DOI: https://doi.org/10.34123/jurnalasks. v18i1.895

POLITEKNIK STATISTIKA STIS

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

Analyzing Medium and Long Text Indonesian Tourism Feedback Using Topic Modeling and Sentiment Analysis

Sulisetyo Puji Widodo[1*], Isnaeni Noviyanti[2]

[1]*BPS-Statistics Indonesia, Jakarta, Indonesia,* [2]*Universitas Indonesia, Depok, Indonesia*
**Corresponding Author: E-mail address:* sulisetyo.widodo@bps.go.id

ARTICLE INFO

Article history:
Received 2 Dec, 2025
Revised 20 Dec, 2025
Accepted 2 June, 2026
Published 30 June, 2026

Keywords:
Feedback, Indonesian Tourism, Natural Language Processing, Sentiment Analysis, Topic Modeling

Abstract

Introduction/Main Objectives: Tourism is a vital sector supporting Indonesia's economic growth, making the effective utilization of public feedback essential for improving service quality. Most feedback is collected through web-based forms in the form of open-text responses that provide rich insights but remain underutilized due to their unstructured nature. Background **Problems**: This study examines the challenge of identifying the most suitable topic modeling and sentiment analysis techniques for analyzing medium- and long-text feedback in the Indonesian tourism context. **Novelty**: The novelty lies in the comparative evaluation of classical topic modeling algorithms against modern embedding-based approaches combined with multiple Indonesian transformer models, which has not been extensively explored in tourism-related datasets. **Research Methods**: The research compares LDA and NMF with BERTopic, Top2Vec, kBERT, and kUSE using coherence scores, and evaluates sentiment analysis using majority voting across transformer architectures. **Finding/Results**: The results show that BERTopic performed best for medium-length text, while NMF was optimal for long text, and a RoBERTa-based model achieved the highest sentiment agreement. Positive sentiment often appeared in feedback on facilities and fees, whereas negative sentiment dominated topics on environmental and governance issues. These findings offer valuable insights for tourism managers and policymakers in prioritizing improvements and refining strategies.

1. Introduction

Tourism is a crucial sector for Indonesia's economic development. Therefore, improving and strengthening this sector requires a data-driven approach. In recent years, web-based digital surveys have successfully collected information on travel patterns and tourist experiences at specific destinations. The information collected is generally structured quantitative data, typically multiple-choice or rating scales.

Furthermore, surveys are often accompanied by feedback forms designed to capture respondents' aspirations, complaints, or input regarding their experiences. Feedback can be about the survey process, application usage, or general views on tourism in Indonesia. The data obtained from these forms is generally unstructured and tends to be narrative, but it contains more in-depth and contextual information. Unfortunately, this type of data is often underutilized in policy formulation and tourism

service development. Yet, open-ended user feedback has strategic potential to uncover real-world issues and capture community expectations that may not be reflected in quantitative survey results.

Processing feedback data can benefit various parties. For survey organizers, user feedback is useful for evaluating and refining survey instrument design, improving questionnaire flow, and identifying technical issues that may have gone undetected during the testing phase. Meanwhile, for governments or authorities managing the tourism sector, feedback can provide a direct picture of public perceptions of existing services, infrastructure, and tourist attractions. This information can be used as a basis for developing more appropriate policies based on community needs. Therefore, utilizing feedback can encourage more targeted and participatory interventions.

However, the narrative and unstructured nature of feedback data makes manual analysis inefficient, especially when the volume of data collected is large. Variations in respondents' communication styles, including language style, length of writing, and focus on specific issues, pose challenges in consistently extracting relevant information. This situation demands an approach capable of classifying and capturing public perceptions without losing context.

In this context, Natural Language Processing (NLP)–based approaches offer an effective solution for analyzing large-scale textual feedback. Techniques such as topic modeling and sentiment analysis are particularly useful for uncovering latent thematic structures and sentiment orientations in user-generated content [1]–[11], [12]–[16], [20]. Topic modeling enables the extraction of hidden thematic patterns from text, including issues related to cleanliness, accessibility, services, and environmental sustainability at tourist destinations [1]–[11]. Meanwhile, sentiment analysis is employed to identify users' attitudes toward these themes, categorizing opinions as positive, negative, or neutral based on contextual cues [12]–[16], [20]. To ensure robust evaluation, model predictions are assessed using majority vote and agreement-based measures, allowing the selection of models that best align with dominant labeling patterns [17]–[19]. The integration of topic modeling and sentiment analysis thus provides a more comprehensive understanding of public perception and supports evidence-based decision-making in the tourism sector.

This study aims to (1) explore the potential use of user feedback data in the context of tourism surveys in Indonesia, focusing on identifying key topics emerging from feedback content and analyzing the accompanying sentiment trends. Furthermore, (2) this study evaluates and selects the best models for both topic modeling and sentiment analysis to recommend the most appropriate model for feedback with medium to long text lengths. Furthermore, (3) by applying topic modeling and sentiment analysis approaches, this study is expected to produce an analytical framework that helps map strategic issues of public concern. Finally, (4) the results of this analysis are not only useful for survey managers in improving the quality of instruments and services, but can also be used as considerations for policymakers in developing more participatory, adaptive, and evidence-based tourism development programs.

2. Material and Methods

The workflow in this study is illustrated in Figure 1, which consists of several main stages. (1) Dataset Preparation, where feedback data collected in CSV format is preprocessed, and only medium-length text (11–30 words) and long text (>30 words) are retained, while short text (<11 words) is removed. (2) Topic Modelling, in which the prepared datasets are processed using various algorithms such as GSDMM, BERTopic, Top2Vec, kBERT, kUSE, NMF, Agglomerative, and LDA. (3) Coherence Score Evaluation, conducted to identify the best-performing topic model for each dataset category. (4) Sentiment Analysis, where the best models are integrated with transformer-based classifiers such as RoBERTa, DistilBERT, BERT, ALBERT, and XLM-RoBERTa. (5) Model Agreement, which ensures consistency and reliability of the sentiment results across models. Finally, the process produces the output in the form of sentiment per topic for both medium and long text datasets.

2.1. Dataset

The data used in this study were obtained from feedback forms distributed concurrently with the Nusantara Tourist Survey (Survei Wisatawan Nusantara) conducted by Badan Pusat Statistik (BPS – Statistics Indonesia) throughout 2024. The Nusantara Tourist Survey focuses on domestic tourism activities, namely Indonesian residents traveling within the country. The feedback forms were administered separately from the main survey questionnaire and were intended to capture respondents'

opinions, experiences, and aspirations regarding domestic tourism and the use of the survey application. A summary of the collected feedback data is presented in Table 1.

