computation

MDPI

Article

Sentiment Analysis of Tourist Reviews About Kazakhstan Using a Hybrid Stacking Ensemble Approach

Aslanbek Murzakhmetov [1,*], Maxatbek Satymbekov [2,3], Arseniy Bapanov [1] and Nurbol Beisov [2,*]

[1] Department of Information Systems, Faculty of Technology, M.Kh. Dulaty Taraz University, Taraz 080001, Kazakhstan; a.bapanov@dulaty.kz
[2] Department of Computer Science, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; m.n.satymbekov@gmail.com
[3] School of STEM Education, Shakarim University, Semey 071410, Kazakhstan
* Correspondence: aslanmurzakhmet@gmail.com (A.M.); beisov.nurbol@gmail.com (N.B.)

Abstract

Tourist reviews provide essential insights into travellers experiences and public perceptions of destinations. In Kazakhstan, however, sentiment analysis, particularly using ensemble learning, remains underexplored for evaluating such reviews. This study proposes a hybrid stacking ensemble for sentiment analysis of English-language tourist reviews about Kazakhstan, integrating four complementary approaches: VADER, TextBlob, Stanza, and Local Context Focus Mechanism with Bidirectional Encoder Representations from Transformers (LCF-BERT). Each model contributes distinct analytical capabilities, including lexicon-based polarity detection, rule-based subjectivity evaluation, generalised star-rating estimation, and contextual aspect-oriented sentiment classification. The evaluation utilised a cleaned dataset of 11,454 TripAdvisor reviews collected between February 2022 and June 2025. The ensemble aggregates model outputs through majority and weighted voting strategies to enhance robustness. Experimental results (accuracy 0.891, precision 0.838, recall 0.891, and F1-score 0.852) demonstrate that the proposed method KazSATR outperforms individual models in overall classification accuracy and exhibits superior capacity for aspect-level sentiment detection. These findings underscore the potential of the hybrid ensemble as a practical and scalable tool for the tourism sector in Kazakhstan. By leveraging multiple analytical paradigms, the model enables tourism professionals and policymakers to better understand traveller preferences, identify service strengths and weaknesses, and inform strategic decision-making. The proposed approach contributes to advancing sentiment analysis applications in tourism research, particularly in underrepresented geographic contexts.

Keywords: tourist reviews; sentiment analysis; ensemble model; VADER; TextBlob; Stanza; LCF-BERT

check for updates

Academic Editor: Francesco Cauteruccio

Received: 18 August 2025
Revised: 24 September 2025
Accepted: 28 September 2025
Published: 13 October 2025

Citation: Murzakhmetov, A.; Satymbekov, M.; Bapanov, A.; Beisov, N. Sentiment Analysis of Tourist Reviews About Kazakhstan Using a Hybrid Stacking Ensemble Approach. *Computation* **2025**, *13*, 240. https://doi.org/10.3390/computation13100240

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Kazakhstan is the largest landlocked country in Central Asia and is becoming an increasingly attractive destination for tourists due to its rich cultural heritage, diverse natural landscapes, and developing infrastructure. The latest statistics from the Ministry of Tourism and Sports [1] show significant growth in the tourism sector, both in terms of international arrivals and domestic mobility. In 2023, Kazakhstan welcomed 9.2 million foreign tourists, rising to 15.3 million in 2024 (+66%). Domestic tourism also grew from

9.6 to 10.5 million trips (+9.4%). Key destinations include Almaty, Astana, Shymkent, and natural sites such as Altyn-Emel, Charyn Canyon, and Kaindy Lake.

Modern travellers increasingly post reviews on TripAdvisor, Yelp, and Google Reviews, making user-generated content a vital resource. Analysing reviews helps identify infrastructure strengths and weaknesses, adapt services to expectations, and improve competitiveness [2]. User-generated content has become a critical component of contemporary tourism research and practice. It not only supports travellers in making evidence-based decisions but also provides tourism organisations with systematic feedback for service improvement. Platforms such as TripAdvisor serve as valuable instruments for analysing tourist perceptions, offering structured insights into satisfaction, shortcomings, and expectations [3].

The TripAdvisor platform has become a particularly significant source of information for analysing the tourist experience, including in relation to tourist destinations in Kazakhstan [4]. An analysis of the tone of reviews on TripAdvisor also proved useful for assessing the economic aspects of tourism [5,6]. In the Kazakh context, reviews of retail outlets, hotels, and natural attractions help shape management and investment decisions [7–9]. Thus, analysing user content not only contributes to understanding subjective perceptions but also to developing sound economic policies. From a technical point of view, sentiment classification on platforms such as TripAdvisor involves identifying opinions, determining their polarity, and extracting aspects. These processes are increasingly being automated using natural language processing (NLP) tools, which is particularly important in the multilingual context of Kazakhstan [10]. Tourists actively shape the digital reputation of Kazakhstani destinations by posting detailed reviews, ratings, and photos. Such active engagement makes the TripAdvisor an important source for analysing tourist behaviour and improving service quality [11,12].

Sentiment analysis, also known as opinion analysis, has become an important tool in tourism research, allowing for the assessment of tourist satisfaction, perception of service quality, and destination image. Numerous studies have applied sentiment analysis to online reviews to identify key aspects of service, detect dissatisfaction, and improve user experience. Early approaches mainly relied on lexical and rule-based methods, such as VADER and TextBlob, due to their simplicity and interpretability [13]. These models perform well with short, informal texts and do not require large computational resources, making them suitable for analysing social media posts or short reviews [14]. With the development of NLP methods, transformer models such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), which have the ability to take into account the semantic and contextual level of the text, have become widespread. Particularly promising in the analysis of tourist reviews are Aspect-Based Sentiment Analysis (ABSA) models, which allow the emotional tone of specific aspects mentioned in the text, such as 'cleanliness,' 'staff,' or 'location,' to be identified [15,16]. However, such models require significant computing power and large training samples. In recent years, more and more research has focused on ensemble approaches that combine several models to improve accuracy and robustness. For example, hybrid systems have shown superiority over single models in areas such as e-commerce and healthcare [17]. Some studies have examined multilingual reviews or focused on popular tourist destinations [18,19], but the Central Asian region, and Kazakhstan in particular, remain understudied, and reviews of tourism in Kazakhstan have been analysed rarely, and mainly using single models [20]. Thus, despite the availability of a significant number of tourist reviews about Kazakhstan and the growing accessibility of sentiment analysis algorithms, existing studies in this area are generally limited to the application of individual models—lexical, rule-based, or transformer-based. At the same time, there is no holistic

