International Journal of Innovative Research and Scientific Studies, 8(3) 2025, pages: 395-404

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

A two-stage sentiment analysis approach on multilingual restaurant reviews in Almaty

Alper Kürşat Uysal[1*], Egemen Güneş Tükenmez [2], Nurzhan Abdirazakov [3], Murat Alper Başaran[4], Kemal Kantarci[2,5]

[1]*Alanya Alaaddin Keykubat University, Rafet Kayis Faculty of Engineering, Computer Engineering Department, Alanya-Antalya/Türkiye.*
[2]*Alanya Alaaddin Keykubat University, Tourism Faculty, Tourism Management Department Alanya-Antalya/Türkiye.*
[3]*International University of Tourism and Hospitality, Turkestan, Kazakhstan.*
[4]*Alanya Alaaddin Keykubat University, Rafet Kayis Faculty of Engineering, Industrial Engineering Department, Alanya-Antalya/Türkiye.*
[5]*Hoca Ahmet Yesevi International Turkish-Kazakh University, Faculty of Economics, Administrative Sciences and Law, Department of Management and Tourism, Turkistan / Kazakhstan.*

Corresponding author: Alper Kürşat Uysal (*Email: alper.uysal@alanya.edu.tr*)

Abstract

Sentiment classification has become one of the most widely studied areas in text classification, especially in recent years. This study presents extensive experiments in sentiment analysis, investigating the performance of seven state-of-the-art sentiment analyzers (TextBlob, VADER, AFINN, Stanza, Nlptown, Sentistrength, and Flair) in Stage 1, and an ensemble approach in Stage 2, using multilingual restaurant reviews from Almaty, Kazakhstan. The reviews, either originally written in English or translated from Russian, are analyzed across various sections, including HEAD, TEXT, and their combinations (HEAD+TEXT). The results of Stage 2 ensemble methods demonstrate clear advantages of carefully selected ensembles over individual sentiment analyzers. Specifically, the highest Micro-F1 score for English reviews was 0.733 in the TEXT section, achieved by the ensemble TextBlob+Stanza+Nlptown+Sentistrength. The highest Macro-F1 score for English reviews was 0.684, achieved by the same ensemble in the TEXT section. For Russian reviews, the highest Micro-F1 score was 0.703 in the HEAD+TEXT combination, and the highest Macro-F1 score was 0.642 in the TEXT section, both achieved by the ensemble TextBlob+Stanza+Nlptown+Sentistrength. These findings highlight that the performance of sentiment analyzers varies depending on the original language and the corresponding review section.

Keywords: Ensemble methods, Kazakhstan, Multilingual data, Restaurant reviews, Sentiment analysis.

DOI: 10.53894/ijirss.v8i3.6512
Funding: This study was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan through (Grant Number: AP19679687 between 2023–2025).
History: Received: 20 March 2025 / **Revised:** 21 April 2025 / **Accepted:** 23 April 2025 / **Published:** 25 April 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

With the increasing prevalence of online platforms, sentiment analysis has emerged as a vital subfield of text classification [1]. As more individuals share their opinions online, particularly through reviews, this user-generated content significantly influences the decisions of others in areas such as purchasing products or choosing services. In this context, sentiment analysis refers to the automated process of detecting and categorizing opinions expressed in text, typically as positive, negative, or neutral.

In the food and beverage industry, alongside the broader hospitality sector, understanding customer sentiment is essential for attracting clientele and maximizing revenue. Restaurants strive to enhance their reputation by offering quality service, diverse menu options, and competitive pricing. Positive customer experience contributes not only to a stronger brand image but also to financial growth. However, capturing and interpreting the nuanced feedback of customers, ranging from praise to complaints, remains a challenge. This is where sentiment analysis becomes particularly valuable, offering data-driven insights into customer satisfaction, service quality, and overall experience.

Given the importance of such insights, sentiment analysis has gained traction in extracting actionable knowledge from restaurant reviews. Platforms like TripAdvisor have become rich sources of user feedback, enabling researchers and businesses to analyze customer opinions on a scale. Several studies have focused on sentiment analysis of restaurant reviews using data from major cities and in commonly used languages like English.

However, limited attention has been given to Central Asian countries, specifically Kazakhstan, the Kyrgyz Republic, Tajikistan, Turkmenistan, and Uzbekistan, despite their growing tourism and hospitality sectors [2]. In Kazakhstan, particularly on TripAdvisor, Russian is one of the dominant languages alongside English. Yet, sentiment analysis studies focusing on multilingual data from this region remain scarce, and little is known about how sentiment analyzers perform across different languages in this context.

To address this research gap, this study investigates the effectiveness of seven state-of-the-art sentiment analyzers with different characteristics, along with ensemble approaches that combine their outputs, using restaurant reviews from Almaty, Kazakhstan. The dataset comprises original English reviews and English translations of Russian-language reviews. Unlike previous works, this study evaluates both individual and combined sentiment analysis methods, comparing their performance across two languages in a region that is largely underexplored in existing literature.

