encoder representations from transformers (BERT), which improves sentiment classification accuracy by capturing contextual information and sequential relationships among words within a sentence. Recent research has demonstrated that deep learning methods, such as BERT, provide strong contextual representations that can enhance ABSA performance when combined with task-specific adaptations [13].

In this study, the term "aspect" refers to coarse-grained product or content categories inferred at the document level, such as movie, music, or cleaning product. This differs from traditional token-level ABSA, which focuses on fine-grained attributes such as "battery" or "price". As such, the task more closely resembles category-level aspect sentiment analysis, which has distinct methodological implications compared with traditional fine-grained ABSA.

However, previous studies have rarely employed topic modeling techniques, such as LDA, to automatically identify aspects in large, diverse datasets. Our approach in this study uses BERT as the sentiment classification model and compares its performance with that of a traditional algorithm, Logistic Regression with term frequency-inverse document frequency (TF-IDF). Additionally, aspect annotation is performed automatically using LDA topic modeling, in which the most probable topic is assigned as the aspect label for each document in the dataset. By combining BERT and LDA, this study aims to develop a more effective approach to ASBA, particularly for analyzing product reviews from online marketplaces.

While previous studies have explored the use of topic modeling for aspect extraction, this study does not claim methodological novelty in employing LDA. Instead, it focuses on critically evaluating its effectiveness and limitations when used as a fully automatic labeling mechanism for ABSA. Unlike prior studies that assume topic-model-based aspect labeling improves ABSA performance, this study systematically investigates whether such improvement holds in a large, heterogeneous, real-world review dataset.

Rather than proposing a new algorithm, this work makes three main contributions. First, it empirically evaluates the reliability of LDA-based automatic aspect labeling through manual validation. Second, it quantitatively analyzes the impact of topic-based data segmentation on downstream sentiment classification using both classical and transformer-based models. Third, it provides an error-driven diagnostic analysis that explains why topic-model-based aspect labeling may fail to improve ABSA performance under realistic conditions. The findings offer practical guidance on when topic modeling is—and is not—appropriate for aspect-oriented sentiment analysis.

In line with the proposed contribution, we define the following research questions:

1. To what extent can LDA-based topic modeling provide reliable automatic aspect labels when compared with manual annotation?
2. Does topic-based data segmentation derived from automatic labeling improve, degrade, or have no significant effect on sentiment classification performance?
3. How do transformer-based models, such as BERT, compare with classical TF-IDF-based models under both segmented and non-segmented settings?
4. What types of errors and data characteristics explain the observed performance outcomes?

2. METHODS

Figure 1 shows the overall workflow used in this study. The collected dataset was directly to build a topic model using the LDA method. The topics created from this process were then automatically labeled and integrated into each review in the dataset.

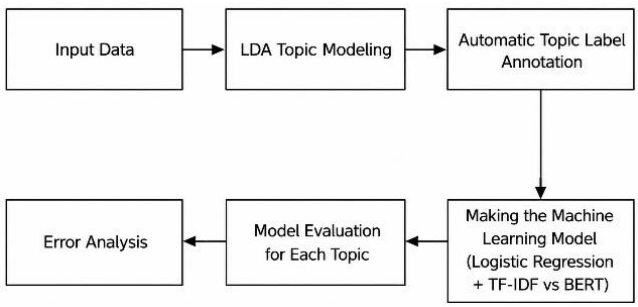

Figure 1. Research flow diagram

After annotation, the dataset was segmented according to the assigned topic labels. Then, training and evaluation of the classification models were conducted on each topic-specific subset using two methods: a Logistic Regression model with TF-IDF feature representation and the BERT model. For comparison, both models were also trained and evaluated on the entire dataset without topic segmentation, serving as a baseline. The baseline evaluation metrics were then compared

with the topic-specific models' results to assess the effectiveness of the topic-based approach.

Finally, an error analysis was performed to identify and understand patterns in the misclassified data. This analysis aimed to identify characteristics or commonalities in model errors that might not be apparent from a quantitative evaluation alone.

2.1 Dataset

The dataset used in this study consists of product reviews labeled as either positive or negative. The dataset was obtained from the public GitHub repository of aakashgoel12, located in the blogs repository. This dataset contains 26,873 rows with two columns. The Review column contains user reviews of various products in English. The User_sentiment column contains sentiment labels classified into two categories: positive (1) and negative (0).

The preprocessing includes removing punctuation and stopwords, converting to lowercase, lemmatization, and removing numbers and extra whitespace. This processed dataset served as the basis for automatic aspect-label annotation through topic modeling. It was also used to train and evaluate the baseline classification models. Figure 2 shows a random sample of five product reviews and their associated sentiment labels.

	Review	user_sentiment
22480	great last attempt godzilla pathetic didnt exp...	1
16053	black rice best rice yet extremely tasty sligh...	0
24478	cold tastic amaxing product	0
26218	leaf hair one amazing conditioner price point ...	1
4730	great purchase although include d version las...	1

Figure 2. Sample of product review and user sentiment

2.2 Topic modeling and automatic annotation

Topic modeling is a technique for uncovering hidden topics in large document collections [14]. In this study, we employed LDA, one of the most widely used statistical algorithms for topic modeling [15]. Several steps were taken to implement LDA-based topic modeling. First, bigrams and trigrams were constructed to capture word pairs and triplets that frequently co-occur in reviews. These phrases were added to the token list for each document using Gensim.

Next, a dictionary was created from the collected tokens, assigning each word a unique ID. Words that appeared too rarely (in fewer than five documents) or too frequently (in more than 20% of documents) were removed to improve the quality of the representation. Then, a corpus was created from the dictionary using the bag-of-words representation with the doc2bow function, which produces (word_id, frequency) pairs for each document. A total of 5,848 unique tokens were identified in the dataset, each assigned a unique ID and counted by frequency.