approach that combines the advantages of several methods in a single architecture. This study fills this scientific and applied gap by proposing an ensemble model that combines TextBlob, VADER, Stanza, and LCF-BERT. The aim of this work is to evaluate the effectiveness of such an ensemble in the context of analysing English-language tourist reviews about Kazakhstan, with an emphasis on improving the accuracy and interpretability of the results. To achieve this aim, we address the following research questions: (1) How can a dataset of English-language tourist reviews about Kazakhstan from TripAdvisor be effectively collected and preprocessed to ensure reliability for sentiment analysis? (2) How do individual sentiment analysis models perform on this dataset, and what are their relative strengths and limitations? (3) Can an ensemble model that integrates lexical, rule-based, and transformer-based approaches provide superior performance compared to single models? (4) To what extent does the proposed ensemble improve accuracy, robustness, and interpretability in aspect-based sentiment analysis of tourist reviews? In response to these research questions, we propose a hybrid stacking ensemble method for sentiment analysis of tourist reviews, applied for the first time to English-language data on attractions in Kazakhstan. The ensemble architecture is formalised as a two-level system: at the first level, predictions and confidence scores from four heterogeneous models (TextBlob, VADER, Stanza, and LCF-BERT) are combined into a feature vector, while at the second level, a Random Forest meta-classifier captures latent dependencies among them to generate the final decision. This approach has been comprehensively validated on a cleaned dataset of 11,454 TripAdvisor reviews, thereby making our study one of the first large-scale empirical investigations into Kazakhstan's tourism discourse. Furthermore, we outline future directions for extending this method to multilingual datasets through the integration of state-of-the-art generative models and in-context learning strategies, opening opportunities for more universal and scalable sentiment analysis across culturally and linguistically diverse contexts.

The remainder of this paper is structured as follows. In Section 2, recent research on sentiment analysis is reviewed, highlighting the advantages of hybrid and ensemble methods that integrate lexical and transformer-based approaches in multicultural contexts. In Section 3, the dataset of 11,454 TripAdvisor reviews and the four baseline models (VADER, TextBlob, Stanza, LCF-BERT) are described, including their performance metrics and limitations. In Section 4, a Stacking-based ensemble with a Random Forest meta-classifier is introduced, designed to overcome the shortcomings of individual models and improve generalisation. In Section 5, experimental findings show that the ensemble significantly outperforms single models and voting schemes across all key metrics, offering more accurate predictions in conflicting cases. In Section 6, the comparative advantages of the ensemble approach are discussed, demonstrating its superior precision, recall, and robustness for classifying tourist reviews in Kazakhstan.

2. Literature Review

Recent studies increasingly emphasise the importance of integrating sentiment analysis methods to gain a deeper understanding of tourist preferences and improve the quality of services provided in the tourism and hospitality sector. Hybrid ensemble approaches, which combine lexical tools and deep learning models, are particularly noteworthy due to their ability to capture subtle emotional nuances contained in user reviews. For example, in [21], Malebary and Abulfaraj developed a stacking ensemble model that combines sentiment lexicons with traditional machine learning algorithms. The model showed improved results in classifying user content, highlighting the effectiveness of combining rule-based features and supervised classification methods for practical tourism analytics problems. Comparative studies of lexical tools have also become an important area of research. In [13],

a study was conducted comparing TextBlob and VADER in the analysis of comments on tourism videos. The results showed that VADER is better at identifying positive emotions, while TextBlob provides a more balanced assessment across all tone classes. These findings confirm that no single method can be universal, and hybrid or ensemble models allow the strengths of individual approaches to be effectively combined. This idea is confirmed in [22], which compared lexical and transformer models: BERT architectures demonstrated superiority on large data volumes, while lexical models remain competitive and interpretable when working with low-annotation and multilingual corpora—particularly relevant for diverse tourism data.

In an effort to go beyond basic sentiment classification, researchers have begun to integrate aspect-oriented analysis and deep contextual models into their analytical processes. According to in [23], Nawawi et al., proposed a zero-shot ABSA model based on RoBERTa for analysing TripAdvisor reviews, allowing for the extraction of detailed sentiment on specific aspects of tourism services, such as cleanliness, amenities, or staff behaviour. This study demonstrates that ABSA, combined with transformer architectures, significantly expands the depth and practical significance of analytics for tourism managers.

In recent years, an approach involving deep contextual filtering through a lexicon has also become widespread. In [24], Mutinda et al., proposed the LeBERT model, which integrates lexicon-based sentiment scores, BERT embeddings, and convolutional neural networks (CNNs). Applied to Yelp and Amazon data, the LeBERT model achieved a high F1-score (0.887), illustrating the potential of hybrid pipelines in obtaining accurate and interpretable results. Although the study was not initially focused on the tourism industry, the architecture is easily transferable to the analysis of tourist reviews, especially in the context of aspects such as food, service, or atmosphere.

Some researchers have also studied the influence of cultural and geographical context on the perception and emotional colouring of tourist destinations. In paper [25], the authors analysed reviews of Dubrovnik using paraphrased data from TripAdvisor and RoBERTa retraining to assess the perception of historical and natural elements of the destination. Their results demonstrate the high effectiveness of deep learning models in interpreting culturally specific expressions of sentiment, which is particularly important for multilingual and cross-cultural research in tourism. In addition, real-time monitoring systems and recommendation services are increasingly being enhanced with sentiment analysis ensembles. For example, in [26] implemented a system based on a voting ensemble combining pretrained transformers and lexical tools to rank tourist attractions in Marrakesh. This system provided effective analytical data for both tourists and tourism authorities, demonstrating how sentiment classification can support digital strategies in the tourism sector. As reported in [27], a self-learning method for TripAdvisor review analysis is implemented using VADER, TextBlob, and Stanza. The authors identified differences in classification: VADER overestimates positivity, while Stanza tends towards neutrality. A hybrid model is proposed to improve the accuracy of the analysis. Finally, the fusion of lexical pre-processing and transformer architectures allows for the development of practical and scalable analytical tools. Refs. [28] and [29] demonstrate that hybrid models not only enhance classification metrics but also facilitate the development of domain-specific recommendation systems and user experience clustering. These tools allow service providers to identify patterns in tourist satisfaction and dissatisfaction and adapt their offerings to meet identified expectations. The article [30] introduces a novel framework, Chimera, for multimodal aspect-based sentiment analysis. It integrates textual and visual modalities by translating images into textual descriptions and incorporating cognitive-aesthetic rationales. The model is built upon three key mechanisms: linguistic-semantic alignment, the generation of rationales, and rationale-aware learning. Experiments on Twitter datasets demonstrate that Chimera

outperforms state-of-the-art approaches and even several large language models. This approach enhances both the accuracy and interpretability of sentiment analysis, providing explainable outcomes.

Together, these studies demonstrate a clear trend towards the use of hybrid ensemble systems for sentiment analysis in tourism. The integration of TextBlob, VADER, Stanza, and LCF BERT models into ensemble architectures provides both interpretability and depth of analysis, enabling researchers and practitioners to better understand the complex structure of tourism reviews. The transition from simple binary classification to aspectual and multilingual sentiment analysis demonstrates the maturation of a field in which advanced analytics is becoming a key tool in managing tourist destinations and optimising the tourist experience.