This study aims to answer the following research questions:

1. What is the highest accuracy achieved by sentiment analyzers on English reviews and translated Russian reviews for restaurants in Almaty, Kazakhstan?
2. Do the most successful sentiment analyzers differ between English reviews and reviews translated from Russian?
3. Can ensemble approaches improve the accuracy of sentiment analysis for multilingual restaurant reviews?

The remainder of this article is structured as follows: Section 2 presents a review of relevant literature. Section 3 describes the methodology, including the sentiment analyzers and ensemble techniques used. Section 4 reports the experimental results. Finally, Section 5 concludes with key findings and implications for future research.

2. Literature Review

The traditional concept of Word of Mouth (WOM), once characterized by face-to-face exchanges of experiences and opinions, has evolved into its digital form: electronic Word of Mouth (eWOM) [2, 3]. Through eWOM, consumers share both positive and negative feedback online, significantly influencing not only potential customers but also business strategies. This communication is particularly impactful in service sectors such as the restaurant industry, where consumer experiences are intangible and highly subjective. As a result, consulting online reviews before dining out has become routine Gao et al. [4]. Litvin et al. [5] further emphasized how online reviews influence consumer behavior in the food and beverage sector.

In tandem with the rise of eWOM, sentiment analysis has gained prominence as a computational approach to extracting and classifying opinions from textual data. Although sentiment analysis has matured over the past two decades, it continues to evolve with the emergence of new methodologies and applications. It generally involves identifying the polarity—positive, negative, or neutral—of textual content. For customer-driven industries like hospitality and dining, understanding consumer sentiment offers critical insights into satisfaction levels, service quality, and areas needing improvement.

Numerous studies have addressed sentiment analysis on restaurant reviews, ranging from traditional machine learning to ensemble approaches:

Kang et al. [6] developed a domain-specific sentiment lexicon for restaurant reviews, observing a classification bias where positive sentiments were identified more accurately than negative ones using standard supervised learning algorithms. To reduce this imbalance, they proposed an enhanced Naïve Bayes classifier incorporating unigram and bigram features. The results showed a substantial reduction in accuracy disparity, down to 3.6% compared to standard Naïve Bayes and 28.5% relative to SVM, alongside notable gains in recall and precision.

Laksono et al. [7] focused on customer satisfaction analysis in Surabaya restaurants using Naïve Bayes and TextBlob. Data collected via WebHarvy showed Naïve Bayes outperforming TextBlob by 2.94%, with an overall accuracy of 72.06%.

Sharif et al. [8] developed a sentiment classification system for Bengali-language restaurant reviews. Their model, based on the multinomial Naïve Bayes algorithm, achieved an accuracy of 80.48% on a dataset of 1,000 reviews.

Adnan et al. [9] used TripAdvisor data to assess customer satisfaction via sentiment classification. Utilizing a Decision Tree (J48) algorithm, their model attained an accuracy of 45.6%, with a precision of 48.7%, a recall of 36.8%, and an F1-score of 41.4%. The relatively modest performance indicates challenges in handling unstructured English-language reviews even on widely used platforms.

Burra and Mishra [10] compared Logistic Regression and SVM for sentiment classification using 1,001 restaurant reviews from Kaggle. While both algorithms performed well, SVM slightly outperformed Logistic Regression with an accuracy of 76.80% versus 76.40%.

Al-Qudah et al. [11] proposed a comprehensive model for analyzing customer sentiment in Jordan's food service sector. Their method combined Extreme Gradient Boosting (XGBoost) with Particle Swarm Optimization (PSO) within an ordinal regression framework to address sentiment misclassification. The PSO-XGB approach achieved a lower RMSE (0.7722) than PSO-SVM (0.9988), highlighting its effectiveness, particularly for multilingual sentiment prediction tasks in Arabic.

In addition to individual classifiers, ensemble methods have been explored to improve sentiment classification performance. For instance, Saleena [12] proposed an ensemble model combining multiple base classifiers for Twitter sentiment analysis. Their method outperformed individual classifiers and majority voting, with F1-scores ranging from 70.28 to 76.85 across configurations.

Similarly, Kazmaier and Van Vuuren [13] advocated for ensemble approaches to mitigate individual model weaknesses. They evaluated multiple ensemble methods on benchmark datasets, including movie and business reviews, and reported median improvements of up to 5.53% over the best-performing single model.

While these studies collectively contribute to the advancement of sentiment analysis in the restaurant domain, most are limited to monolingual datasets or focus on regions with well-established digital infrastructures and predominantly English-language content. Notably, there is a relative scarcity of research that investigates sentiment classification in multilingual or underrepresented regions such as Central Asia, particularly Kazakhstan, where both Russian and English are widely used. Furthermore, although ensemble learning has shown promise in general sentiment analysis tasks, there remains a limited number of studies that apply ensemble-based methods specifically to sentiment analysis of restaurant reviews. Even fewer have compared the performance of individual sentiment analyzers and their combinations across multiple languages. This study aims to bridge these gaps by exploring both individual and ensemble sentiment analysis approaches on bilingual restaurant review data.