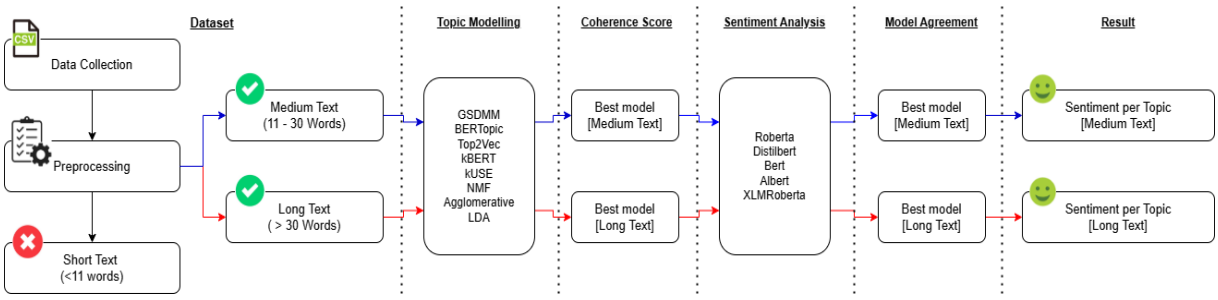

Figure 1. Research workflow

The primary variable in the dataset is "feedback," which contains text entries in the form of open-ended sentences or paragraphs with no length limit or specific structure. All entries were written in Indonesian, reflecting the respondents' population of domestic tourists. Due to its free-form and non-standardized nature, this data displays a variety of language usage, including non-standard words, abbreviations, and informal expressions. Therefore, a preprocessing stage based on Natural Language Processing (NLP) techniques is crucial for filtering, normalizing, and structuring the data before further analysis, such as topic modeling and sentiment analysis.

Table 1. Details of Feedback data

Stage	Feedbacks	Max Words	Avg Words
Data Collection	28.996	1.183	10 - 11
Preprocessing	21.171	851	9 - 10

2.2. Preprocessing

Before further analysis, the text data from the feedback column underwent a pre-processing phase to remove irrelevant elements and adjust the text formatting for uniformity. The first step was to reduce all letters to lowercase to avoid differences in word representation caused solely by capitalization. Next, the text was cleaned of non-linguistic elements such as emojis, symbols, numbers, and links—including internal URLs of the survey system—that were deemed not to contribute to the sentence's core semantic meaning. Redundant characters, such as repeated vowels or consonants, were also removed to normalize variations in informal expressions commonly found in open-ended input data.

Table 2. Details of Feedback data after data separation

Length Category	Feedbacks	Max Words	Avg Words
Short	16.255	10	4
Medium	4.084	30	16
Long	832	851	70

After the structural cleaning phase was completed, the text was separated into word units and filtered using a list of common stopwords that tend not to carry significant semantic weight in the analysis. The remaining words were then stemmed, reducing words to their base or lexical form to unify various morphological variations of the same word. The final results of the pre-processing process were then counted and grouped based on the length of each feedback. Based on a literature review, this study divides feedback into three categories: small (≤10 words), medium (11–30 words), and long (>30 words). This categorization aims to facilitate comparative analyses, such as identifying topic patterns or sentiment tendencies based on feedback length. Details of the dataset grouping are presented in Table 2. Furthermore, Chen et al. [6] highlighted that a large proportion of short texts with generic content may introduce noise and obscure meaningful patterns found in more contextually rich feedback.

Therefore, this study focuses only on the medium and long categories, as these are considered to contain more comprehensive contextual information than short texts.

2.3. Topic Modelling

This stage aims to obtain a thematic structure so that feedback can be grouped based on common themes. In this study, various topic modeling approaches were compared to determine the most effective method for extracting key topics from the text data in the feedback column. Classic statistical models such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were applied with word frequency-based text representations or TF-IDF, each of which groups documents into topics based on the distribution of dominant words. Meanwhile, more modern, semantic distribution-based approaches, such as Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), are designed to handle short and varied documents—such as open-ended feedback data—without requiring predefined topic distributions.

This study also employed deep learning-based approaches (neural embedding), including BERTopic, Top2Vec, kBERT, and kUSE, which utilize transformer models such as Sentence-BERT and Universal Sentence Encoder to obtain vector representations of documents. These models enable more contextual topic grouping based on semantic proximity and generate topics with higher meaning cohesion through clustering algorithms such as K-Means and Agglomerative Clustering.

All models are then evaluated using a coherence score to assess the semantic consistency of the resulting topics. In this study, the C_V metric was used to evaluate and identify the best-performing models for both medium-length text (11–30 words) and long-length text (>30 words). C_V was chosen because this metric has proven to be stable, interpretable, and capable of capturing semantic relationships between topics across both text size categories. The use of a uniform metric across both datasets also facilitates objective and consistent performance comparisons between models.

To improve the relevance and accuracy of topic extraction, the entire topic modeling process in this study was applied separately to only two text length categories: medium (11–30 words) and long (>30 words). This separation was made because the linguistic and semantic characteristics of short feedback tend to differ significantly compared to longer text. By dividing the topic modeling process into length categories, each model can be calibrated to suit the structure and density of information within each data set. This strategy is expected to produce more precise and applicable topic mapping and enable the identification of thematic differences or similarities between categories, which will be useful for supporting decision-making in managing the national tourism sector.

2.4. Sentiment Analysis

The sentiment analysis in this study was conducted using a transformer-based approach, a state-of-the-art deep learning model that has proven effective in understanding semantic and syntactic context in natural text. Five pre-trained Indonesian language models from the Huggingface website were utilized. Each selected model had a different architecture, as shown in Table 3, to determine the best model based on its architecture. At this stage, each entry in the feedback data was modeled separately to identify sentiment polarity, such as positive, neutral, or negative for each model. This process was automated using a sentiment classification pipeline, which generated labels and confidence scores for each prediction, enabling the visualization of opinion trends and analysis of public opinion about the tourism sector.

Table 3. Overview of Indonesian Sentiment Analysis Models

Model	Architectures	Language
w11wo/indonesian-roberta-base-sentiment-classifier	**RoBERTa** For Sequence Classification	Indonesian
afbudiman/distilled-optimized-indobert-classification	**DistilBert** For Sequence Classification	Indonesian
ayameRushia/bert-base-indonesian-1.5G-sentiment-analysis-smsa	**Bert** For Sequence Classification	Indonesian
tyqiangz/indobert-lite-large-p2-smsa	**Albert** For Sequence Classification	Indonesian
cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual	**XLM-RoBERTa** For Sequence Classification	Multilingual

2.5. Model Agreement

This study used a dataset without ground truth labels, so model evaluation for sentiment analysis was conducted using a model agreement approach to assess the level of prediction agreement between models. This study used majority voting, a method that determines the final label based on the highest consensus among models. The level of agreement between a model and the majority label indicates the consistency of predictions on the same data. The model with the highest agreement with the majority label is considered the most stable and representative of the data characteristics and is therefore selected as the primary candidate for further sentiment analysis. This evaluation also considered variations in text length (medium and long) to determine its effect on prediction consistency.