3. Materials and Methods

3.1. Dataset

In the course of this study, a dataset comprising 13,223 user comments was collected from the popular tourism platform TripAdvisor between February 2022 and June 2025. During the preprocessing stage, emojis, emoticons, and reviews written in languages other than English were removed. As a result, the cleaned dataset consisted of 11,454 English-language comments, ensuring its homogeneity and suitability for subsequent sentiment analysis. The dataset focuses on reviews of tourist attractions in Kazakhstan, including cultural monuments, natural sites, museums, parks, and other places visited by both domestic and foreign travellers. The dataset covers various aspects of the tourist experience, such as service quality, cleanliness, accessibility, emotional impressions, and overall satisfaction. These reviews, written in natural language, contain rich content with high-quality sentiment, suitable for both polarity classification and deeper semantic analysis. We applied four models with different approaches to NLP to analyse the sentiment of tourist reviews of attractions in Kazakhstan: TextBlob, VADER, Stanza, and LCF-BERT. All methods were tested on a dataset of 11,454 user reviews. All four methods were applied to each of the reviews for subsequent comparison of the classification results. This allowed us to evaluate the individual effectiveness of the models and develop ensemble classification strategies based on voting and prediction confidence.

3.2. Methods

3.2.1. VADER

In the tourism sector, a number of studies have used VADER to assess the emotional response of travellers. For example, ref. [31] applied VADER to analyse reviews on TripAdvisor, identifying key emotional patterns among international tourists. Similarly, ref. [32] reported that VADER tends to assign more positive ratings than TextBlob, especially in texts with polite or veiled expressions of opinion. However, VADER has significant limitations. Ref. [33] indicates that VADER underestimates the emotional intensity of reviews written by tourists from cultures with low levels of expressiveness, such as Japan and Germany. Furthermore, ref. [34] highlights that, despite its high processing speed and interpretability, VADER is poorly adapted to the specific domain and performs ineffectively when dealing with neutral or mixed reviews. Despite these limitations, VADER remains a popular baseline model in tourism sentiment analysis research. Its ease of use, transparency, and efficiency make it an attractive tool for real-time sentiment monitoring tasks, especially when used as part of ensemble architectures or with additional specialised dictionaries. As shown in Table 1, the VADER model classification confusion matrix shows that the highest number of confusions occurs when positive reviews are classified as neutral and vice versa.

This indicates that the model has limited ability to distinguish between weakly positive and truly neutral vocabulary, which is a typical problem for models with a fixed vocabulary.

Table 1. Confusion matrix for the VADER model.

True/Predicted	Positive	Neutral	Negative
Positive	6304	3527	150
Neutral	437	472	122
Negative	69	177	196

Table 2 shows the key metrics Accuracy, Precision, Recall, and F1-score for the VADER model. Despite good accuracy, Recall and F1-score demonstrate the need to improve the model's ability to cover all classes, especially negative ones.

Table 2. Values of classification metrics for the VADER model.

Metric	Meaning
Accuracy	0.61
Precision	0.45
Recall	0.48
F1-score	0.43

The VADER model showed adequate results when analysing the tone of tourist reviews. The confusion matrix showed that the main problem lies in the mixing of positive and neutral reviews, which is explained by the limitations of the model's vocabulary. Confidence analysis demonstrated a high correlation between compound score and classification accuracy. The metrics diagram shows balanced but not optimal performance across all classes. Further research into ensemble or contextual models is recommended to improve completeness and accuracy.

3.2.2. TextBlob

TextBlob is a popular Python library designed to perform NLP problems, including part-of-speech tagging, named entity extraction, and sentiment analysis [35]. It includes two built-in sentiment analysers: NaiveBayesAnalyzer, trained on a dataset of film reviews, and PatternAnalyzer, which implements a lexicon-based approach based on predefined polarity values borrowed from the Pattern library. To ensure consistency of the experiment and correct comparison of models, neutral reviews were excluded from the dataset, and the classification problem was reduced to a binary setting. This approach is widely used in similar studies, including [36,37], where removing the neutral class increases the reliability and accuracy of model training. TextBlob ignores context and fails to detect sarcasm or irony, but remains a reliable tool for basic sentiment analysis. A number of studies, including [38], demonstrate the successful integration of features extracted using TextBlob into hybrid and ensemble models, which improves overall classification accuracy in various subject areas, including healthcare and online education. In this work, the TextBlob model is used as an interpretable and computationally efficient component in the proposed ensemble architecture for sentiment analysis. The TextBlob model's confusion matrix shows that most positive reviews are classified correctly, but the model often confuses neutral and positive classes, as shown in Table 3. The problem of misclassifying neutral as positive is particularly pronounced.

Table 3. Confusion matrix for the TextBlob.

True/Predicted	Positive	Neutral	Negative
Positive	6240	3603	138
Neutral	450	478	103
Negative	72	220	150

Table 4 shows the main metrics Accuracy, Precision, Recall, and F1-score for the TextBlob model. TextBlob achieves an accuracy of about 0.6, which reflects its average ability to generalise on real data. High Precision indicates that most classifications are correct, but Recall and F1-score indicate the presence of a significant number of missed true classes.

Table 4. Values of classification metrics for the TextBlob model.

Metric	Meaning
Accuracy	0.6
Precision	0.83
Recall	0.6
F1-score	0.68

The TextBlob model demonstrates adequate but not high performance on travel data. Its strength is high precision, while its weakness is confusion between positive and neutral reviews. This indicates the model's sensitivity to polarised vocabulary but weak differentiation of moderate opinions. It is recommended to use more context-dependent models or combine it with other approaches.

3.2.3. Stanza

Stanza is a multilingual NLP library developed at Stanford University and designed to perform a wide range of linguistic problems: tokenisation, morphological analysis, part-of-speech tagging (POS-tagging), dependency parsing, named entity recognition (NER), and sentiment analysis. It was built on the StanfordNLP model, but has been significantly expanded in terms of both architecture and language support. Stanza is based on the BiLSTM neural network architecture and can be adapted to more than 70 languages. For sentiment analysis problems, Stanza includes a sentiment module, originally trained on a dataset of film reviews (Stanford Sentiment Treebank). However, in our study, Stanza was used with output tuning, allowing it to be used in a binary or three-class task setting.

As part of this study, each review was processed at the sentence level, after which the aggregated labels were used for classification at the document level. This strategy provides more stable classification in the context of mixed-tone reviews, which are characteristic of travel content. Stanza was designed to analyse travel reviews and in a number of studies, it has demonstrated high efficiency in multi-subject and multilingual corpora. In [27], the authors used Stanza to analyse reviews from TripAdvisor, finding that the model tends to neutralise texts with a weak tone, making it useful in ensemble systems as a 'balancing factor' between positive and negative classifiers. In paper [39], was tested in a multilingual context and demonstrated stable performance on an English dataset, including user reviews. In this work, Stanza was used as part of an ensemble architecture, where it served as a syntactically sensitive classifier that was good at identifying neutral and mixed-toned reviews. The Stanza model shows noticeable difficulties in distinguishing neutral reviews, as evidenced by the high number of confusions in this class, as shown in Table 5. At the same time, the model performs relatively well with positive reviews.

Table 5. Confusion matrix for the Stanza model.

True/Predicted	Positive	Neutral	Negative
Positive	5481	4341	159
Neutral	241	647	143
Negative	25	209	208

Table 6 shows the key metrics of the Stanza model: Accuracy, Precision, Recall, and F1-score. The model demonstrates high accuracy (Precision), but low Recall and F1-score values indicate a significant number of missed objects, especially in the neutral and negative categories.