3. Methodology

This study adopts a two-stage sentiment analysis framework utilizing seven state-of-the-art sentiment analyzers. In Stage 1, each analyzer is applied independently to evaluate its standalone performance. In Stage 2, ensemble learning is employed through majority voting to improve overall classification accuracy.

Each of the seven sentiment analyzers provides a sentiment label or score for a given input, enabling the comparison of their individual and combined effectiveness. Details of each sentiment analyzer are presented in the next subsection.

3.1. Stage 1: Individual Sentiment Analyzers

The seven sentiment analyzers used in Stage 1—TextBlob, VADER, AFINN, Stanza, Nlptown, Sentistrength, and Flair—comprise a mix of four lexicon-based approaches (TextBlob, VADER, AFINN, and Sentistrength) and three machine learning-based models (Stanza, Nlptown, and Flair), providing a balanced perspective that leverages both rule-based precision and the contextual understanding of deep learning.

3.1.1. TextBlob

TextBlob is a widely used Python library for lexicon-based sentiment analysis and general natural language processing (NLP) tasks [14]. Built on top of NLTK and Pattern, it offers a user-friendly API ideal for both beginners and experienced developers. In addition to sentiment analysis, TextBlob supports tasks such as POS tagging, tokenization, noun phrase extraction, and translation [14, 15].

For sentiment classification, TextBlob uses a rule-based system with polarity and subjectivity scores derived from a predefined lexicon. Polarity ranges from -1 (very negative) to +1 (very positive), and subjectivity ranges from 0 (objective) to 1 (subjective). While simpler than deep learning approaches, it is effective where speed, transparency, and ease of use are priorities. It also supports multilingual sentiment analysis when combined with translation tools.

3.1.2. VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based, lexicon-driven sentiment analyzer developed specifically for social media text, but also performs well on other informal content such as reviews and emails

[16]. It uses a polarity lexicon with word scores from –1 to +1 and aggregates these scores to classify the overall sentiment of a sentence.

VADER is particularly effective for English text, lightweight, and requires no pre-training, making it suitable for real-time or low-resource applications [17].

3.1.3. AFINN

AFINN is a lexicon-based sentiment analysis tool, Nielsen [18], consisting of a predefined list of words scored from –5 (most negative) to +5 (most positive). It provides a straightforward, numerically weighted approach for sentiment detection and is especially suitable for short, informal English texts such as tweets or user reviews.

3.1.4. Stanza

Stanza is an advanced NLP library developed by the Stanford NLP Group, Qi et al. [19], supporting over 70 languages. For sentiment classification, it uses a convolutional neural network (CNN) to assign sentiment labels on a scale of 0 (negative), 1 (neutral), or 2 (positive). Its multilingual capabilities make it a powerful tool for cross-lingual sentiment analysis.

3.1.5. Nlptown

Nlptown is a sentiment classifier based on the BERT-base-multilingual-uncased model, fine-tuned on product reviews in six languages, including English, German, and French [20, 21]. Leveraging BERT's transformer-based architecture, it captures contextual nuances in text and performs well on multi-language review classification tasks, especially in domains like e-commerce and customer feedback.

3.1.6. Sentistrength

Sentistrength is a lexicon-based tool developed to capture sentiment in informal, user-generated content such as social media posts [22]. It handles emoticons, misspellings, and punctuation to better assess emotional tone in noisy data. Its rule-based structure enables effective detection of both positive and negative sentiment strengths within a given text.

3.1.7 Flair

Flair is an NLP framework known for its easy integration and high-performance models across various tasks [23]. For sentiment analysis, Flair utilizes contextual word embeddings, often from transformer models like BERT, allowing for accurate and nuanced sentiment classification. It supports multiple languages and is well-suited for handling complex text inputs from domains such as social media, news, and product reviews.

3.2. Stage 2: Ensemble Sentiment Analysis Method

In Stage 2, an ensemble sentiment classification approach is applied using majority voting. Each of the seven sentiment analyzers outputs a sentiment label (e.g., positive, negative, or neutral), and the most frequent label among them is selected as the final prediction. This approach assumes that aggregating predictions from multiple models can yield more reliable and accurate results.

In addition to using all seven analyzers, a second ensemble strategy explores subsets of four analyzers at a time. This alternative setting allows the study to assess whether smaller ensembles, carefully selected, can outperform or match the accuracy of full ensembles by leveraging the strengths of certain models while minimizing potential noise from weaker ones. In total, 37 ensemble combinations are tested:

- 1 combination using all seven sentiment analyzers
- 36 combinations formed by selecting all possible subsets of four analyzers

The two-stage approach is applied to three types of input representations for each review:

- HEAD (title of the review)
- TEXT (body of the review)
- HEAD+TEXT (combined title and body)

Figure 1 illustrates the overall workflow of the two-stage sentiment analysis schema. In this schema, *N* represents the number of sentiment analyzers used in the process, where it is 7 for this study.