3. Results and Discussion

3.1. Medium dataset

3.1.1 Topic Modelling

At this stage, all feedback that falls into the medium text length category will be assigned topics using each model (GSDMM, BERTopic, Top2Veck, BERT, kUSE, NMF, Agglomerative, and LDA). Each model is tested with several topics and word count settings. Table 4 shows an example of feedback with 5 topics and 10 words assigned topics.

Table 4. Example of Topic Modeling Results Using Various Models (Medium Dataset)

Model	Topic	Feedback
GSDMM	4	Prices for tourists are often excessively high and are deliberately inflated for profit. In some tourist destinations in Indonesia, visitors may also encounter coercive practices and irresponsible behavior from local service providers. (*Harga untuk wisatawan yang sangat mahal dan di sengaja mengambil kesempatan untuk meraih keuntungan dan terkadang tempat wisata di indonesia orang-orang di sana suka memaksa dan tidak bertanggung jawab*)
BERTopic	**0**	
Top2Vec	0	
kBERT	0	
kUSE	0	
NMF	4	
Agglomerative	0	
LDA	3	

3.1.2 Coherence Score

Evaluation results for the medium-length feedback data category show that BERTopic dominates as the model with the best performance across most configurations, Table 5. This model recorded the highest coherence scores for the configurations of 5 topics, 5 words (0.819), 5 topics, 10 words (0.808),

10 topics, 5 words (0.819), and 10 topics, 10 words (0.761). This consistent score demonstrates that the contextual embedding-based approach used by BERTopic is effective in capturing semantic relationships between words, facilitating theme interpretation in medium-sized data.

On the other hand, Non-negative Matrix Factorization (NMF) demonstrated very competitive and even superior performance for the configuration of 15 topics, 5 words (0.816), while maintaining a high score for the configuration of 15 topics, 10 words (0.745). This confirms that NMF has an advantage when the number of topics is expanded, although overall it remains slightly behind BERTopic in terms of consistency across configurations.

Table 5. Coherence Scores of Topic Models (Medium Dataset)

Model	5 Topics, 5 Words	5 Topics, 10 Words	10 Topics, 5 Words	10 Topics, 10 Words	15 Topics, 5 Words	15 Topics, 10 Words
GSDMM	0.597	0.464	0.563	0.470	0.576	0.489
BERTopic	***0.819**	**0.808**	**0.819**	**0.761**	0.762	0.760
Top2Vec	0.453	0.334	0.453	0.344	0.384	0.291
kBERT	0.574	0.405	0.608	0.467	0.604	0.468
kUSE	0.542	0.388	0.574	0.433	0.588	0.458
NMF	0.775	0.697	0.766	0.698	**0.816**	**0.745**
Agglomerative	0.560	0.392	0.497	0.417	0.511	0.402
LDA	0.511	0.414	0.531	0.416	0.574	0.463

Other models, such as GSDMM (maximum 0.576), kBERT (0.608), and kUSE (0.588), produced moderate performance—better than both LDA (0.574) and Agglomerative Clustering (0.560)—but still lagged far behind BERTopic and NMF. Top2Vec, on the other hand, achieved the lowest score (maximum 0.453), making it less reliable in this context.

Figure 2. Top Terms per Topic (5 Topics × 5 Words) by BERTopic on Medium Dataset

Overall, these results confirm that BERTopic is recommended as the primary model due to its high and consistent coherence across various configurations. However, NMF can be a strong alternative, especially for scenarios with a larger number of topics, while the other models are more appropriate for comparison or complementarity in exploratory analysis.

Based on the evaluation in Table 5, BERTopic with a configuration of five topics and five keywords successfully identified distinct themes. Figure 2 illustrates that Topic #1 includes terms such as ticket, visit, and parking, which relate to tourist activities and destination facilities. Topic #2 emphasizes survey, BPS, and SMS, reflecting methods of data collection and respondent outreach. Topic #3 highlights survey, tourist survey, and nusantara, indicating the participation of domestic tourists and their travel behavior, and showing thematic connections with Topic #2.

Meanwhile, Topic #4 is less specific, containing terms such as front and house, making its interpretation more abstract and requiring further contextual analysis. Topic #5 underscores governance and development issues, represented by human resource quality, investment, and environmental aspects. While these keywords are useful for distinguishing topics, thematic overlaps remain between some categories, particularly between Topics #2 and #3, requiring manual interpretation. To clarify the thematic context, Table 6 presents topic interpretations and descriptions, formulated with expert input and intended as a reference for further analysis.

Table 6. Topic Interpretations and Top Keywords from BERTopic (Medium Dataset)

Topic	Interpretation	Description	Keywords
1	Tourist Facilities and Comfort	Tourist experiences related to facilities, cleanliness, tickets, and parking at tourist destinations.	Ticket, Visit, Nusantara, Clean, Parking (*Tiket, Kunjung, Nusantara, Bersih, Parkir*)
2	Survey and Data Collection Methods	Survey data collection methods via BPS and SMS to reach respondents.	Survey, BPS, SMS, Respondent, Sold (*Survei, BPS, SMS, Responden, Laku*)
3	Survei Wisatawan Nusantara	Participation of domestic tourists in the survey to understand travel behavior.	Survey, Tourist Survey, Nusantara, Domestic tourist, Sold (*Survei, Survei Wisatawan, Nusantara, Wisatawan Nusantara, Laku*)
4	Respondents' Perceptions and Preferences	Respondents' perceptions and preferences related to daily experiences.	Front, House, Long, Give, Year (*Depan, Rumah, Lama, Kasih, Tahun*)
5	Investment and Environmental Management	Management of human resources, investment, and environmental aspects with potential overlaps.	Human Resource Quality, Investment, Environmental Aspect, Overlap, Attention aspect *(Kualitas SDM, Investasi, Aspek Lingkung, Tumpang Tindih, Perhati Aspek*)

After identifying the themes, the next step was to examine the distribution of feedback on each topic. Figure 3 shows that the distribution of responses was uneven: Topic #1 (Tourist Facilities and Convenience) was the top performer, followed by Topic #2 (Survey and Data Collection Methods). Conversely, topics like Topic #4 (Respondent Perceptions and Preferences) and Topic #5 (Investment and Environmental Management) received relatively few responses. This imbalance indicates that issues related to tourist facilities and the survey process garnered more attention, while personal perceptions and environmental governance received less attention. This could signal that low-intensity topics require further review, especially if they are assumed to be hidden complaints.