Later, a TF-IDF model was employed to assign higher weights to important words that appear less frequently across documents. To determine the optimal number of topics, the coherence score was calculated for various numbers of topics. This score measures the semantic similarity of words within each topic, with higher scores indicating greater semantic coherence. Based on these results, an LDA model was built using the optimal number of topics.

Each resulting topic includes a set of key terms that are highly relevant to that topic. To facilitate interpretation and analysis, the topic modeling results were visualized interactively using pyLDAvis. This tool displays topic distributions, inter-topic distances, and the most important keywords in an intuitive format.

Each preprocessed document, represented using TF-IDF weights, was then assigned topic probability scores by the LDA model. For each document, the topic with the highest probability score was chosen as its dominant topic label. This assignment used the get_document_topics() function from Gensim's LDA model. The function returns a list of (topic_id, probability) pairs for each document, and the topic with the highest probability is selected as the dominant label. The topic with the highest probability in this list was selected as the main topic label.

The topic labels were then added as a new column to the dataset, enabling each review to be automatically annotated with its dominant topic based on the LDA model's topic distribution. This process was automated and used to divide the dataset into topic-specific subsets for training and evaluating topic-specific classification models.

To assess the reliability of the automatic annotation, a validation process was conducted by comparing its results with those of manual annotation. A total of 100 samples were randomly chosen from the dataset. The researcher manually labeled these samples using the same aspect categories as those used in the automatic annotation. The manually assigned labels were then compared to the automatically assigned labels to determine accuracy. This accuracy score indicates how well the automatic annotation aligns with the manually assigned labels.

It is important to note that the proposed pipeline in Figure 1 introduces a two-stage dependency, *i.e.*, topic identification followed by sentiment classification. Errors in the first stage (automatic aspect labeling) may propagate to the second stage, potentially degrading overall performance. Therefore, this study explicitly evaluates not only classification accuracy but also annotation reliability and error patterns, acknowledging that topic-based segmentation may not necessarily show performance gains.

2.3 Machine learning models

The machine learning models used for sentiment analysis in this study were Logistic Regression with TF-IDF and BERT. Initially, all documents in the dataset were used as input to train both models as a baseline comparison. The data were split into training and testing sets using a 70:30 ratio. The Review column served as the input feature, while user_sentiment was the label to be predicted. However, because the sentiment label distribution was imbalanced (88% positive and 12% negative), the dataset was balanced via oversampling.

The first model utilized logistic regression with TF-IDF to convert reviews into numerical data. TF-IDF assigns weights to each word in a document, helping the model determine the importance of each term within the document's context. The numerical representation of reviews was then used as input to the logistic regression algorithm. Since the sentiment labels were binary (positive or negative), the model estimated the

probability that each review belonged to one of the two sentiment categories.

The second model used fine-tuning of the BERT architecture for sentiment classification. The data were first converted to the Dataset format provided by the Hugging Face library to enable efficient processing during transformer-based model training. The reviews were then tokenized into sequences of tokens or numerical IDs that BERT can interpret, using the bert-base-uncased tokenizer. The BERT model was configured with TrainingArguments to define hyperparameters, such as the number of epochs and batch size. The Trainer object from Hugging Face was then used to fine-tune the BERT model by combining it with training arguments and evaluation datasets.

The same training process was then repeated for each topic identified through LDA-based topic modeling. As a result, multiple classification models were created, using both Logistic Regression and BERT, with each model trained specifically on review data segmented by topic.

2.4 Model evaluation

After training all models, the models were evaluated using the F1 score to assess classification performance on the dataset. Several evaluation metrics are used in this study [16]:

- Precision

Precision measures how well the model correctly predicts positive cases. It indicates the percentage of true positives among all predictions labeled as. Eq. (1) presents the formula for calculating precision. Precision is calculated as the number of True Positives (TP) divided by the total number of positive predictions (TP + False Positives (FP)). True Positives are elements that the model correctly identifies as positive. False Positives are elements that the model predicts as positive even though they are actually negative.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- Recall

Recall measures how well the model finds all positive examples in the dataset. It is the ratio of true positive instances to the total number of actual positive instances. Eq. (2) provides the formula for recall. Recall is calculated as the number of true positives divided by the total number of positive instances. False Negatives are elements that the model predicts as negative even though they are actually positive.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- Accuracy

Accuracy measures the overall correctness of the model's predictions on the entire dataset. Eq. (3) presents the formula for calculating accuracy. It is computed by adding the True Positives (TP) and True Negatives (TN), then dividing the sum by the total number of predictions. These values correspond to the entries in the confusion matrix (TP, TN, FP, and False Negatives (FN)).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- F1-Score

The F1-score is a metric used to evaluate classification performance by combining Precision and Recall into a single value through the harmonic mean. The F1-score provides a balanced measure between Precision and Recall. Eq. (4) presents the formula for calculating the F1-score. It reaches a maximum value of 1 when both Precision and Recall are perfect, and drops to 0 if either Precision or Recall is zero.

$$F1 - score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{4}$$

In addition to the F1-score, this study also evaluated whether the performance differences between models are statistically significant. This evaluation was conducted using the Student's t-test, a statistical method for hypothesis testing that compares the means of two groups [17]. Specifically, we applied a two-sample equal-variance (homoscedastic) t-test. This test compares the means of two groups whose data are assumed to come from different subjects [18].

While a student's t-test is used to assess performance differences, it has limitations when applied to binary correctness outcomes. Therefore, the statistical results are interpreted conservatively and are used only as supplementary evidence. The primary conclusions of this study are based on consistent trends observed in F1-scores across models rather than on statistical significance testing alone.