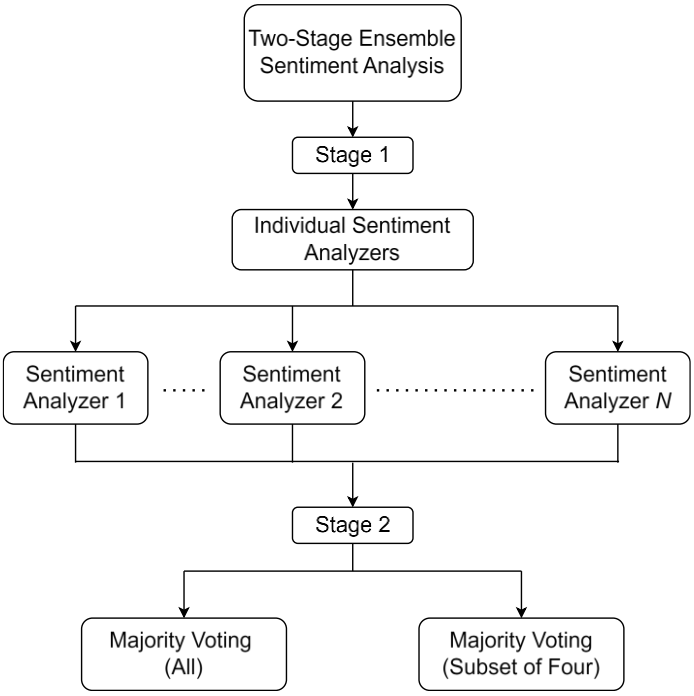

Figure 1.
Overview of the two-stage sentiment analysis framework.

4. Results

This section describes the data collection process, outlines the preprocessing steps applied to the original customer reviews, and presents the experimental results of the proposed two-stage sentiment analysis algorithm across different sections of the restaurant reviews.

4.1. Data Collection and Preprocessing

The text data used in this study was sourced from the TripAdvisor platform, which hosts a substantial number of restaurant and accommodation reviews. These user-generated reviews offer valuable insights based on personal dining experiences. A custom-developed web crawler was employed to collect and process the data. The dataset focuses specifically on restaurant reviews from Almaty, Kazakhstan, where English and Russian are the most commonly used languages.

In total, the dataset includes 600 reviews in English and 600 reviews in Russian. To standardize the input for analysis, the Russian reviews were translated into English. Only restaurants with a minimum of five reviews were considered, resulting in a final dataset of 1,200 reviews.

Rather than using manually labeled data, the study relied on the review scores provided by customers as sentiment labels. To enhance label reliability, only reviews expressing strong opinions were included—specifically, ratings of 1 (strongly negative), 3 (neutral), and 5 (strongly positive). Reviews with weakly positive or weakly negative sentiment were excluded.

Tables 1 and 2 summarize the class distribution for English and translated Russian reviews, respectively, each consisting of an equal number of samples across the three sentiment categories. For experimental analysis, three input types were evaluated separately: the review HEAD, the review TEXT, and their combination (HEAD+TEXT).

Table 1.
Distribution of Restaurant reviews in English.

No	Class Label	Samples
1	Negative	200
2	Neutral	200
3	Positive	200

Table 2.
Distribution of Restaurant reviews translated from Russian.

No	Class Label	Samples
1	Negative	200
2	Neutral	200
3	Positive	200

During preprocessing, all review texts were lemmatized using spaCy's English lemmatizer [24]. Lemmatization transforms words into their base or dictionary forms, enabling the grouping of words with similar meanings for example, both "orders" and "order" are reduced to "order," ensuring consistency in analysis. Lemmatization was chosen over stemming due to its ability to retain meaningful word forms.

4.2. Success Measures

In this study, both Micro-F1 and Macro-F1 metrics are used to evaluate classification performance. The Micro-F1 score is computed by considering all classification decisions across the entire dataset, without taking class distribution into account. As a result, it tends to favor larger classes in imbalanced datasets, potentially overshadowing the performance on smaller classes [1]. The Micro-F1 score is defined as follows:

$$Micro-F1=\frac{2\times p\times r}{p+r}, \quad (1)$$

In this formula, *p* and *r* represent the overall precision and recall values across all classification decisions, respectively. In contrast, the Macro-F1 score is calculated by first computing the F1 score for each class individually, then averaging these scores across all classes. This approach assigns equal weight to each class, regardless of its frequency in the dataset, making it more suitable for evaluating performance on imbalanced datasets. The Macro-F1 score is defined as follows:

$$Macro-F1=\frac{\sum_{k=1}^{C}F_k}{C}, \qquad F_k=\frac{2\times p_k\times r_k}{p_k+r_k}, \quad (2)$$

In this formula, p_k and r_k represent the precision and recall values of class *k*, respectively.