3.1.3 Sentiment Analysis

In this stage, feedback belonging to the medium text length category is analyzed using several sentiment analysis models (RoBERTa, DistilBert, Bert, Albert, XLM-RoBERTa). These models are employed to classify the feedback into three categories: positive, negative, and neutral. Table 7 provides an example of sentiment analysis results, showing how each model assigns different sentiment labels to the same feedback text.

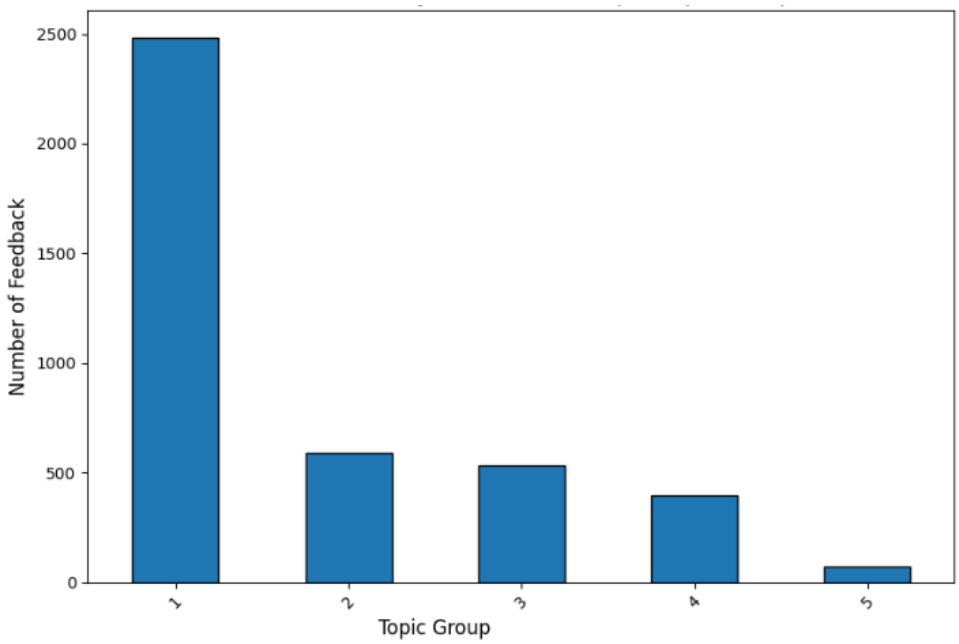

Figure 3. Number of Feedback Entries per Topic by BERTopic (Medium Dataset)

Table 7. Example of Sentiment Analysis Results Using Various Models (Medium Dataset)

Model	Sentiment	Feedback
RoBERTa	(-)	Prices for tourists are very expensive and are deliberately taken advantage of to gain profits and sometimes in tourist attractions in Indonesia the people there like to force and are irresponsible.
DistilBert	(+)	
Bert	(+)	(*Harga untuk wisatawan yang sangat mahal dan di sengaja mengambil kesempatan untuk meraih keuntungan dan terkadang tempat wisata di indonesia orang-orang di sana suka memaksa dan tidak bertanggung jawab*)
Albert	(-)	
XLM-RoBERTa	(-)	

3.1.4 Model Agreement

The evaluation results in Table 8 show that the RoBERTa model achieved the highest agreement with the majority predictions at 82.32%, closely followed by ALBERT at 82.07%. BERT and XLM-RoBERTa also performed competitively, with agreement scores of 79.43% and 78.26%, respectively, while DistilBERT, as a more lightweight model, obtained the lowest agreement at 68.29%.

Table 8. Model Agreement Results for Sentiment Analysis on the Medium Dataset

Model	Agreement with the majority	%
RoBerta	**3.362**	**82.32**
DistilBert	2.789	68.29
Bert	3.244	79.43
Albert	3.352	82.07
XLM-RoBerta	3.196	78.26

These results suggest that architectures with enhanced contextual modeling capabilities, such as RoBERTa and ALBERT, provide more consistent predictions than baseline or distilled models. While BERT and XLM-RoBERTa remain reliable, architectural refinements appear to offer a clear advantage in agreement stability. The lower performance of DistilBERT reflects the trade-off inherent in model

distillation, where reduced complexity improves efficiency but limits the capacity to capture deeper semantic nuances. Overall, model selection for sentiment analysis should balance computational efficiency with the need for contextual depth and prediction stability.

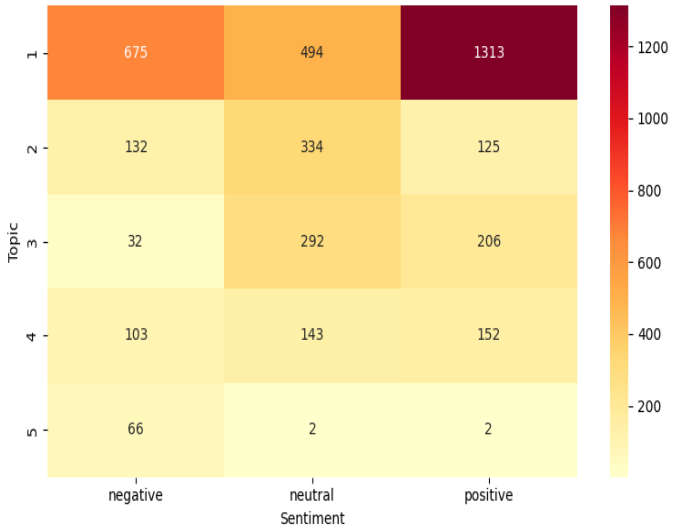

Figure 4. Sentiment Distribution Across Topics by BERTopic (Medium Dataset)

After identifying the most stable sentiment model, the next step was to analyze sentiment distribution across topics using predictions from the RoBERTa model, which achieved the highest agreement score of 82.32%. As shown in Figure 4, the results indicate that:

- Topic 1 (Tourist Facilities and Convenience) was the most frequently discussed topic by respondents and a key focus of attention. This topic was characterized by keywords such as "tickets," "visit," "Indonesian archipelago," "clean," and "parking," and was dominated by positive sentiment, indicating satisfaction with the tourist facilities provided. However, a significant amount of negative sentiment also indicated that there were still a number of complaints regarding these facilities.
- Topic 2 (Survey and Data Collection Methods) and Topic 3 (Domestic Tourist Survey) were dominated by neutral sentiment, with keywords such as "survey," "bps," "sms," "respondents," and "domestic tourist survey." These topics were more informative because many respondents referred to the survey process itself, rather than services or direct experiences.
- Topic 4 (Respondent Perceptions and Preferences) showed a relatively balanced distribution of sentiment. Keywords such as "front," "home," "long," "love," and "year" indicated a variety of opinions regarding respondents' personal perceptions or preferences. This topic reflects a mix of satisfaction and dissatisfaction.
- Topic 5 (Investment and Environmental Management) has the least amount of data, but almost all of the sentiment is negative. Keywords such as human resource quality, investment, environmental aspects, overlapping, and attention to aspects indicate that this topic relates to criticism of environmental management or policy aspects that are not yet optimal. Therefore, this topic needs to be prioritized for future improvement.