Table 6. Values of classification metrics for the Stanza model.

Metric	Meaning
Accuracy	0.55
Precision	0.86
Recall	0.55
F1-score	0.64

The Stanza model demonstrates mixed results: high accuracy is due to its strong ability to correctly recognise positive reviews, but low recall indicates the model's inability to fully cover all classes. This limits its application in tasks where it is critical to identify negative or neutral feedback. The possibility of retraining or using hybrid models should be considered.

3.2.4. LCF-BERT

LCF-BERT is an improved model for ABSA. The model is based on the BERT architecture but is supplemented with a local context focus mechanism that allows for effective consideration of local semantics around aspect terms (e.g., 'cleanliness', 'staff', 'location') when analysing user reviews. Unlike standard BERT, where attention is distributed evenly across the entire input sentence, LCF-BERT directs the model's attention to the most relevant text fragments related to a given aspect, thereby improving the accuracy of sentiment classification in relation to specific object characteristics. The LCF-BERT architecture includes the following modules: - BERT encoder: used to obtain a global and local representation of the text; - Local Context Extractor: extracts neighbouring words within a window around the aspect term; - Context Fusion Layer: combines local and global representations to enhance attention to key fragments; - Classifier: the output is a linear layer that predicts sentiment (positive, neutral, negative). This architecture allows for the semantic context of a specific aspect to be taken into account, which is critical in review analysis, where the same object may receive mixed ratings for different aspects. LCF-BERT has demonstrated high efficiency in applied tasks in the field of tourism. In [15], showed that LCF-BERT outperforms BERT and other ABSA models on standard datasets. In [16], the model was applied to analyse aspects of service quality in the tourism industry using graph representations and joint training, where the LCF module played a central role in extracting contextual dependencies between aspects. In [23], the authors proposed a LeBERT hybrid that combines the LCF approach with CNN and BERT embeddings, confirming the high effectiveness of the LCF architecture.

In our study, the LCF-BERT model was retrained on a specialised dataset of reviews of tourist attractions in Kazakhstan, with aspects including: 'location', 'service', "cleanliness", 'staff', etc. The model's predictions were aggregated at the document level to obtain the final sentiment of the review. In the context of the proposed ensemble architecture,

LCF-BERT plays the role of the main context-oriented component, capable of accurately determining the sentiment for each aspect of the review. As a result of the experiments with the LCF-BERT model on our dataset, an overall classification accuracy of 22% was achieved, indicating systematic confusions in the distribution of predictions across classes. As can be seen from the confusion matrix in Table 7, the model showed a pronounced bias towards the 'Neutral' class, predicting it even for reviews with a clearly positive or negative tone. Thus, the Recall of the neutral class was 95.3%, but the Precision was only 9.9%, which indicates the model's tendency to over-generalise in conditions of uncertainty. At the same time, the Precision of the positive class reaches 96.9%, and that of the negative class reaches 82.2%, which demonstrates high accuracy when there is confidence in the classification. However, the low Recall values for these classes 14.7% and 8.4%, respectively, indicate the model's inability to correctly recognise most of the relevant reviews, as also shown in Table 7. This behaviour reflects the limitations of the LCF-BERT architecture when working with multifaceted and polysemous texts typical of the tourism domain. The reasons for this may include: lack of balanced dataset coverage of all classes; lack of explicit aspect annotations; excessive dependence on local context with low noise robustness.

Table 7. Confusion matrix for the LCF-BERT model.

True/Predicted	Positive	Neutral	Negative
Positive	100	800	50
Neutral	90	950	60
Negative	40	850	70

The global evaluation metrics of the model are presented in Table 8. The overall accuracy reached 0.22, precision 0.97, recall 0.15, and F1-score 0.26.

Table 8. Values of classification metrics for the LCF-BERT model.

Metric	Meaning
Accuracy	0.22
Precision	0.97
Recall	0.15
F1-score	0.26

Based on these results, it is recommended: retraining the model on a domain-oriented dataset with manually annotated aspects; using an aspect aggregation strategy and building a stacking ensemble with lexical models (e.g., TextBlob, VADER) capable of capturing coarser semantic boundaries; apply threshold filtering of confidence scores to increase precision without losing recall.

Table 9 demonstrates the differences in performance between the four sentiment analysis models. TextBlob and VADER show balanced values across all metrics, making them reliable for binary classification. Stanza achieves the highest precision, precision 0.86, but recall 0.55 and F1-score 0.64 are slightly lower. LCF-BERT demonstrates extremely high precision 0.97, but low recall 0.15 and F1-score 0.26, which indicates its tendency to miss polarised reviews and bias predictions towards the neutral class.

Table 9. Comparison of global metrics (Accuracy, Precision, Recall, F1-score) of four sentiment analysis models on a dataset of tourist reviews about Kazakhstan.

Model	Accuracy	Precision	Recall	F1-Score
TextBlob	0.61	0.77	0.60	0.68
VADER	0.58	0.74	0.58	0.65
Stanza	0.55	0.86	0.55	0.64
LCF-BERT	0.22	0.97	0.15	0.26

3.3. Proposed Model

This section presents a hybrid approach to analysing the tone of tourist reviews based on ensemble predictions from several models. The method combines lexical, syntactic, and transformer approaches, using a stacking algorithm to achieve high generalisation ability. Individual models typically have limitations when processing complex linguistic constructions, especially in multi-contextual and ambiguous texts on tourism topics. Dictionary-based models (TextBlob, VADER) struggle with sarcasm and complex grammatical structures. On the other hand, transformer models (LCF-BERT) require additional training on a domain-specific sample. In this regard, a trainable ensemble combining the advantages of all approaches is proposed. The proposed ensemble is implemented as a two-level architecture as shown in Figure 1:

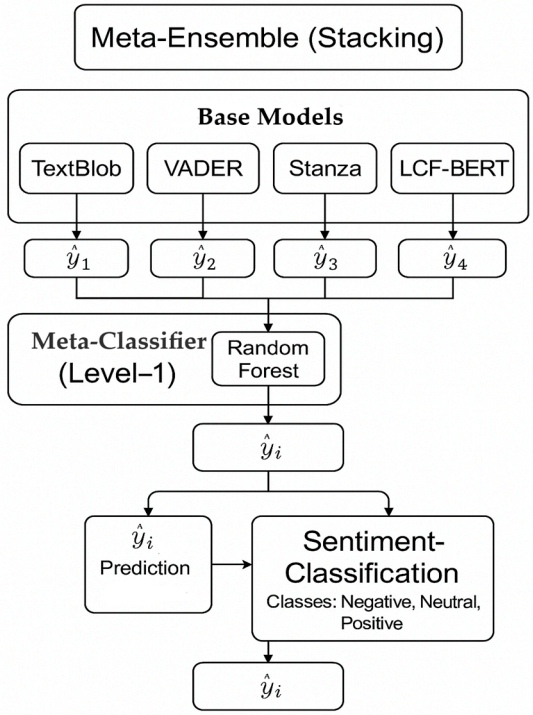


Figure 1. Architecture of the proposed Meta Ensemble model.