4.3. The Results of Stage 1 Experiments

4.3.1. Analysis of the Customer Reviews Originally in English

Table 3 displays the performance of seven sentiment analyzers across both Micro-F1 and Macro-F1 scores. The best performance in both metrics was achieved by Stanza. Notably, the highest Micro-F1 score was obtained using the TEXT section of the reviews, while the highest Macro-F1 score was obtained using the HEAD section. This suggests that despite the HEAD section containing shorter text, it proves valuable for distinguishing between the three sentiment classes (positive, negative, and neutral). Sentistrength and AFINN also performed well, following closely behind Stanza.

Table 3.
Success measures (%) for Restaurant reviews in English (Single Sentiment Analyzer).

	Micro-F1			Macro-F1		
Methods	**HEAD**	**TEXT**	**HEAD+TEXT**	**HEAD**	**TEXT**	**HEAD+TEXT**
TextBlob	0.576	0.520	0.561	0.580	0.484	0.533
VADER	0.628	0.568	0.581	0.631	0.485	0.490
AFINN	0.640	0.551	0.560	0.643	0.485	0.483
Stanza	**0.678**	***0.711***	**0.706**	***0.678***	**0.649**	**0.636**
Nlptown	0.410	0.503	0.513	0.299	0.402	0.411
Sentistrength	0.646	0.606	0.621	0.655	0.571	0.584
Flair	0.608	0.650	0.653	0.488	0.525	0.527

4.3.2. Analysis of the Customer Reviews Originally in Russian

Table 4 presents the performance of seven sentiment analyzers on Russian restaurant reviews. The best performance in both Micro-F1 and Macro-F1 scores was achieved by Stanza. The highest results for both metrics were obtained using the TEXT section of the reviews. In contrast to the English reviews, the HEAD section of the Russian reviews did not perform as well. Sentistrength, Flair, and VADER followed closely behind Stanza in performance.

Table 4.
Success measures (%) for Restaurant reviews in Russian (Single Sentiment Analyzer).

	Micro-F1			Macro-F1		
Methods	**HEAD**	**TEXT**	**HEAD+TEXT**	**HEAD**	**TEXT**	**HEAD+TEXT**
TextBlob	0.488	0.548	0.583	0.494	0.520	0.560
VADER	0.565	0.576	0.581	0.573	0.488	0.488
AFINN	0.555	0.561	0.581	0.561	0.481	0.498
Stanza	**0.585**	***0.708***	**0.685**	**0.591**	***0.626***	**0.591**
Nlptown	0.388	0.471	0.471	0.275	0.375	0.374
Sentistrength	0.581	0.606	0.630	0.590	0.566	0.588
Flair	0.573	0.653	0.655	0.457	0.527	0.530

4.4. The Results of Stage 2 Experiments

4.4.1. Analysis of the Customer Reviews Originally in English

Tables 5-7 present the performance of the top-10 two-stage ensemble sentiment analysis approaches for reviews originally in English. As shown in Table 5, the best performance in terms of both Micro-F1 and Macro-F1 scores for the HEAD section of the reviews was achieved by the combination of AFINN+Stanza+Sentistrength+Flair. In Table 6, the best

performance for the TEXT section was obtained with the combination of TextBlob+Stanza+Nlptown+Sentistrength, again in terms of both Micro-F1 and Macro-F1 scores. Similarly, Table 8 shows that the combination of TextBlob+Stanza+Nlptown+Sentistrength achieved the highest performance for the HEAD+TEXT combination of reviews in both metrics.

Stanza and Sentistrength appear consistently across all settings, with Stanza's strong individual performance explaining its frequent inclusion in the top combinations. While Majority Voting (7 methods) ranks third in Table 5, it does not appear among the top 10 in Tables 6 and 7, suggesting that ensembling a selected set of methods can outperform using all available sentiment analyzers together.

Table 5.
Top 10 Success measures (%) for Restaurant reviews in English (Majority Voting-HEAD).

Methods	Micro-F1	Macro-F1
AFINN+Stanza+Sentistrength+Flair	0.683	0.682
VADER+Stanza+Sentistrength+Flair	0.680	0.678
Majority Voting (7 methods)	0.678	0.679
VADER+AFINN+Stanza+Flair	0.675	0.672
AFINN+Stanza+Nlptown+Sentistrength	0.671	0.673
TextBlob+AFINN+Sanza+Fair	0.670	0.667
VADER+AFINN+Stanza+Nlptown	0.670	0.669
TextBlob+Stanza+Sentistrength+Flair	0.668	0.667
VADER+Stanza+Nlptown+Sentistrength	0.668	0.670
TextBlob+VADER+Stanza+Sentistrength	0.666	0.671

Table 6.
Top 10 Success measures (%) for Restaurant reviews in English (Majority Voting-TEXT).