Overall, Topic 1 is an area that needs to be maintained and continuously improved because it indicates satisfaction with the majority of respondents, while Topic 5 is the most critical area and requires immediate follow-up improvements.

3.2. Long dataset

3.2.1 Topic Modelling

In this stage, all long text category feedback is classified as assigned to topics using various models, including GSDMM, BERTopic, Top2Vec, kBERT, kUSE, NMF, Agglomerative, and LDA. Each model is evaluated with different configurations of topic number and word count. Table 9 presents an example of feedback from the long data set that has been assigned to five topics with ten representative words.

Table 9. Example of Topic Modeling Results Using Various Models (Long Dataset)

Model	Topic	Feedback
GSDMM	4	This feedback is provided in response to the 2024 Digital Nusantara Tourist Survey conducted by Badan Pusat Statistik (BPS – Statistics Indonesia). The survey is planned to be implemented by BPS in July 2024. The survey is expected to generate comprehensive information related to tourism activities in Indonesia, including detailed tourist profiles, travel motivations, purposes of travel, types of accommodation used during trips, … (*Masukan terhadap Survei Digital Wisatawan Nusantara 2024 oleh Badan Pusat Statistik (BPS), namun berikut adalah beberapa informasi mengenai survei tersebut: Survei Digital Wisatawan Nusantara 2024 akan dilakukan oleh BPS pada Juli 2024. Survei ini akan menghasilkan informasi terkait wisata, seperti profil wisatawan, maksud perjalanan, akomodasi yang digunakan, dan rata-rata lama perjalanan.*)
BERTopic	0	
Top2Vec	0	
kBERT	0	
kUSE	0	
NMF	4	
Agglomerative	0	
LDA	3	

3.2.2 Coherence Score

Referring to Table 10, the topic modeling evaluation results for the long feedback data category show that Non-negative Matrix Factorization (NMF) consistently performed best compared to the other seven models. NMF recorded the highest coherence scores in almost all configurations: 0.680 for 5 topics of 5 words, 0.698 for 10 topics of 5 words, and 0.696 for 15 topics of 5 words. Even in other configurations, NMF remained above 0.58, indicating consistent performance in maintaining semantic relationships between words. These values strengthen evidence that NMF is effective in mapping thematic structures in long texts, which generally have higher vocabulary variety and meaning complexity.

Table 10. Coherence Scores of Topic Models (Medium Dataset)

Model	5 Topics, 5 Words	5 Topics, 10 Words	10 Topics, 5 Words	10 Topics, 10 Words	15 Topics, 5 Words	15 Topics, 10 Words
GSDMM	0.522	0.441	0.519	0.453	0.529	0.481
BERTopic	0.602	0.473	0.576	0.473	0.602	0.473
Top2Vec	0.418	0.333	0.444	0.329	0.444	0.329
kBERT	0.569	0.435	0.561	0.446	0.563	0.450
kUSE	0.551	0.451	0.541	0.437	0.545	0.463
NMF	**0.680**	**0.581**	***0.698**	**0.625**	**0.696**	**0.609**
Agglomerative	0.476	0.385	0.538	0.453	0.534	0.419
LDA	0.563	0.450	0.528	0.391	0.541	0.419

Under NMF, BERTopic and kBERT performed quite competitively. BERTopic achieved the highest score of 0.602 for the 5 topics of 5 words and 15 topics of 5 words. Furthermore, kBERT achieved a maximum score of 0.569 for the 5 topics of 5 words. Both models demonstrated the ability to capture semantic representations, although their consistency was still not on par with NMF. Meanwhile, GSDMM recorded the highest score of 0.529 on 15 5-word topics, but its performance remained below the three main models. Models such as Top2Vec, kUSE, and Agglomerative Clustering showed fluctuating results, with average scores below 0.55, indicating limitations in maintaining coherence across configurations. LDA, while relatively stable, only achieved a maximum score of 0.563 on 5 5-word topics, making it less than ideal for analyzing long text data.

Overall, for the long feedback category, NMF is the most recommended choice because it consistently provides the highest scores across almost all configurations. A configuration of 10 5-word topics can be considered the most optimal representation, while 5 5-word topics and 15 5-word topics remain strong options for maintaining a balance between topic number and semantic cohesion. Models such as BERTopic and kBERT can be considered as secondary alternatives or comparisons in exploratory analysis.

Topic modeling results on the long dataset yielded ten main topics with more targeted keywords that illustrate diverse issues in tourism, as shown in Figure 5. Topic #6 highlights travel costs with the terms "ticket prices," "airplane tickets," and "expensive flights," while Topic #2 focuses on destination quality and accessibility with the terms "tourist attractions," "Indonesian tourism," and "access to places." Topic #3 addresses environmental and governance issues, Topic #4 emphasizes the digitalization of tourism, while Topic #5 addresses geographic aspects with the terms "Java island," "island country," and "sea route."