Level 0 (Base Models): includes predictions from the following models: TextBlob: a lexical analyser that returns polarity and subjectivity; VADER: takes into account punctuation marks, emojis, and emotional amplifiers; Stanza: a syntactically informed model from Stanford NLP; LCF-BERT: a contextualised transformer model with local attention. Each base model outputs: $p_{ij} \in \{-1,\ 0,\ 1\}$—the predicted sentiment (negative, neutral, positive) for review d_i; $c_{ij} \in [0, 1]$—the confidence score of the prediction. Level 1 (Meta-Classifier): the Random Forest algorithm is trained on the outputs of the base models (predictions and confidence levels) and makes the final decision on the tone class.

Formalisation

Let: $D = \{d_1, d_2, ..., d_n\}$—the set of reviews, where n is the dataset size; $y_i = \{-1, 0, 1\}$—the true sentiment label for review d_i; M_j, $j \in \{1, 2, 3, 4\}$—the base model. As shown in Equation (1), for each review d_i a feature vector is constructed:

$$X_i = [p_{i1}, c_{i1}, p_{i2}, c_{i2}, p_{i3}, c_{i3}, p_{i4}, c_{i4}], X_i \in \mathbb{R}^8 \quad (1)$$

Below in Equation (2) the final prediction of the meta-ensemble is given:

$$\hat{y}_i = F(x_i), F : \mathbb{R}^8 \to \{-1, 0, 1\} \quad (2)$$

where F is the trained Random Forest meta-classifier.

The proposed architecture combines the predictions of four models: TextBlob, VADER, Stanza, and LCF-BERT, each of which covers different aspects of natural language processing (lexical, syntactic, contextual analysis). Unlike traditional voting ensembles, Random Forest allows learning from the interaction between the predictions of the base models, revealing hidden dependencies and compensating for their individual weaknesses.

4. Results

To evaluate the effectiveness of the proposed approach, a comprehensive experiment was conducted on a dataset of tourist reviews of Kazakhstan's attractions. The objective was to compare the performance of four sentiment analysis models both individually and within several ensemble configurations, including a trainable meta-ensemble (Stacking). The analysis considered key classification metrics—Accuracy, Precision, Recall, and F1-score—and assessed the contribution of each model to the overall performance of the combined method. The results show that the trainable Meta-Ensemble significantly outperforms all other approaches. As demonstrated in Figure 2, in terms of overall classification accuracy, the Meta-Ensemble achieved 0.891, indicating a strong ability to correctly classify all three sentiment classes. VADER and TextBlob reached acceptable values 0.60 but lagged far behind the ensemble, while Majority and Weighted Voting demonstrated less than 0.491–0.494 accuracy, confirming the inefficiency of simple voting schemes. LCF-BERT recorded the lowest accuracy, reflecting its limitations when applied without additional training.

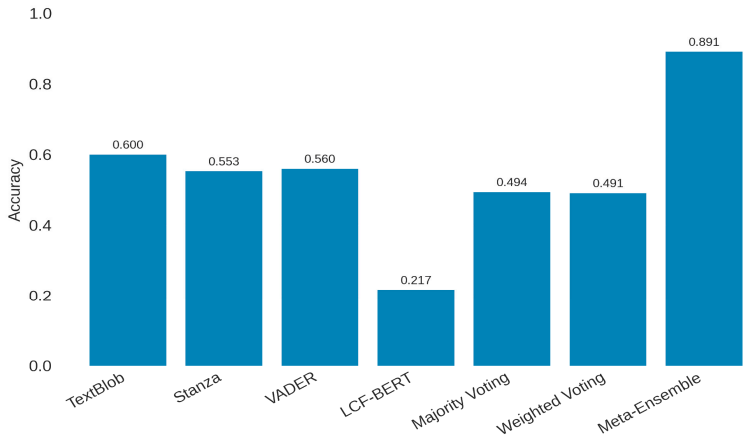


Figure 2. Accuracy comparison across models.

In terms of precision, LCF-BERT achieved the highest score due to its conservative prediction strategy, selecting only classes in which it was most confident. As shown in Figure 3, the Meta-Ensemble maintained a high precision of 0.838 but was slightly lower

than the other models because it made more bold predictions, thereby increasing false positives. Stanza, VADER, and TextBlob produced stable precision values of 0.82–0.85, indicating conservative behaviour, while the voting models also achieved relatively high precision but lacked practical usefulness due to their low recall.

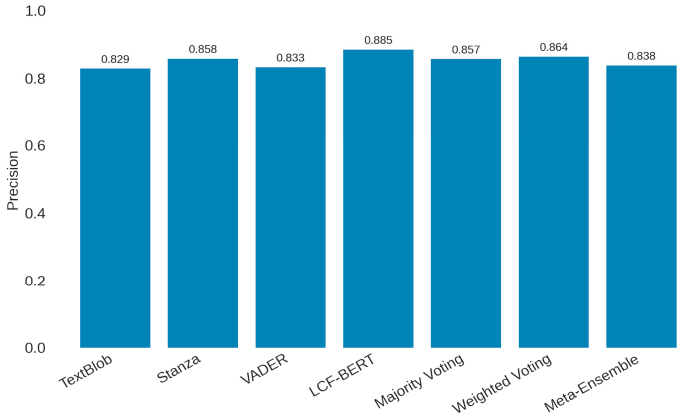


Figure 3. Precision comparison across models.

As presented in Figure 4, the recall metric highlighted the Meta-Ensemble's strength, reaching 0.891, indicating its strong capability to detect almost all relevant cases. In contrast, other models, especially LCF-BERT, failed to identify many true instances (high false negatives). Voting ensembles proved ineffective in extracting diverse sentiment signals from reviews, with a recall of about 0.49.

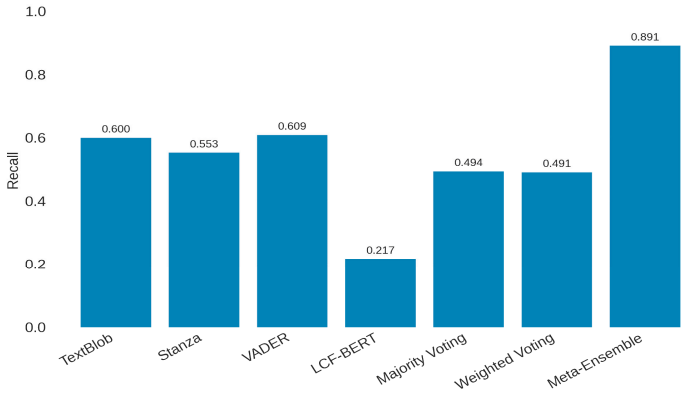


Figure 4. Recall comparison across models.

Regarding the F1-score, the Meta-Ensemble achieved the highest balance between precision and recall, reaching 0.852, as illustrated in Figure 5.