Methods	Micro-F1	Macro-F1
TextBlob+Stanza+Nlptown+Sentistrength	0.733	0.684
TextBlob+AFINN+Stanza+Nlptown	0.723	0.664
TextBlob+VADER+Stanza+Nlptown	0.720	0.657
VADER+Stanza+Nlptown+Sentistrength	0.716	0.650
AFINN+Stanza+Nlptown+Sentistrength	0.713	0.646
VADER+AFINN+Stanza+Nlptown	0.708	0.632
Stanza+Nlptown+Sentistrength+Flair	0.703	0.631
TextBlob+Stanza+Nlptown+Flair	0.701	0.627
AFINN+Stanza+Nlptown+Flair	0.701	0.621
VADER+Stanza+Nlptown+Flair	0.698	0.617

Table 7.
Top 10 Success measures (%) for Restaurant reviews in English (Majority Voting -HEAD + TEXT).

Methods	Micro-F1	Macro-F1
TextBlob+Stanza+Nlptown+Sentistrength	0.730	0.676
TextBlob+VADER+Stanza+Nlptown	0.721	0.660
TextBlob+AFINN+Stanza+Nlptown	0.718	0.656
VADER+Stanza+Nlptown+Sentistrength	0.708	0.636
AFINN+Stanza+Nlptown+Sentistrength	0.700	0.623
TextBlob+Stanza+Nlptown+Flair	0.696	0.619
TextBlob+Nlptown+Sentistrength+Flair	0.695	0.611
TextBlob+AFINN+Nlptown+Flair	0.693	0.606
VADER+AFINN+Stanza+Nlptown	0.693	0.604
TextBlob+VADER+Nlptown+Flair	0.690	0.599

4.4.2. Analysis of the Customer Reviews Originally in Russian

Tables 8-10 display the performance of the top-10 two-stage ensemble sentiment analysis approaches for reviews translated from Russian. As seen in Table 8, the combination of AFINN+Stanza+Sentistrength+Flair achieved the best performance for the HEAD section, in terms of both Micro-F1 and Macro-F1 scores. In Table 9, the combination of TextBlob+Stanza+Nlptown+Sentistrength yielded the highest performance for the TEXT section, again in both metrics. Likewise, Table 10 shows that TextBlob+Stanza+Nlptown+Sentistrength performed best for the HEAD+TEXT combination.

Stanza and Sentistrength consistently appear in top-performing combinations, with Stanza's strong individual performance accounting for its frequent inclusion. Notably, while Majority Voting (7 methods) ranks fifth in Table 8, it does not make the top 10 in Tables 9 and 10, suggesting that a carefully selected ensemble of methods can outperform the use of all available sentiment analyzers combined.

Table 8.
Top 10 Success measures (%) for Restaurant reviews in Russian (Majority Voting-HEAD).

Methods	Micro-F1	Macro-F1
AFINN+Stanza+Sentistrength+Flair	0.593	0.597
VADER+Stanza+Sentistrength+Flair	0.588	0.590
TextBlob+AFINN+Stanza+Flair	0.583	0.587
TextBlob+VADER+Stanza+Flair	0.581	0.584
Majority Voting (7 methods)	0.578	0.586
TextBlob+VADER+Stanza+Sentistrength	0.576	0.585
TextBlob+VADER+Sentistrength+Flair	0.576	0.583
VADER+AFINN+Stanza+Flair	0.576	0.580
TextBlob+VADER+AFINN+Stanza	0.575	0.584
VADER+AFINN+Sentistrength+Flair	0.575	0.583

Table 9.
Top 10 Success measures (%) for Restaurant reviews in Russian (Majority Voting-TEXT).

Methods	Micro-F1	Macro-F1
TextBlob+Stanza+Nlptown+Sentistrength	0.701	0.642
TextBlob+AFINN+Stanza+Nlptown	0.693	0.619
TextBlob+VADER+Stanza+Nlptown	0.690	0.614
AFINN+Stanza+Nlptown+Flair	0.690	0.592
VADER+Stanza+Nlptown+Flair	0.688	0.591
AFINN+Stanza+Nlptown+Sentistrength	0.688	0.607
VADER+AFINN+Stanza+Nlptown	0.686	0.590
VADER+Stanza+Nlptown+Sentistrength	0.686	0.606
TextBlob+Stanza+Nlptown+Flair	0.685	0.602
Stanza+Nlptown+Sentistrength+Flair	0.683	0.596

Table 10.
Top 10 Success measures (%) for Restaurant reviews in Russian (Majority Voting -HEAD + TEXT).