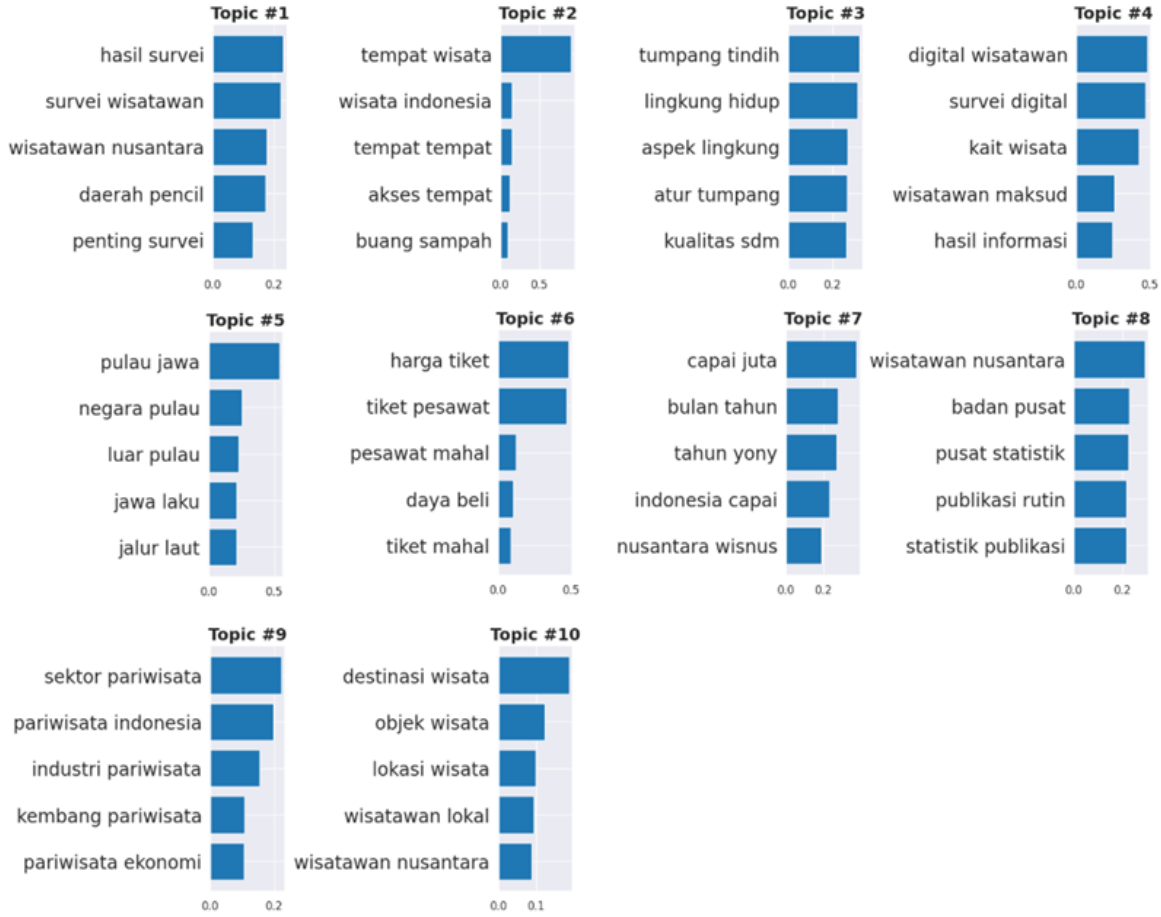

Figure 5. Top Terms per Topic (10 Topics × 5 Words) by NMF on Long Dataset

In contrast, Topics #9 and #10 reflect a macro perspective on tourism development. The first emphasizes industry aspects with the terms "tourism sector," "tourism industry," and "economic tourism," while the second highlights destination issues and the role of domestic tourists with the terms

"tourist destinations," "tourist attractions," and "local tourists." This separation of themes is quite clear, although there is potential for overlap, for example, between Topics #2 and #10, which both discuss destinations. These findings indicate that the model successfully maps major issues into more specific topics, which is further strengthened by expert interpretations in Table 11 to provide a more comprehensive context for subsequent analysis.

Table 11. Topic Interpretations and Top Keywords from NMF (Long Dataset)

Topic	Interpretation	Description	Keywords
1	Tourist Survey and Data Collection	Highlights the implementation of the domestic tourist survey, including the importance of reaching remote areas for data completeness.	survey results, tourist survey, domestic tourists, remote areas, survey importance (*hasil survei, survei wisatawan, wisatawan nusantara, daerah terpencil, penting survei*)
2	Quality and Accessibility of Tourist Destinations	Discusses access to destinations, cleanliness of tourist environments, and waste issues affecting tourist experiences.	tourist sites, Indonesian tourism, places, access to places, throw garbage. (*tempat wisata, wisata Indonesia, tempat tempat, akses tempat, buang sampah*)
3	Environment and Governance	Raises issues of tourism governance related to human resource quality, environmental sustainability, and regulatory overlap.	overlap, environment, environmental aspects, overlap regulation, human resource quality (*tumpang tindih, lingkungan hidup, aspek lingkungan, atur tumpang, kualitas SDM*)
4	Digitalization in Tourist Surveys	Focuses on the use of digital technology to facilitate tourist surveys and information processing.	digital tourists, digital survey, related tourism, tourist intention, survey results (*digital wisatawan, survei digital, kait wisata, wisatawan maksud, hasil informasi*)
5	Geographical Dimension (Islands & Transportation Routes)	Emphasizes the importance of geographical factors such as Java island, outer islands, and sea routes in tourist mobility.	Java Island, island nation, outside islands, Java trend, sea routes (*pulau Jawa, negara pulau, luar pulau, Jawa laku, jalur laut*)
6	Travel Costs & Transportation	Highlights the high price of airline tickets and their relation to tourists' purchasing power.	ticket price, plane ticket, expensive plane, purchasing power, expensive ticket (*harga tiket, tiket pesawat, pesawat mahal, daya beli, tiket mahal*)
7	Tourist Visit Statistics	Presents tourist visit data based on numbers, time periods, and annual achievement trends.	reach millions, month year, year on year, Indonesia reached, domestic tourists (*capai juta, bulan tahun, tahun yoy, Indonesia capai, nusantara wisnus*)
8	Data Publication & Official Statistics	Emphasizes the role of BPS in providing and publishing tourism statistics regularly.	domestic tourists, central agency, statistics center, regular publication, statistics publication (*wisatawan nusantara, badan pusat, pusat statistik, publikasi rutin, publikasi statistik*)
9	Tourism Industry & Economy	Describes tourism's contribution as a growing industry sector supporting the national economy.	tourism sector, Indonesian tourism, tourism industry, tourism development, tourism economy (*sektor pariwisata, pariwisata Indonesia, industri pariwisata, kembang pariwisata, pariwisata ekonomi*)
10	Destinations & Local Tourists	Discusses visits by local tourists to various objects and tourist locations in the archipelago.	tourist destinations, tourist attractions, tourist locations, local tourists, domestic tourists (*destinasi wisata, objek wisata, lokasi wisata, wisatawan lokal, wisatawan nusantara*)

Once the main topics were identified, the next step was to analyze the feedback on each topic group. Figure 6 shows an uneven distribution of respondents' responses. Topic #10 (Destinations & Local Tourists) topped the list with the highest number of feedbacks, followed by Topic #1 (Tourist Surveys

and Data Collection) and Topic #2 (Destination Quality and Accessibility). In contrast, several other topics such as Topic #4 (Digitalization of Tourist Surveys), Topic #5 (Geographic Dimension), and Topic #7 (Tourist Visit Statistics) received only a few responses, each below 50. This imbalance indicates that issues related to tourist destinations, service quality, and tourist surveys received more frequent attention in the feedback, thus requiring more attention as they likely represent some of the most relevant to respondents.