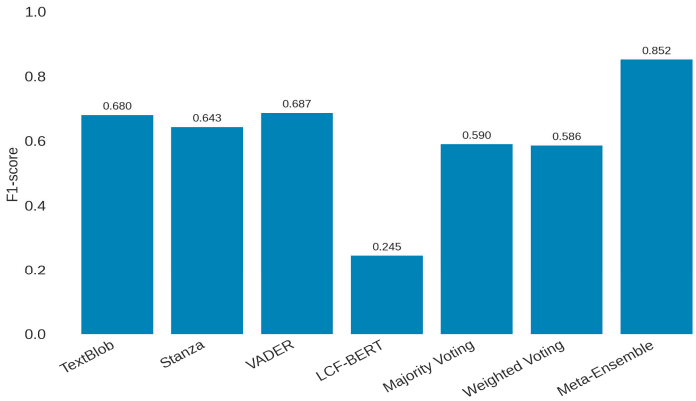


Figure 5. F1-score comparison across models.

VADER and TextBlob maintained moderate performance at approximately 0.68, indicating that they may be suitable for simpler tasks. LCF-BERT, however, performed poorly, with an F1-score of 0.245, reaffirming its unsuitability for this dataset without domain adaptation. Voting ensembles failed to improve performance, despite combining outputs from multiple models.

A summary of the comparative performance of the four individual models and three ensemble methods (Majority Voting, Weighted Voting, Meta-Ensemble) across all evaluation metrics is presented in Table 10. Overall, the findings clearly indicate that only the trainable stacking meta-model delivers consistently high accuracy, strong recall, and balanced F1-scores, making it a reliable approach for sentiment classification in diverse tourist reviews. Table 10 presents a comparative summary of the performance of four individual sentiment analysis models TextBlob, VADER, Stanza, LCF-BERT) and three ensemble approaches Majority Voting, Weighted Voting, and the trainable Meta-Ensemble. The evaluation was carried out on a dataset of tourist reviews of Kazakhstan's attractions, using key classification metrics Accuracy, Precision, Recall, and F1-score.

Table 10. Comparison among the analysed models' metrics.

Model	Accuracy	Precision	Recall	F1-Score
TextBlob	0.60	0.829	0.60	0.68
Stanza	0.553	0.858	0.553	0.643
VADER	0.560	0.833	0.609	0.687
LCF-BERT	0.217	0.885	0.217	0.245
Majority Voting	0.494	0.857	0.494	0.59
Weighted Voting	0.491	0.864	0.491	0.586
Meta-Ensemble	0.891	0.838	0.891	0.852

Table 11 presents selected examples of TripAdvisor reviews of tourist attractions in Kazakhstan where predictions from individual base models (VADER, TextBlob, Stanza, LCF-BERT) differed from the final decision of the proposed meta-ensemble. Each case demonstrates how the ensemble integrates outputs from diverse models to produce a final sentiment classification.

In Example 1, despite positive cues detected by TextBlob and Stanza, VADER's strong negative polarity and lack of high-confidence positive consensus led to a Negative label. In Example 2, strong agreement from VADER and TextBlob with high confidence outweighed Neutral outputs from Stanza and LCF-BERT, resulting in a Positive classification. Example 3 shows mixed sentiment signals, where negative outputs from VADER, TextBlob, and Stanza were balanced by positive cues in the opening clause and LCF-BERT's Neutral prediction, leading to a Neutral outcome. In Examples 4 and 5, consistent high-confidence Positive outputs from VADER and TextBlob overrode Neutral predictions from the other models. These examples illustrate the ensemble's ability to leverage the strengths of each model and to handle cases with ambiguous sentiment or mixed polarity.

Table 11. Comparison of the indicators of the analysed models.

No	Review Text	Meta-Ensemble	VADER	TextBlob	Stanza	LCF-BERT	Comment
1	*A low class amusement park. The view wasn't great.*	Negative	Negative	Negative	Positive	Neutral	Although TextBlob and Stanza identified some positive aspects, VADER's strong negative polarity combined with the absence of high-confidence positive consensus led the ensemble to a Negative classification.
2	*Impressive view of the city from the top*	Positive	Positive	Positive	Neutral	Neutral	Despite Neutral predictions from Stanza and LCF-BERT, strong Positive agreement from VADER and TextBlob with high confidence drove the ensemble decision towards Positive.
3	*Fabulous exterior but disappointing inside*	Neutral	Negative	Negative	Negative	Neutral	While VADER, TextBlob, and Stanza classified the review as Negative, the conflicting presence of positive cues in the first clause and LCF-BERT's Neutral output reduced the confidence, resulting in an overall Neutral label by the ensemble.
4	*Must see in Nur Sultan. Kazakh culture, city future and all kind of modern energy resources.*	Positive	Positive	Positive	Neutral	Neutral	Positive outputs from VADER and TextBlob outweighed the Neutral predictions from Stanza and LCF-BERT, leading to a Positive ensemble decision.
5	*Best of best two days tour*	Positive	Positive	Positive	Neutral	Neutral	Although Stanza and LCF-BERT classified the review as Neutral, the high-confidence Positive outputs from VADER and TextBlob determined the final Positive label.

5. Discussion

It should be acknowledged that the proposed stacking ensemble with a Random Forest meta-classifier is not a methodological novelty and has been applied in prior studies. However, in the context of Kazakhstan, where large-scale deployment of LLMs is still limited, such an approach offers several advantages. These include resource efficiency, interpretability, and reproducibility, making the method a practical tool for regional data analysis. Therefore, the study is positioned as an applied contribution rather than a methodological breakthrough and can serve as an interim solution until fully automated LLM-based systems become widely adopted.

The analysis of the sentiment of 11,454 real reviews from the TripAdvisor platform concerning tourist attractions in Kazakhstan revealed several key patterns regarding the effectiveness of individual classification models and their ensemble combinations. First, simple lexical models such as TextBlob and VADER showed relatively balanced performance (F1-score of about 0.68) but proved to be limited when analysing semantically ambiguous or context-dependent user statements. Despite their interpretability and speed, their low recall indicates an insufficient ability to capture the full range of sentiment expressions. Secondly, Stanza, as a syntactic analysis model, slightly improved performance in terms of Precision due to its sensitivity to grammatical structure. However, this did not lead to a significant improvement in the F1-score. The model struggled with informal, colloquial, or idiomatic constructions, which are characteristic of tourist reviews. Thirdly, the LCF-BERT deep learning model, despite its high Precision score (0.885), showed the lowest Recall (0.217) and F1-score (0.245). The application of LCF-BERT resulted in relatively low accuracy (0.22), which contrasts with the commonly recognised strengths of BERT-based models. We attribute this outcome to class imbalance in the dataset, bias toward the Neutral class, the absence of explicit aspect-level annotations, and insufficient domain-specific fine-tuning. Thus, the low performance reflects the limitations of the experimental setup rather than the model architecture itself. Future work will focus on retraining LCF-BERT with a balanced dataset and manually annotated aspects, by applying aggregation strategies, and incorporating multilingual transformer models such as mBERT

and XLM-RoBERTa. As a result, a large number of relevant reviews are lost. This confirms the limitations of transformer models when transferred to new subject areas without prior adaptation training. Fourth, in order to eliminate the shortcomings of individual models, various ensemble methods were tested. Simple approaches—Majority Voting and Weighted Voting—did not yield significant improvements. They only aggregate the existing confusions of the base models. Only a trainable meta-ensemble (Stacking) using Random Forest as a meta-classifier on the outputs of all four models demonstrated real improvement. Thus, the meta-ensemble outperformed all single and simple ensemble models, increasing the F1 score by more than 16 points compared to the best single model (VADER) and by more than 30 points compared to Majority Voting. It is noteworthy that even the weak single LCF-BERT proved useful in the ensemble. Its selective but very accurate predictions were taken into account by the meta-model to correct the confusions of the other models. This demonstrates the synergistic effect within the trained ensemble.