The data used in this test consisted of prediction outcomes, with each instance marked as 1 if correctly predicted by the model and 0 if incorrectly predicted. A comparison was made between the baseline and aspect-based models, using both BERT and Logistic Regression with TF-IDF, on 300 randomly selected samples from each model.

The purpose of this test was to determine whether the performance difference between models is statistically significant. If the resulting p-value was less than 0.05, the test was considered significant, indicating a meaningful difference in performance. Conversely, if the p-value was greater than or equal to 0.05, the test was not statistically significant. In this case, there was insufficient evidence to conclude that the performance difference between the models is statistically meaningful.

2.5 Error analysis

After evaluating model performance using metrics such as accuracy, precision, recall, and F1-score, this study also conducted an error analysis to gain a deeper understanding of the model's prediction errors. The analysis focused on two main error types: FP and FN. An FP occurs when the model predicts a positive sentiment, but the true label is negative. Conversely, an FN occurs when the model predicts a negative sentiment, but the true label is positive [19].

Error analysis was performed on all developed models using 30% of the dataset as test data. The variable y_test, which contains the actual sentiment labels from the dataset, served as the ground truth, whereas the model's predictions after training were used as the predicted labels. These two variables were compared, and any mismatch was classified as a misclassification. This process identified misclassified documents and enabled further examination of words or phrases that were likely to cause misclassification in the sentiment prediction process.

3. RESULTS AND DISCUSSION

3.1 Topic modeling results with the visualization

The number of topics used in this study was determined based on the coherence score. Figure 3 shows the coherence scores for different numbers of topics, with the highest score obtained for four topics. Therefore, this study chose four topics as the optimal number for the subsequent topic modeling process. Although the coherence score was used as the primary criterion for selecting the number of topics, coherence alone does not guarantee interpretability or aspect purity.

The coherence curve showed blurred topic boundaries, and manual inspection revealed overlapping and mixed-topic keywords. This suggests that relying solely on coherence optimization may be insufficient for determining aspect-representative topics in multi-domain review datasets.

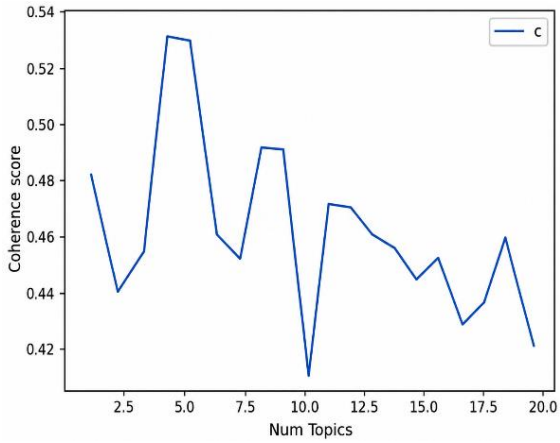

Figure 3. Coherence score visualization based on topic numbers

A visualization was then created using pyLDAvis to aid in understanding the topic distribution. Figure 4 shows the pyLDAvis visualization for the four-topic LDA model. The size of each circle indicates the number of documents linked to that topic.

Topic 2 is the largest, with 8,973 documents, followed by Topic 1 with 8,087, Topic 4 with 5,081, and Topic 3 with 4,732. Additionally, the distance between topics reflects their similarity, with closer topics sharing more similar keywords.

To facilitate identification, a descriptive name was assigned based on the most common words, each representing a specific aspect of the topic. The aspects from the four topics are: cleaning products, music, hair products, and movies. Table 1 presents the most frequent words for each topic and the corresponding aspect labels.

To compare and identify the optimal number of topics, a five-topic model was constructed and visualized using pyLDAvis. The reason for selecting five topics was that they achieved the second-highest coherence score among the models. Figure 5 shows the pyLDAvis visualization for the five-topic LDA model.

The pyLDAvis visualization indicates that circles 1 and 5 share similar keywords, such as good, godzilla, and watch, suggesting overlap between these two topics. The smaller circle (Topic 5) could be merged with circle 1 to create a single, clearer topic. As a result, this study used four topics, each with more focused and representative keywords, providing greater semantic clarity.

The presence of semantically unrelated terms within several topics (e.g., food-related terms within the "music" topic) indicates that LDA captured co-occurrence patterns rather than semantically coherent aspects. This ambiguity directly affected the quality of automatic labeling and provides empirical evidence that topic purity is a critical bottleneck when LDA is used for aspect identification.