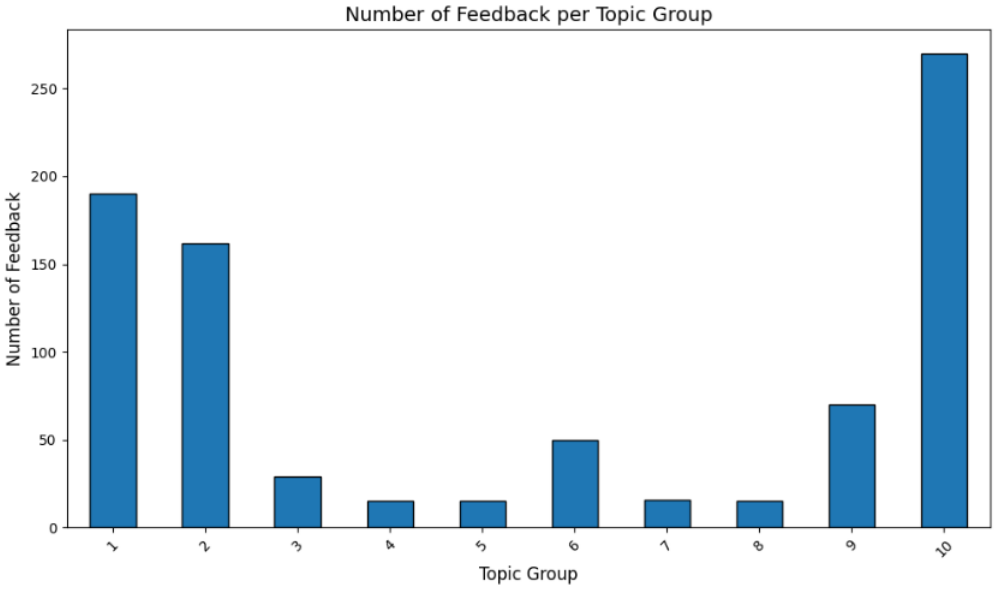

Figure 6. Number of Feedback Entries per Topic by NMF (Long Dataset)

3.2.3 Sentiment Analysis

Similarly to the medium dataset, feedback classified within the long text length category is analyzed using a range of sentiment analysis models, namely RoBerta, DistilBert, Bert, Albert, XLM-RoBERTa. The purpose of employing these models is to systematically categorize the feedback into three sentiment classes: positive, negative, and neutral. Table 12 provides an illustrative example of the analysis results, highlighting how each model may produce different sentiment labels when applied to the same feedback text.

Table 12. Example of Sentiment Analysis Results Using Various Models (Long Dataset)

Model	Sentiment	Feedback
RoBERTa	(-)	MasyaAllah, since 2024, the results of the BPS tourism to Madiun have been satisfactory. We took our family on a trip to Madiun. Enjoying the beauty of Madiun. The food and recreation areas are really economical and pocket-friendly. We are truly satisfied. Every trip in the city center of Madiun, we are treated to several beautiful statues that are truly similar to the original. The government is truly top notch. Vacations now don't have to go far anymore.
DistilBert	(+)	
Bert	(+)	
Albert	(-)	(*MasyaAllah dari tahun 2024 hasil BPS wisata ke Madiun hasilnya memuaskan. Kami mengajak keluarga jalan jalan ke madiun.. Menikmati indahnya kota madiun. Makanan dan tempat rekreasinya bener bener hemat dan ramah dikantong. Kami bener bener puas.setiap perjalanan dikota pusat madiunnya disuguhi beberapa patung patung yg bagus dan bener bener mirip banget dengan aslinya. Benerbener pemerintahanya Top markotop dah. Liburan sekarang gak perlu jauh jauh lagi*)
XLM-RoBERTa	(-)	

3.2.4 Model Agreement

In the sentiment analysis of the long-category dataset, the results presented in Table 13 demonstrate clear performance differences among the evaluated models. RoBERTa achieved the highest agreement with the majority vote at 84.38%, indicating its strong capability in capturing sentiment nuances in long Indonesian texts. BERT followed with an agreement score of 82.09%, showing that it remains effective in modeling contextual information, although its performance is slightly lower than that of RoBERTa. ALBERT and XLM-RoBERTa exhibited moderate performance, with agreement levels of 75.96% and 75.24%, respectively, reflecting a balance between model efficiency and, in the case of XLM-RoBERTa, multilingual generalization. DistilBERT, as a lightweight distilled model, recorded the lowest agreement at 65.75%, suggesting that while computationally efficient, model distillation reduces sentiment classification accuracy for longer texts.

Table 13. Model Agreement Results for Sentiment Analysis on Long Dataset

Model	Agreement with the majority	%
RoBERTa	**702**	**84.38**
DistilBert	547	65.75
Bert	683	82.09
Albert	632	75.96
XLM- RoBERTa	626	75.24

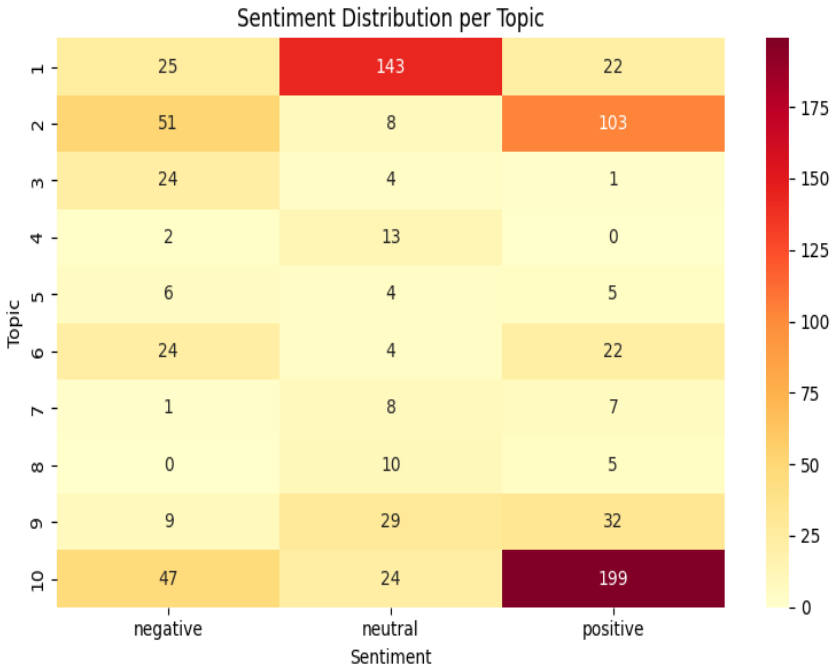

Figure 7. Number of Feedback Entries per Topic by NMF (Long Dataset)

After identifying the most consistent sentiment model, the next step was to examine the distribution of sentiment within each topic based on the predictions of the RoBERTa model. The analysis revealed the following results:

- Topic 10 (Destinations and Local Tourists) generated the most positive sentiment, reflecting respondents' appreciation for domestic tourist destinations. This topic was characterized by keywords such as tourist destinations, tourist attractions, and local tourists, confirming enthusiasm for the domestic tourism sector.
- Topic 2 (Quality and Accessibility of Tourist Destinations) and Topic 6 (Travel & Transportation Costs) also showed predominantly positive sentiment. However, in Topic 6,

complaints persisted regarding high travel costs or ticket prices, which impact respondents' purchasing power.