Methods	Micro-F1	Macro-F1
TextBlob+Stanza+Nlptown+Sentistrength	0.703	0.641
TextBlob+AFINN+Stanza+Nlptown	0.701	0.631
TextBlob+VADER+Stanza+Nlptown	0.691	0.615
TextBlob+Stanza+Nlptown+Flair	0.686	0.601
Stanza+Nlptown+Sentistrength+Flair	0.686	0.597
TextBlob+Nlptown+Sentistrength+Flair	0.685	0.611
AFINN+Stanza+Nlptown+Sentistrength	0.685	0.601
VADER+Stanza+Nlptown+Sentistrength	0.683	0.599
VADER+Stanza+Nlptown+Flair	0.681	0.577
TextBlob+AFINN+Nlptown+Flair	0.678	0.588

4.3. The Results of Stage 2 Experiments

This section presents a comparison of the results from Stage 1 and Stage 2 experiments to evaluate the effectiveness of two-stage ensemble methods versus single sentiment analyzers in sentiment analysis of restaurant reviews.

For the reviews originally written in English, in Stage 1, Stanza consistently achieved the best performance across both Micro-F1 and Macro-F1 scores. Notably, Stanza achieved the highest Micro-F1 score in the TEXT section and the highest Mac-ro-F1 score in the HEAD section. This suggests that Stanza is highly effective in distinguishing sentiment classes while also handling different review sections effectively. However, in Stage 2, the performance of ensemble methods surpassed that of individual analyzers. Specifically, in Table 5, the combination of AFINN+Stanza+Sentistrength+Flair achieved the highest performance in the HEAD section. Similarly, in Table 6 and Table 7, ensemble methods like TextBlob+Stanza+Nlptown+Sentistrength dominated in the TEXT and HEAD+TEXT sections, respectively. This indicates that the two-stage ensemble approaches provide a more nuanced understanding of sentiment, benefiting from the diversity of individual analyzers in each combination.

For the reviews translated from Russian, Stanza also led the Stage 1 experiments, particularly in the TEXT section (Table 4), where it achieved the highest performance in both Micro-F1 and Macro-F1 scores. However, in Stage 2, the best performance for the HEAD section was again obtained by the combination of AFINN+Stanza+Sentistrength+Flair (Table 8), while the combination of TextBlob+Stanza+Nlptown+Sentistrength excelled in both the TEXT and HEAD+TEXT sections (Tables 9 and 10). These results align with the English reviews, showing that two-stage ensemble methods outperform individual analyzers, even when Stanza performed well as a standalone method.

An interesting trend observed across reviews in both languages is the performance of Majority Voting (7 methods), which did not perform the best in any setting and rarely appeared in the top-10 results. This suggests that a more strategic

ensemble of methods outperforms the inclusion of all available analyzers, supporting the notion that carefully selecting a set of effective methods can provide superior results compared to using all available options in a majority vote.

Considering the maximum performances for reviews originally in English, the highest Micro-F1 score for English reviews was 0.733, achieved by the ensemble method TextBlob+Stanza+Nlptown+Sentistrength for the TEXT section in Table 6. This indicates that this combination provides the most accurate sentiment classification when analyzing the textual content of the reviews. The highest Macro-F1 score for English reviews was 0.684, achieved by the same ensemble method TextBlob+Stanza+Nlptown+Sentistrength, specifically in the TEXT section in Table 6. This suggests that this combination not only provides the best overall accuracy but also demonstrates a well-balanced classification of all sentiment classes (positive, negative, and neutral).

Considering the maximum performances for reviews translated from Russian, the highest Micro-F1 score for Russian reviews was 0.703, achieved by the combination TextBlob+Stanza+Nlptown+Sentistrength for the HEAD+TEXT combination in Table 10. This indicates that, when both the HEAD and TEXT sections are considered together, this ensemble method outperforms others in accurately identifying the sentiment of Russian reviews. The highest Macro-F1 score for Russian reviews was 0.642, achieved by the same ensemble method TextBlob+Stanza+Nlptown+Sentistrength, again for the TEXT section in Table 9. This result shows that, while the method excels in overall accuracy, it also demonstrates a strong ability to classify the different sentiment classes effectively.

5. Conclusions

Research on ensemble sentiment analysis approaches for restaurant reviews is limited in the literature. This study contributes to filling this gap and provides valuable insights into bilingual restaurant reviews. The performances of both individual sentiment analyzers and ensemble decision approaches vary between reviews originally written in English and those translated from another language, such as Russian. To clearly observe these differences, only reviews indicating strong sentiment were considered in this study. Research on restaurant customer reviews in Central Asia, particularly in Kazakhstan, remains limited. This study represents one of the first efforts to examine customer priorities and concerns in Almaty's restaurant industry, offering valuable insights for local stakeholders.

Overall, the Stage 2 ensemble results demonstrate the clear advantages of using carefully selected ensembles over single sentiment analyzers. While individual methods like Stanza performed well in Stage 1, the combination of multiple analyzers in Stage 2 led to improved performance, particularly in the TEXT and HEAD+TEXT sections. The consistent appearance of Stanza and Sentistrength in the top combinations further highlights their effectiveness in sentiment classification across both languages.