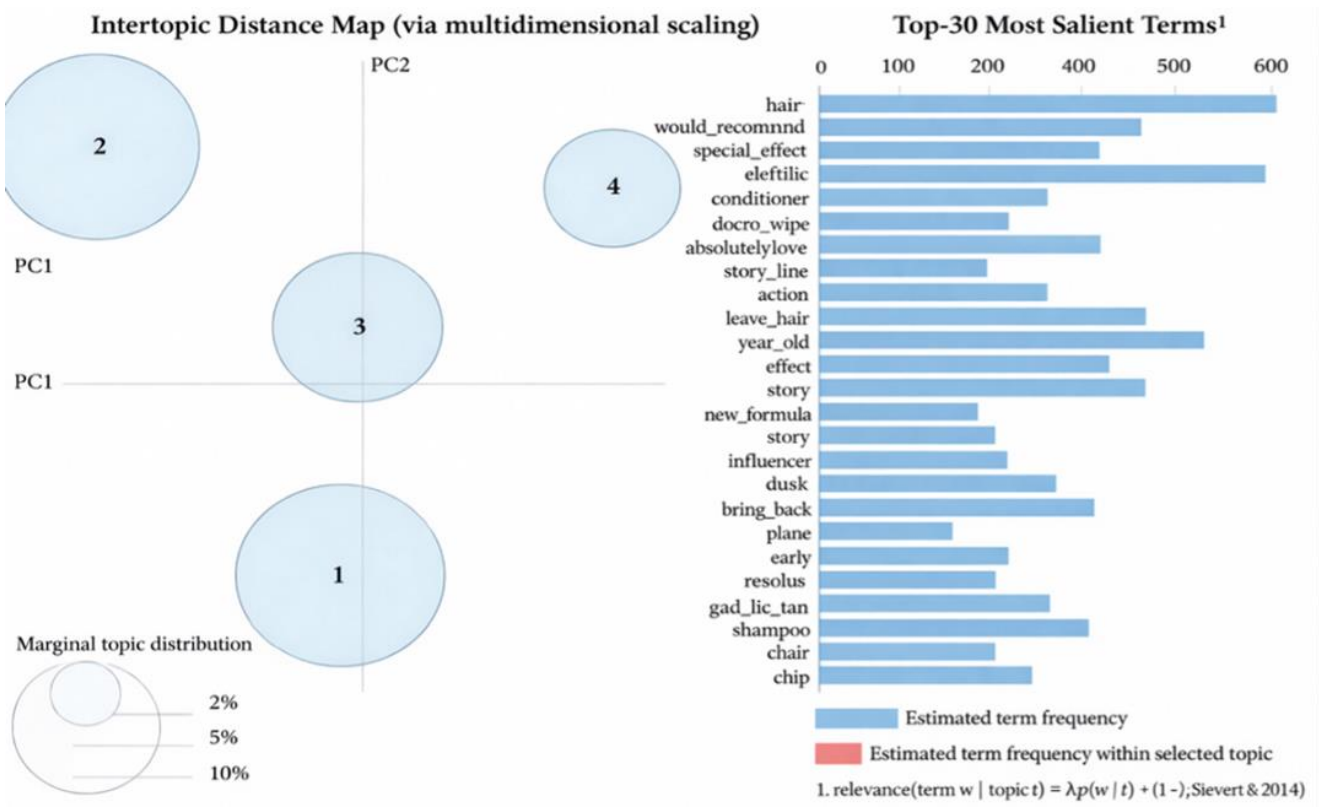

Figure 4. pyLDAvis visualization based on four topics

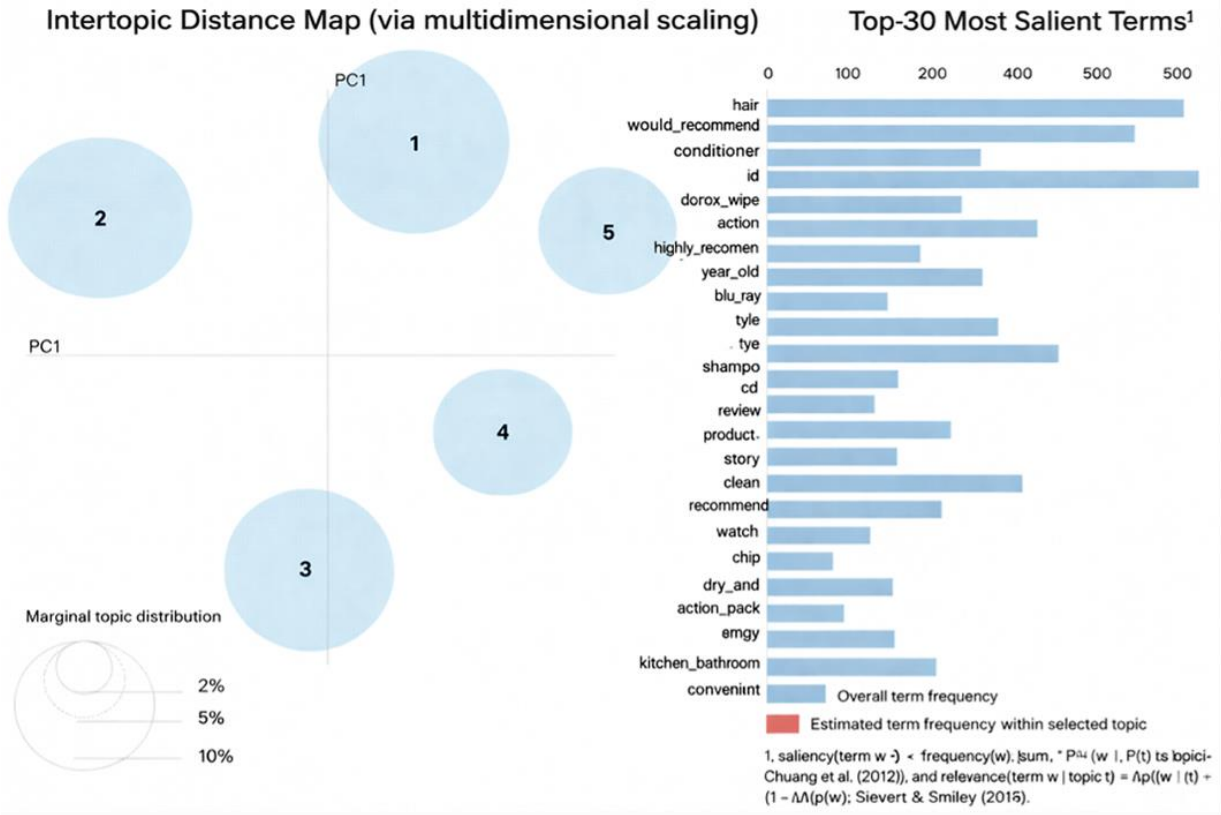

Figure 5. pyLDAvis visualization for five topics

Table 1. Aspect determination based on dominant keywords

Topic	Aspect	Words per Topic
1	Cleaning Product	clorox_wipe, kid, clean, year_old, smell, wipe, absolutely_love, clorox, easy, pretty_good, new_formula, work, chair, clorox_disinfect, car, good, house, weigh_hair, disinfect, every_day
2	Music	would_recommend, put_together, cd, good, son, chip, taste, grandson, music, would, recommend, digital_copy, long_time, kid, quick_easy, kitchen_bathroom, rice, godzilla, look_forward, buy
3	Hair Product	hair, conditioner, leave_hair, receive, soft, influenster, good, shampoo, oily, free, like, free_influenster, end, really, receive_product, opinion, dry_end, feel, receive_free, oily_root
4	Movie	godzilla, special_effect, d, story_line, action, good, plane, story, effect, watch, godzilla_fan, awesome, bring_back, monster, kid, fun, back, king_monster, fan, special

3.2 Automatic label annotation

Automatic annotation was performed by estimating the probability of keyword occurrence for each topic in individual documents. The topic with the highest probability for a specific document was assigned as its label. Figure 6 shows example topic probability scores for several documents in the dataset. Figure 7 shows the topic labels assigned to the reviews based on the dominant LDA topic for each document.

In the first document, the second topic has the highest probability; therefore, it is labeled Topic 2, corresponding to music. The same process is applied to the following documents, with each labeled according to the topic with the highest probability score.

```
[[[(0, 0.05758376), (1, 0.5177093), (2, 0.06273545), (3, 0.36197144)],
  [(0, 0.10759092), (1, 0.67231196), (2, 0.10916908), (3, 0.11092803)],
  [(0, 0.04132955), (1, 0.6723225), (2, 0.1091803), (3, 0.11090642)]],
  [(0, 0.04132955), (1, 0.03712526), (2, 0.58091915), (3, 0.3380837)]],
  [(0, 0.07182444), (1, 0.06916529), (2, 0.79415955), (3, 0.06949284)],
  [(0, 0.21235228), (1, 0.05184841), (2, 0.685575), (3, 0.050224327)])],
  [(0, 0.05362083), (1, 0.05383517), (2, 0.8376882), (3, 0.054585652)],
  [(0, 0.05362083), (1, 0.05383517), (2, 0.8376882), (3, 0.054585652)],
  [(0, 0.066420265), (1, 0.06594669), (2, 0.8206729), (3, 0.04696613)],
  [(0, 0.38921687), (1, 0.06364701), (2, 0.48449925), (3, 0.06263691)],
  [(0, 0.38921687), (1, 0.06364701), (2, 0.48449925), (3, 0.06263691],
  [(0, 0.05974536), (1, 0.48529854), (2, 0.06530165), (3, 0.39145476)]],