- Topic 3 (Environment and Governance) and Topic 5 (Geographic Dimension/Islands & Transportation Routes) displayed largely negative sentiment, indicating concerns regarding environmental issues, unequal access between regions, and limited infrastructure.
- Topic 1 (Tourist Survey & Data Collection) and statistical topics such as Topic 7 (Tourist Visit Statistics) and Topic 8 (Official Data & Statistics Publication) tended to be neutral, with respondent feedback being more informative regarding the survey mechanism and tourism data.

Overall, Topics 10, 2, and 6 were perceived positively and should be maintained, while Topics 3 and 5 should be prioritized for improvement due to numerous complaints regarding governance and limited access to the tourism sector.

Limitations

However, this study has several limitations that should be acknowledged. First, the dataset used in this research exhibited an imbalance across both sentiment categories and topic groups, which may have influenced the representativeness and stability of the sentiment distribution. Second, short texts were deliberately excluded from the analysis due to their sparse linguistic features and limited contextual information. Although this decision helped improve topic coherence and sentiment consistency, it may limit the generalization of findings to shorter forms of feedback. Third, the sentiment evaluation relied solely on the model agreement approach since no human-annotated ground truth was available for validation. While this approach provides a reliable estimation of model consistency, it cannot fully replace human judgment in assessing nuanced emotional expressions. Lastly, all models were trained and tested using domain-specific data related to the tourism sector, which may constrain the transferability of the results to other domains or datasets.

4. Conclusion

On the medium-sized dataset, the evaluation results showed that BERTopic was the superior model with the highest coherence score. This model was able to group respondent feedback into more structured topics, such as tourist facilities, surveys, domestic tourists, and tourism governance. The distribution of feedback revealed that the topic of tourist facilities received the most attention and was largely positive, indicating respondents' appreciation for the well-established destination management.

Furthermore, on the long-sized dataset, NMF performed more consistently than the other models. NMF was more effective at capturing a more complex vocabulary, resulting in a wider variety of topics, including travel costs, destination quality and accessibility, environmental issues, and tourism digitalization. Among these topics, tourist destinations were the most frequently discussed theme and were generally viewed positively, although some complaints about travel costs and access were still found.

It is also noteworthy that several topics within the medium and long text datasets contained only a small number of feedback entries because the phenomena they represented naturally occurred less frequently in the survey. Removing these topics could risk eliminating minor issues that are substantively significant, such as specific complaints or technical suggestions from respondents. Therefore, these topics were retained as part of the thematic diversity in the medium and long text datasets.

In terms of sentiment analysis, RoBERTa proved the most stable across both datasets, followed closely by ALBERT and BERT, while DistilBERT was the least consistent. The combination of topic models (BERTopic for medium-length text and NMF for long-form text) with RoBERTa as the sentiment model represents the most optimal configuration.

In conclusion, this study (1) demonstrates the usefulness of user feedback data in the context of Nusantara Tourist Survey, showing its potential to capture valuable insights for evaluation and development. Furthermore, (2) the analysis successfully identified key topics emerging from feedback content along with their sentiment trends, where tourist destinations and facilities were generally assessed positively, while travel costs, environmental concerns, and governance issues generated recurring negative sentiments. In addition, (3) the evaluation of several models showed that RoBERTa was the most consistent sentiment model, and when combined with BERTopic for medium-length texts and NMF for long-form texts, it provided the most optimal configuration. Finally, (4) these findings establish an analytical framework that can guide survey managers in improving instruments and

services, while also supporting policymakers in prioritizing actions such as enhancing accessibility, controlling costs, and strengthening tourism governance to ensure more participatory, adaptive, and evidence-based tourism development.

Ethics approval

The study was conducted in accordance with the ethical guidelines, and informed consent was obtained from all individual participants included in the study.

Acknowledgments

The authors would like to thank BPS-Statistics Indonesia for providing the data utilized in this research.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

The research received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Sulisetyo Puji Widodo: Conceptualization, Methodology, Investigation, Data Curation, Software Development**. Isnaeni Noviyanti:** Formal Analysis, Writing – Original Draft Preparation.

References

[1] D. Angelov, "Top2Vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.

[2] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. First Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA, Jul. 2010, pp. 80–88.

[3] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in Proc. 22nd Int. World Wide Web Conf. (WWW '13), Rio de Janeiro, Brazil, May 2013, pp. 1445–1456.

[4] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: A survey," IEEE Trans. Knowl. Data Eng., vol. 34, no. 3, pp. 1427–1445, Mar. 2022.

[5] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in Advances in Information Retrieval (ECIR 2011), LNCS, vol. 6611. Berlin, Germany: Springer, 2011, pp. 338–349.

[6] Z. Ji, Z. Lu, and H. Li, "An information retrieval approach to short text conversation," arXiv preprint arXiv:1408.6988, 2014.

[7] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," in Proc. 20th ACM Int. Conf. Inf. Knowl. Manag. (CIKM '11), Glasgow, U.K., Oct. 2011, pp. 2333–2336.

[8] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF–IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
[9] B. Bianchi, G. Lami, and F. Sebastiani, "CombinedTM: Combining topic models for improved short text modeling," Inf. Process. Manage., vol. 58, no. 2, 2021.
[10] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "The embedded topic model," arXiv preprint arXiv:1707.01417, 2020.
[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM), Shanghai, China, 2015, pp. 399–408.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[14] [P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[15] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063, 2019.
[16] M. A. Jahin, M. N. Uddin, and M. A. Hossain, "TRABSA: Transformer and attention-based bidirectional LSTM for sentiment analysis," Sci. Rep., vol. 14, 2024.
[17] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," Comput. Linguist., vol. 34, no. 4, pp. 555–596, 2008.
[18] K. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Gaithersburg, MD: Advanced Analytics, LLC, 2014.
[19] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE Publications, 2018.
[20] J. M. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," Inf. Sci., vol. 311, pp. 18–38, 2015.