5.1. Theoretical Implications

The rise of digital platforms has enabled customers to rapidly share opinions and complaints, shaping perceptions of service quality and customer satisfaction. This study contributes to the growing body of research on online restaurant reviews by applying sentiment analysis to assess the performances of various individual and ensemble sentiment analysis methods for reviews originating from two different languages.

5.2. Practical Implications

This study offers valuable insights for restaurant managers, operators, and marketing teams in Almaty, as well as for the academic community working on sentiment analysis. Understanding sentiment trends and customer preferences can inform targeted marketing strategies and improve the customer experience, both of which are critical in a competitive market like the restaurant industry.

5.3. Limitations and Suggestions for Future Research

Different preprocessing settings and sentiment analysis methods can be integrated into similar studies, as well as reviews from other languages, if studies from different regions worldwide are available. In this study, reviews originally written in English and translations of reviews in Russian were used, as these are the dominant languages in Kazakhstan's tourism industry.

Future research could expand the dataset by incorporating restaurant reviews from other travel platforms, allowing for broader comparative analyses. Exploring restaurant reviews from various regions within Kazakhstan may reveal regional variations in customer preferences and sentiment distributions. Moreover, leveraging advanced text mining, natural language processing (NLP), and machine learning techniques could further enhance insights into customer expectations and industry trends.

References

[1] A. K. Uysal, "Comparative Analysis of Recent Feature Selection Methods for Sentiment Classification," *Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering,* vol. 19, no. 3, pp. 645-659, 2018.

[2] A. K. Uysal *et al.*, "Analysis of Almaty's Restaurant Reviews through Topic Modelling," *Academica Turistica-Tourism and Innovation Journal,* vol. 17, no. 2, 2024.

[3] J.-y. Kim and J. Hwang, "Who is an evangelist? Food tourists' positive and negative eWOM behavior," *International Journal of Contemporary Hospitality Management,* vol. 34, no. 2, pp. 555-577, 2022.

[4] S. Gao, O. Tang, H. Wang, and P. Yin, "Identifying competitors through comparative relation mining of online reviews in the restaurant industry," *International Journal of Hospitality Management,* vol. 71, pp. 19-32, 2018.

[5] S. W. Litvin, J. E. Blose, and S. T. Laird, "Tourists' use of restaurant webpages: Is the internet a critical marketing tool?," *Journal of Vacation Marketing,* vol. 11, no. 2, pp. 155-161, 2005.

[6] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications,* vol. 39, no. 5, pp. 6000-6010, 2012.

[7] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," presented at the In 2019 12th International Conference on Information & Communication Technology and System (ICTS) (pp. 49-54). IEEE, 2019.

[8] O. Sharif, M. M. Hoque, and E. Hossain, "Sentiment analysis of Bengali texts on online restaurant reviews using multinomial Naïve Bayes," presented at the In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-6). IEEE, 2019.

[9] M. Adnan, R. Sarno, and K. R. Sungkono, "Sentiment analysis of restaurant review with classification approach in the decision tree-j48 algorithm," presented at the In 2019 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 121-126). IEEE, 2019.

[10] H. Burra and P. Mishra, "Restaurant reviews sentimental analysis using machine learning approach," presented at the International Conference on Emerging Techniques in Computational Intelligence (ICETCI) (pp. 414-417). IEEE, 2024.

[11] D. A. Al-Qudah, A.-Z. Ala'M, A. I. Cristea, J. J. Merelo-Guervós, P. A. Castillo, and H. Faris, "Prediction of sentiment polarity in restaurant reviews using an ordinal regression approach based on evolutionary XGBoost," *PeerJ Computer Science,* vol. 11, p. e2370, 2025.

[12] N. Saleena, "An ensemble classification system for twitter sentiment analysis," *Procedia Computer Science,* vol. 132, pp. 937-946, 2018.

[13] J. Kazmaier and J. H. Van Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Systems with Applications,* vol. 187, p. 115819, 2022.

[14] D. Sarkar, *Text analytics with python*. New York, NY, USA: Apress, 2016.

[15] P. Deitel and H. Deitel, "Intro to python for computer science and data science." London, UK: Pearson education, 2020, p. 17.

[16] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," presented at the In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225), 2014.

[17] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," *Expert Systems with Applications,* vol. 162, p. 113746, 2020.

[18] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903,* 2011.

[19] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082,* 2020.

[20] Nlptown, "Bert-base-multilingual-uncased-sentiment hugging face," Retrieved: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment2022.

[21] J. A. Lossio-Ventura *et al.*, "A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data," *JMIR Mental Health,* vol. 11, p. e50150, 2024.

[22] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology,* vol. 63, no. 1, pp. 163-173, 2012.

[23] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *In Proceedings of NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 54–59),* 2019.

[24] M. Kundu *et al.*, "Modulation of the tumor microenvironment and mechanism of immunotherapy-based drug resistance in breast cancer," *Molecular Cancer,* vol. 23, no. 1, p. 92, 2024.