  [(0, 0.05974536), (1, 0.48529854), (2, 0.06530165), (3, 0.39145476)]]
```

Figure 6. Topic probabilities for each document

text	user_sentiment	label_st1
awesome love album good hip hop side current p...	1	1
good good flavor review collect part promotion	1	1
good good flavor	1	1
disappoint read review look buy one couple lub....	0	2
irritation husband buy gel us gel caused irrit...	0	2
worth boyfriend bought spice thing bedroom irri...	0	2
disappoint buy earlier today excite check base...	0	2
happy buy product husband try impress tingle w...	1	2
happy buy product husband try impress tingle w...	1	2
disappointing husband buy extra fun werevboth ...	0	2
dont buy get surprise husband nothing special ...	0	1
dont buy get surprise husband nothing special ...	0	1

Figure 7. Topic labeling in the dataset

3.3 Validation results of automatic vs manual annotation

The validation involved randomly selecting 100 samples

from the dataset, which were then annotated using both manual and automatic methods. A comparative analysis of the two annotation methodologies revealed that the automated approach achieved an accuracy of 66%. This level of agreement indicates moderate alignment between automatic and manual annotation. At the same time, it highlights substantial ambiguity that limits the reliability of fully automatic aspect labeling.

On the other hand, some reviews included specific keywords, such as product names (e.g., clorox), which helped the automatic model map the review to a specific topic. This suggests that automatic annotation tends to perform more effectively when the data are well-defined and contain obvious keywords, particularly product names.

Nevertheless, annotation errors persist due to lexical ambiguity and overlapping topic representations, limiting the reliability of fully automatic labeling. Additionally, several reviews lacked clear keywords for specific aspects, making it challenging for the model to assign them to the correct topic. In some cases, manual annotation was also based on the researcher's subjective judgment, as no clear link was found between the review and the predefined topics.

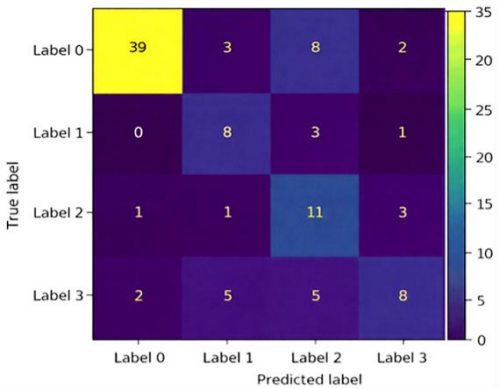

Figure 8. Confusion matrix for each topic label

Figure 8 shows the confusion matrix for each topic label produced by the automatic annotation. Label 0 (Topic 1) performed best, correctly classifying 39 reviews. This is likely due to the presence of highly specific keywords in Topic 1, such as the product name "clorox," which enabled the model to readily identify and assign reviews to the correct topic.