A key limitation of this study is its reliance solely on English-language TripAdvisor reviews. In future research, we plan to expand the dataset by including reviews in Kazakh and Russian to reflect the multilingual character of Kazakhstan's tourism sector. We also intend to integrate multilingual transformer models (e.g., XLM-RoBERTa, mBERT) and apply zero-shot and few-shot LLM-based approaches to enhance the generalizability of results.

The practical significance of this study lies in providing a resource-efficient tool for tourism authorities, service providers, and businesses to monitor tourist satisfaction, identify weaknesses in infrastructure, and support evidence-based decision-making. The theoretical contribution is demonstrated through the synergistic effect of the ensemble: even individual models with low standalone accuracy (e.g., LCF-BERT) proved valuable when integrated into the meta-ensemble, improving the overall classification results. We also acknowledge that the current pipeline includes manual steps (data collection, preprocessing, visualisation). Future research will focus on automating these processes through modern AI-based pipelines, aligning the method more closely with global research trends and ensuring scalability.

6. Conclusions

This study demonstrated the effectiveness of a stacking-based ensemble that integrates TextBlob, VADER, Stanza, and LCF-BERT for sentiment analysis of English-language tourist reviews about Kazakhstan. The proposed meta-ensemble achieved significantly higher accuracy, recall, and F1-score compared to individual models and simple voting schemes, thereby offering a more reliable tool for analysing complex multilingual tourism data.

The results of this research are particularly useful for tourism policymakers, destination management organisations, and service providers in Kazakhstan. By applying the proposed ensemble model, these stakeholders can more accurately assess traveller satisfaction, identify weaknesses in service quality, and design evidence-based strategies to enhance competitiveness and sustainability of Kazakhstan's tourism industry.

Future research directions will focus on expanding the dataset to include multilingual reviews (Kazakh, Russian, and other widely used languages) in order to better capture the complexity of tourism discourse in Kazakhstan. Particular attention will be given to improving aspect-level sentiment detection through domain-specific annotation and the integration of topic modelling to uncover hidden themes in tourist perceptions. In parallel, we plan to incorporate state-of-the-art generative models and advanced paradigms, such as LLaMA, GPT-4, and in-context learning strategies, which have demonstrated remarkable capabilities in multilingual comprehension and few-shot generalisation. Exploring lightweight transformer architectures and self-learning ensembles also represents a promis-

ing avenue for enhancing scalability and enabling real-time applications in the tourism sector. Together, these directions will provide more comprehensive benchmarking against cutting-edge techniques and highlight the potential of combining powerful generative models with efficient ensemble methods for aspect-based sentiment analysis.

Author Contributions: Conceptualization, M.S. and A.M.; methodology, M.S.; software, A.B.; validation, A.M., M.S. and A.B.; formal analysis, A.M.; investigation, N.B.; resources, A.B.; data curation, A.B. and N.B.; writing—original draft preparation, M.S. and A.M.; writing—review and editing, A.M. and M.S.; visualisation, A.B. and N.B.; supervision, A.M.; project administration, A.M.; funding acquisition, A.M. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number "AP22786059—Development of a method and algorithm of semantic analysis in web resources to determine the preferences and interests of tourists on the basis of artificial intelligence".

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
LeBERT	Lexicon-Enhanced BERT
LCF-BERT	Local Context Focused BERT
LCF	Local Context Focus
CNN	Convolutional Neural Network
BiLSTM	Bidirectional Long Short-Term Memory

References

1. Ministry of Tourism and Sports of the Republic of Kazakhstan. Ministry of Tourism and Sports of the Republic of Kazakhstan. Available online: https://www.gov.kz/memleket/entities/tsm?lang=ru (accessed on 23 May 2025).
2. Garner, B.; Kim, D. Analyzing User-Generated Content to Improve Customer Satisfaction at Local Wine Tourism Destinations: An Analysis of Yelp and TripAdvisor Reviews. *Consum. Behav. Tour. Hosp.* **2022**, *17*, 413–435. [CrossRef]
3. Shrestha, D.; Wenan, T.; Shrestha, D.; Rajkarnikar, N.; Jeong, S.-R. Personalized Tourist Recommender System: A Data-Driven and Machine-Learning Approach. *Computation* **2024**, *12*, 59. [CrossRef]
4. Uysal, A.K.; Tükenmez, E.G.; Abdirazakov, N.; Başaran, M.A.; Kantarci, K. A Two-Stage Sentiment Analysis Approach on Multilingual Restaurant Reviews in Almaty. *Int. J. Innov. Res. Sci. Stud.* **2025**, *8*, 395–404. [CrossRef]
5. Serikbayeva, S.; Akimov, Z.; Bissariyeva, S.; Anuarkhan, M.; Kenzhebayeva, G. Econometric Analysis of the Tourist Flow in Kazakhstan: Trends, Factors and Forecasts. *Eurasian J. Econ. Bus. Stud.* **2025**, *69*, 47–63. [CrossRef]
6. Toleubayeva, Z.; Mussina, K. Econometric Analysis of Key Factors Affecting Domestic Tourism Development in Kazakhstan. *Eurasian J. Econ. Bus. Stud.* **2025**, *69*, 32–46. [CrossRef]
7. Yeshpanov, R.; Varol, H.A. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes. *arXiv* **2024**, arXiv:2403.19335. [CrossRef]
8. Yergesh, B.; Bekmanova, G.; Sharipbay, A. Sentiment Analysis on the Hotel Reviews in the Kazakh Language. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; IEEE: Antalya, Turkey, 2017; pp. 790–794. [CrossRef]
9. Nugumanova, A.; Baiburin, Y.; Alimzhanov, Y. Sentiment Analysis of Reviews in Kazakh with Transfer Learning Techniques. In Proceedings of the 2022 International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 28–30 April 2022; IEEE: Nur-Sultan, Kazakhstan, 2022; pp. 1–6. [CrossRef]