In contrast, the other labels performed adequately, although they still exhibited certain misclassifications. These errors suggest overlapping keywords across topics or reviews that lack relevant keywords, making it difficult for the model to assign them correctly and resulting in incorrect classifications.

3.4 Machine learning model evaluation

To assess the effectiveness of the sentiment classification approach, two main comparisons were made. First, the performance of models trained on the entire dataset (without topic segmentation) was compared with that of models trained separately on topic-specific subsets created through topic modeling. Second, a comparison was conducted between the traditional model based on TF-IDF and Logistic Regression and the BERT-based model.

The evaluation was conducted using standard performance metrics, including accuracy, precision, recall, and F1-score. Additionally, a student's t-test was conducted to assess the statistical significance of performance differences between the models.

3.4.1 Results without topic (baseline)

Table 2 presents the evaluation results for the baseline models trained without topic segmentation. The baseline results indicated that the BERT model outperformed the TF-IDF + Logistic Regression model, achieving an accuracy of 0.94 (vs. 0.89) and an F1-score of 0.83 (vs. 0.75). These results show that BERT's contextual representation more effectively captures the semantic meaning of sentences than the frequency-based representation used by TF-IDF.

3.4.2 Results for each topic

Table 3 presents the evaluation results for the topic-specific models. Overall, BERT consistently outperformed the TF-IDF + Logistic Regression model in both accuracy and F1-score across all topics. For Topic 1, the BERT model achieved an accuracy of 0.94 and an F1-score of 0.82, surpassing the Logistic Regression model, which achieved an accuracy of 0.89 and an F1-score of 0.72, respectively. Similar patterns were observed in Topics 2, 3, and 4, where the BERT-based model achieved higher accuracy and F1-scores than the Logistic Regression model.

Table 2. Model evaluation without topic

Model	Acc.	Precision		Recall		F1
		0	1	0	1	
Logistic Regression + TF-IDF	0.89	0.53	0.95	0.58	0.93	0.75
BERT	0.94	0.79	0.95	0.61	0.98	0.83

Table 3. Model evaluation for each topic

Topic	Model	Acc.	Precision		Recall		F1
			0	1	0	1	
1	LR + TF-IDF	0.89	0.51	0.94	0.48	0.94	0.72
	BERT	0.94	0.82	0.95	0.57	0.98	0.82
2	LR + TF-IDF	0.86	0.46	0.90	0.33	0.94	0.65
	BERT	0.91	0.77	0.92	0.48	0.98	0.77
3	LR + TF-IDF	0.88	0.47	0.93	0.43	0.94	0.69
	BERT	0.92	0.67	0.94	0.53	0.97	0.77
4	LR + TF-IDF	0.87	0.43	0.93	0.44	0.93	0.68
	BERT	0.93	0.73	0.95	0.54	0.98	0.79

When we compared the performance of the baseline model with that of the topic-specific models, we found that the baseline model's accuracy and F1-score are generally higher than the average scores of the topic-specific models. This indicated that topic-based data separation does not consistently improve sentiment classification performance. One possible reason for this outcome is the two-stage prediction process, which consists of aspect identification followed by sentiment classification. Errors in the first stage can propagate to the second stage, leading to overall performance degradation.

3.4.3 Student's t-test

This test was conducted by comparing the actual sentiment labels with the model's predicted outputs. A total of eight tests were performed, covering each topic and the baseline for both model types. Table 4 presents the student's t-test results comparing the baseline and topic-specific models.

All test results exceeded 0.05, indicating that the

performance differences between the baseline models and the topic-specific models were not statistically significant. This suggests insufficient evidence to conclude that segmenting data by topic leads to significantly better or worse model performance.

However, aspect-based data segmentation still offers additional benefits, particularly in enabling more detailed analyses of specific aspects or product categories. This can be used to improve data interpretation tailored to each category's unique needs.

Table 4. Student's t-test results

Topic	Model	P-Value
1	LR + TF-IDF	0.14
	BERT	1
2	LR + TF-IDF	1
	BERT	1
3	LR + TF-IDF	0.48
	BERT	0.34
4	LR + TF-IDF	0.06
	BERT	0.74

Table 5. Number of incorrect predictions

Topic	Model	FP	FN	Total
Baseline	LR + TF-IDF	385	472	857
	BERT	345	199	544
1	LR + TF-IDF	151	136	287
	BERT	125	36	161
2	LR + TF-IDF	129	75	204
	BERT	100	28	128
3	LR + TF-IDF	157	135	292
	BERT	130	71	201
4	LR + TF-IDF	93	99	192
	BERT	76	33	109

3.5 Error analysis

Error analysis was performed to identify FP and FN in the sentiment prediction results. The purpose was to identify specific words or conditions that are likely to cause misclassifications. This analysis was conducted for all models developed in the study.

Table 5 presents the number of incorrect predictions. The results indicate that the BERT model tends to produce fewer errors than the Logistic Regression with TF-IDF model, with a total of 544 errors versus 857 in the baseline model trained on the full dataset. A similar trend is observed in the aspect-based models, where each model also resulted in fewer misclassifications.

Regarding FP and FN, BERT consistently makes fewer FN errors than FP errors, indicating that it is more effective at identifying positive sentiment in sentences.

3.5.1 Identified error patterns

Analysis of misclassified samples reveals several recurring error patterns across the models:

1. Dominance of high-polarity keywords

Both models tend to prioritize strong sentiment words such as *good*, *amazing*, *bad*, and *hate*, which often outweigh the overall context of the sentence. This frequently leads to FP errors when a positive cue appears early in an otherwise negative review, and FN errors when a negative cue disrupts an overall positive message.

2. Conflicting or mixed-sentiment expressions

Several reviews contain both positive and negative cues within the same sentence. LR + TF-IDF misclassifies such cases due to reliance on isolated keywords, while BERT occasionally fails when contradictory sentiment signals appear close together or without clear discourse markers.

3. Aspect or topic ambiguity

Some reviews contain terms associated with multiple topics or aspects, leading to inconsistent automatic aspect assignments. This ambiguity introduces noise into the training labels, leading to polarity errors when sentiment is tied to a specific aspect that the topic model mislabels.

4. Fragmented or incomplete text due to preprocessing

A portion of the dataset contains incomplete or grammatically fragmented sentences. This reduces BERT's ability to infer context from sentence structure, and while LR + TF-IDF is less sensitive to syntax, it still misclassifies such inputs when sentiment cues appear in unexpected positions.

5. Class imbalance

The dataset contains more positive than negative samples, which increases the likelihood of FP errors for models that become biased toward the majority sentiment class. This imbalance reduces the model's sensitivity to negative cues, particularly in ambiguous or mixed-sentiment sentences.

3.5.2 Examples of misclassifications

Table 6 presents examples of FP and FN predictions from the models. Overall, the Logistic Regression + TF-IDF model frequently misclassifies because it focuses on highly weighted words without considering sentence context. In comparison, the BERT model demonstrates overall superior performance, although it remains vulnerable to ambiguous sentences or contradictory words within a single phrase.

Table 6. Samples of False Positives (FP) and False Negatives (FN) predicted by the models

	Model	Text	Act. Label	Predict. Label	Description
FP	LR + TF-IDF	good collection really nice collection bad part come digital copy ultraviolet state	Neg.	Pos.	The model focused on the words "good" and "nice" at the beginning of the sentence, causing it to overlook the word "bad", and thus misclassified the sentiment as positive.
	BERT	watch movie understand hate movie amazing	Neg.	Pos.	The words "amazing" and "hate" both carry strong sentiment, which misled the model into predicting a positive sentiment.
FN	LR + TF-IDF	excellent collection collection amazing must buy resident evil fan	Pos.	Neg.	The word "evil" caused the model to classify the document's sentiment as negative, disregarding the presence of "excellent" and "amazing".
	BERT	well expective always fan godzilla high hope movie disappoint	Pos.	Neg.	The model focused on the word "disappoint", which carries a stronger negative connotation compared to the positive word "well", leading it to predict a negative sentiment.

Additionally, the preprocessed dataset may also cause prediction errors, as incomplete or fragmented sentences can create confusion, especially for context-aware models like BERT, which depend on full sentence structures to interpret meaning accurately.

3.6 Discussion

This study was conducted to automate label annotation for identifying aspects in ABSA. Four optimal topics were identified based on the highest LDA coherence score. When visualized with pyLDAvis, the four topics appeared as roughly equal-sized circles, indicating a balanced distribution of topics within the dataset.

However, some keyword overlap was observed between topics, such as Topic 2 (music) and Topic 4 (movies), with Topic 2 containing terms indicative of films, including "godzilla," "monster," and "watch." Moreover, Topic 2 also included words from multiple categories, such as chip, taste, and rice, which are commonly associated with food. This indicates topic ambiguity, likely due to documents that encompass multiple aspects.

When automatic annotation was compared with manual annotation, an accuracy of 66% was obtained. This level of accuracy was influenced by ambiguity in the review data and by overlapping keywords across multiple topics. These findings suggest that automatic annotation can provide a moderate approximation of manual labels, particularly for reviews that contain specific product names. Nevertheless, data imbalance and the manual annotator's subjective judgment also significantly affect labeling results.

Evaluation results revealed that the BERT model without topic separation achieved the highest performance, with an accuracy of 0.94 and an F1-score of 0.83. It surpassed the average performance of topic-specific BERT models, which had an accuracy of 0.925 and an F1-score of 0.7875. Meanwhile, the Logistic Regression + TF-IDF model demonstrated relatively stable performance, with an accuracy of 0.89 on the full dataset and an average of 0.875 across topic-based models. These findings suggested that topic segmentation does not consistently improve sentiment classification performance.

Several factors may have impacted these results. First, the smaller dataset size per topic resulted in limited training data for topic-specific models. Second, keyword overlaps among topics could have caused ambiguities in automatic topic labeling. These challenges highlight the need to reassess the selection of four topics, as relying on coherence scores alone is insufficient. It's essential to determine whether the selected topics accurately represent the entire dataset or if choosing a different number of topics would alleviate confusion. Conversely, BERT's strong ability to comprehend context enables it to handle topic variation effectively, even without explicit segmentation.

Overall, the results suggest that the aspect-based approach performs comparably to the baseline, with no statistically significant difference. Accordingly, the lack of statistically significant differences suggests that topic-based segmentation does not consistently yield performance improvements under the conditions examined. This interpretation goes beyond the possibility of merely insufficient statistical power. For sentiment classification tasks involving multi-aspect review data, general-purpose BERT-based classifiers remain the most reliable option.

Finally, this study focuses on comparing topic-segmented and non-segmented learning settings rather than benchmarking against the full spectrum of state-of-the-art ABSA architectures. While sequence labeling and lexicon-assisted models are widely used in ABSA, the chosen baselines were selected to isolate the effect of topic-based segmentation under comparable classification architectures. Future work should extend this analysis to more specialized ABSA models. More appropriate statistical tests, such as McNemar's test or repeated cross-validation-based metric comparisons, should be employed in future work to enhance inferential validity.