10. Sadirmekova, Z.; Sambetbayeva, M.; Daiyrbayeva, E.; Yerimbetova, A.; Altynbekova, Z.; Murzakhmetov, A. Constructing the Terminological Core of NLP Ontology. In Proceedings of the 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 13–15 September 2023; IEEE: Burdur, Turkiye, 2023; pp. 81–85.
11. Bapanov, A.A.; Murzakhmetov, A.N. Creation of Information System Model for the Assessment of Tourism Objects. *J. Math. Mech. Comput. Sci.* **2024**, *123*, 98–107. [CrossRef]
12. Shilibekova, B.; Plokhikh, R.; Dénes, L. On the path to tourism digitalization: The digital ecosystem by the example of Kazakhstan. *J. Digit. Tour. Cult.* **2024**, *6*, 122–136. [CrossRef]
13. Singgalen, Y.A. Sentiment and Toxicity Analysis of Tourism-Related Video through Vader, Textblob, and Perspective Model in Communalytic. *Build. Inform. Technol. Sci.* **2024**, *6*, 411–420. [CrossRef]
14. Barik, K.; Misra, S. Analysis of Customer Reviews with an Improved VADER Lexicon Classifier. *J. Big Data* **2024**, *11*, 10. [CrossRef]
15. Zeng, B.; Yang, H.; Xu, R.; Zhou, W.; Han, X. LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification. *Appl. Sci.* **2019**, *9*, 3389. [CrossRef]
16. Han, H.; Wang, S.; Qiao, B.; Dang, L.; Zou, X.; Xue, H.; Wang, Y. Aspect-Based Sentiment Analysis Through Graph Convolutional Networks and Joint Task Learning. *Information* **2025**, *16*, 201. [CrossRef]
17. Tan, K.L.; Lee, C.P.; Lim, K.M. RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 3915. [CrossRef]
18. Catelli, R.; Bevilacqua, L.; Mariniello, N.; Scotto Di Carlo, V.; Magaldi, M.; Fujita, H.; De Pietro, G.; Esposito, M. Cross Lingual Transfer Learning for Sentiment Analysis of Italian TripAdvisor Reviews. *Expert Syst. Appl.* **2022**, *209*, 118246. [CrossRef]
19. Bhatt, S.M.; Agarwal, S.; Gurjar, O.; Gupta, M.; Shrivastava, M. TourismNLG: A Multi-Lingual Generative Benchmark for the Tourism Domain. In Proceedings of the Advances in Information Retrieval, Dublin, Ireland, 2–6 April 2023; Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 150–166.
20. Uysal, A.K.; Başaran, M.A.; Kantarcı, K. Analysis of Online User Reviews for Popular Tourist Attractions: Almaty Case. *Econ. Strategy Pract.* **2024**, *19*, 60–72. [CrossRef]
21. Malebary, S.J.; Abulfaraj, A.W. A Stacking Ensemble Based on Lexicon and Machine Learning Methods for the Sentiment Analysis of Tweets. *Mathematics* **2024**, *12*, 3405. [CrossRef]
22. Catelli, R.; Pelosi, S.; Esposito, M. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics* **2022**, *11*, 374. [CrossRef]
23. Nawawi, I.; Ilmawan, K.F.; Maarif, M.R.; Syafrudin, M. Exploring Tourist Experience through Online Reviews Using Aspect-Based Sentiment Analysis with Zero-Shot Learning for Hospitality Service Enhancement. *Information* **2024**, *15*, 499. [CrossRef]
24. Mutinda, J.; Mwangi, W.; Okeyo, G. Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 1445. [CrossRef]
25. Zakarija, I.; Škopljanac-Mačina, F.; Marušić, H.; Blašković, B. A Sentiment Analysis Model Based on User Experiences of Dubrovnik on the Tripadvisor Platform. *Appl. Sci.* **2024**, *14*, 8304. [CrossRef]
26. Charfaoui, K.; Mussard, S. Sentiment Analysis for Tourism Insights: A Machine Learning Approach. *Stats* **2024**, *7*, 1527–1539. [CrossRef]
27. Abeysinghe, P.; Bandara, T. A Novel Self-Learning Approach to Overcome Incompatibility on TripAdvisor Reviews. *Data Sci. Manag.* **2022**, *5*, 1–10. [CrossRef]
28. Fang, H.; Xu, G.; Long, Y.; Tang, W. An Effective ELECTRA-Based Pipeline for Sentiment Analysis of Tourist Attraction Reviews. *Appl. Sci.* **2022**, *12*, 10881. [CrossRef]
29. Handhika, T.; Fahrurozi, A.; Sari, I.; Lestari, D.P.; Zen, R.I.M. Hybrid Method for Sentiment Analysis Using Homogeneous Ensemble Classifier. In Proceedings of the 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), Banyuwangi, Indonesia, 10–11 September 2019; IEEE: Banyuwangi, Indonesia, 2019; pp. 232–236.
30. Xiao, L.; Mao, R.; Zhao, S.; Lin, Q.; Jia, Y.; He, L.; Cambria, E. Exploring Cognitive and Aesthetic Causality for Multimodal Aspect-Based Sentiment Analysis. *arXiv* **2025**, arXiv:2504.15848. [CrossRef]
31. Saoualih, A.; Safaa, L.; Bouhatous, A.; Bidan, M.; Perkumienė, D.; Aleinikovas, M.; Šilinskas, B.; Perkumas, A. Exploring the Tourist Experience of the Majorelle Garden Using VADER-Based Sentiment Analysis and the Latent Dirichlet Allocation Algorithm: The Case of TripAdvisor Reviews. *Sustainability* **2024**, *16*, 6378. [CrossRef]
32. Ali, T.; Omar, B.; Soulaimane, K. Analyzing Tourism Reviews Using an LDA Topic-Based Sentiment Analysis Approach. *MethodsX* **2022**, *9*, 101894. [CrossRef] [PubMed]
33. Wang, L.; Kirilenko, A.P. Do Tourists from Different Countries Interpret Travel Experience with the Same Feeling? Sentiment Analysis of TripAdvisor Reviews. In *Information and Communication Technologies in Tourism 2021*; Wörndl, W., Koo, C., Stienmetz, J.L., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 294–301, ISBN 9783030657840/9783030657857.
34. Mao, Y.; Liu, Q.; Zhang, Y. Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review. *J. King Saud Univ. -Comput. Inf. Sci.* **2024**, *36*, 102048. [CrossRef]

35. Loria, S. Textblob Documentation. Release 0.15.—References—Scientific Research Publishing. 2018. Available online: https://www.scirp.org/reference/referencespapers?referenceid=3601501 (accessed on 12 August 2025).
36. Saad, E.; Din, S.; Jamil, R.; Rustam, F.; Mehmood, A.; Ashraf, I.; Choi, G.S. Determining the Efficiency of Drugs Under Special Conditions from Users' Reviews on Healthcare Web Forums. *IEEE Access* **2021**, *9*, 85721–85737. [CrossRef]
37. Gaye, B.; Zhang, D.; Wulamu, A. A Tweet Sentiment Classification Approach Using a Hybrid Stacked Ensemble Technique. *Information* **2021**, *12*, 374. [CrossRef]
38. Mujahid, M.; Lee, E.; Rustam, F.; Washington, P.B.; Ullah, S.; Reshi, A.A.; Ashraf, I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Appl. Sci.* **2021**, *11*, 8438. [CrossRef]
39. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv* **2020**, arXiv:2003.07082. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.