4. CONCLUSIONS

This study was conducted to develop a method for implementing an automatic labeling system to enhance aspect classification in ASBA. After successfully identifying and segmenting the dataset by these aspects, the study examined how incorporating them affects sentiment analysis. Additionally, the study compared the performance of TF-IDF and Logistic Regression models with BERT in sentiment classification. Based on the results of the research, several conclusions can be drawn.

An automatic method has been successfully implemented to label the aspects. This is demonstrated by using the highest LDA topic probability as the topic or aspect label for each review in the dataset. This labeling method enabled systematic dataset segmentation by topic and laid the groundwork for developing more targeted sentiment classification models.

Dividing the data by topic in this study did not improve model performance. In fact, the BERT model trained without topic segmentation achieved higher accuracy and F1-score than the models trained on individual topics. This indicated that topic-based segmentation does not automatically improve the quality of sentiment analysis. However, the Student's t-test results showed that the performance differences between the two model types were not statistically significant. Thus, there was no strong evidence that aspect-based segmentation caused worse performance. The evaluation findings also demonstrated that the BERT model achieved superior accuracy and F1-scores compared with both the TF-IDF and Logistic Regression approaches.

Although the automatic annotation mechanism was successfully applied in this study, the positive effect of topic segmentation on classification performance was not fully supported. Therefore, this finding opens opportunities for further research into more precise annotation or segmentation methods. Additionally, techniques for addressing data imbalance warrant further investigation. Random oversampling alone still resulted in a noticeable disparity in precision and recall across sentiment labels.

ACKNOWLEDGMENT

This work is fully supported by the Institute for Research and Community Services, Maranatha Christian University, Bandung, Indonesia.

REFERENCES

[1] Sagnika, S., Pattanaik, A., Mishra, B.S.P., Meher, S.K. (2020). A review on multi-lingual sentiment analysis by machine learning methods. Journal of Engineering Science and Technology Review, 13(2): 154. https://doi.org/10.25103/jestr.132.19

[2] Yang, L., Li, Y., Wang, J., Sherratt, R.S. (2020). Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access, 8: 23522-23530. https://doi.org/10.1109/ACCESS.2020.2969854

[3] Jurek, A., Mulvenna, M.D., Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. Security Informatics, 4(1): 9. https://doi.org/10.1186/s13388-015-0024-x

[4] Manek, A.S., Shenoy, P.D., Mohan, M.C., Venugopal, K.R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 20(2): 135-154. https://doi.org/10.1007/s11280-015-0381-x

[5] Chen, H., Li, S., Wu, P., Yi, N., Li, S., Huang, X. (2018). Fine-grained sentiment analysis of Chinese reviews using LSTM network. Journal of Engineering Science & Technology Review, 11(1): 174-179. https://doi.org/10.25103/jestr.111.21

[6] Xing, F.Z., Pallucchini, F., Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. Information Processing & Management, 56(3): 554-564. https://doi.org/10.1016/j.ipm.2018.11.002

[7] Mackey, T.K., Miner, A., Cuomo, R.E. (2015). Exploring the e-cigarette e-commerce marketplace: Identifying Internet e-cigarette marketing characteristics and regulatory gaps. Drug and Alcohol Dependence, 156: 97-103. https://doi.org/10.1016/j.drugalcdep.2015.08.032

[8] Shayaa, S., Jaafar, N.I., Bahri, S., Sulaiman, A., et al. (2018). Sentiment analysis of big data: Methods, applications, and open challenges. IEEE Access, 6: 37807-37827. https://doi.org/10.1109/ACCESS.2018.2851311

[9] Cambria, E. (2013). An introduction to concept-level sentiment analysis. In Mexican International Conference on Artificial Intelligence, Berlin, pp. 478-483. https://doi.org/10.1007/978-3-642-45111-9_41

[10] Nazir, A., Rao, Y., Wu, L., Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. IEEE Transactions on Affective Computing, 13(2): 845-863. https://doi.org/10.1109/TAFFC.2020.2970399

[11] Maroof, A., Wasi, S., Jami, S.I., Siddiqui, M.S. (2024). Aspect-based sentiment analysis for service industry. IEEE Access, 12: 109702-109713. https://doi.org/10.1109/ACCESS.2024.3440357

[12] Rana, T.A., Cheah, Y.N., Letchmunan, S. (2016). Topic modeling in sentiment analysis: A systematic review. Journal of ICT Research & Applications, 10(1): 76-93. https://doi.org/10.5614/itbj.ict.res.appl.2016.10.1.6

[13] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10): 1872-1897. https://doi.org/10.1007/s11431-020-1647-3

[14] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112: 102131. https://doi.org/10.1016/j.is.2022.102131

[15] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., et al. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In Computational Methods for Communication Science, pp. 13-38. https://doi.org/10.1080/19312458.2018.1430754

[16] Tharwat, A. (2021). Classification assessment methods. Applied Computing and Informatics, 17(1): 168-192. https://doi.org/10.1016/j.aci.2018.08.003

[17] Mishra, P., Singh, U., Pandey, C.M., Mishra, P., Pandey, G. (2019). Application of student's t-test, analysis of variance, and covariance. Annals of Cardiac Anaesthesia, 22(4): 407-411. https://doi.org/10.4103/aca.ACA _94_19

[18] Fiandini, M., Nandiyanto, A.B.D., Al Husaeni, D.F., Al Husaeni, D.N., Mushiban, M. (2024). How to calculate statistics for significant difference test using SPSS: Understanding students comprehension on the concept of steam engines as power plant. Indonesian Journal of Science and Technology, 9(1): 45-108. https://doi.org/10.17509/ijost.v9i1.64035

[19] Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6): 599-606. https://doi.org/10.14569/IJACSA.2021.0120670