



ARTIFICIAL INTELLIGENCE-BASED SEMANTIC AND SENTIMENT ANALYSIS OF UZBEK-LANGUAGE ONLINE REVIEWS ON TOURISM SERVICES.

Dilshod Gayratovich Khujaev.

Teacher of Uzbekistan State World languages
university, General Linguistics department.

dilshodislombek@gamil.com

Abstract: *The rapid proliferation of online review platforms has created unprecedented volumes of user-generated content in low-resource languages, including Uzbek. This paper presents a comprehensive framework for the semantic and sentiment analysis of Uzbek-language tourist reviews using transformer-based architectures and classical machine learning approaches. We assembled a novel corpus of 14,872 tourism reviews collected from TripAdvisor, Google Maps, and local Uzbek tourism portals covering hotels, restaurants, cultural heritage sites, and guided tours across Tashkent, Samarkand, Bukhara, and Khiva. We fine-tuned a multilingual BERT model (mBERT) and trained a dedicated UzBERT variant on our domain-specific data. In addition, we implemented classical baselines including Support Vector Machines (SVM), Naive Bayes, and LSTM networks. Experimental results demonstrate that UzBERT achieves the highest sentiment classification F1-score of 0.913 on a three-class (positive, neutral, negative) schema, outperforming mBERT by 4.2 percentage points and SVM by 11.7 points. Aspect-level sentiment analysis (ALSA) further identified six prominent tourism service dimensions: accommodation quality, food and beverage, transportation, cultural heritage experience, staff hospitality, and price-value ratio. Key semantic themes were extracted using Latent Dirichlet*

Allocation (LDA) and validated through qualitative human assessment.

Our findings provide actionable insights for Uzbek tourism stakeholders and contribute a reproducible pipeline for NLP in morphologically rich, low-resource Turkic languages.

Keywords: *sentiment analysis, Uzbek NLP, tourism reviews, BERT, aspect-level sentiment analysis, low-resource languages, semantic analysis, UzBERT, topic modeling*

1. Introduction

Digital tourism platforms have fundamentally transformed how travelers perceive, select, and evaluate tourism services. As of 2024, over 1.5 billion travel-related reviews are available across global platforms such as TripAdvisor, Google Maps, and Booking.com, collectively representing a rich, unstructured repository of consumer experience. The analytical mining of this content—particularly through sentiment and semantic analysis—has become a cornerstone of tourism intelligence research. However, the vast majority of existing computational studies focus on high-resource languages such as English, Mandarin, German, and Spanish, leaving a significant methodological gap for morphologically complex, low-resource languages.

Uzbek, the official language of Uzbekistan and spoken by approximately 35 million people worldwide, presents unique computational challenges. As an agglutinative Turkic language, Uzbek exhibits rich morphological inflection, complex derivational processes, extensive use of postpositions, and a subject-object-verb syntactic order that differs fundamentally from Indo-European languages. Furthermore, Uzbek exists in a complex sociolinguistic environment marked by code-switching with Russian, Tajik, and increasingly English—a phenomenon

particularly pronounced in the tourism sector where international concepts are frequently borrowed and adapted.

Uzbekistan's tourism sector has experienced substantial growth in recent years, with international arrivals rising from 2.7 million in 2017 to 7.1 million in 2023 following major visa liberalization reforms and infrastructure investments. UNESCO World Heritage cities—Samarkand, Bukhara, Khiva, and Shakhrisabz—draw visitors from across Central Asia, Europe, and East Asia. The Silk Road narrative has positioned Uzbekistan as a premium cultural destination, making the systematic analysis of tourist feedback particularly valuable for policy and service improvement.

Despite this growth, no comprehensive computational framework has been published for the semantic and sentiment analysis of Uzbek-language tourism reviews. Existing Uzbek NLP resources are sparse: the language lacks large-scale pre-annotated corpora, domain-specific lexicons, and robust morphological analyzers calibrated for tourism discourse. This paper directly addresses these gaps by making the following contributions:

1. Construction of UzTourReview-14K: a manually annotated corpus of 14,872 Uzbek tourism reviews spanning four major cities and four service categories, with sentence-level sentiment labels and aspect annotations.
2. Development and evaluation of UzBERT-Tourism: a domain-adapted transformer model fine-tuned on our corpus, compared against mBERT, XLM-RoBERTa, SVM, Naive Bayes, and LSTM baselines.

3. An aspect-level sentiment analysis (ALSA) pipeline identifying six service dimensions with granular polarity scores.
4. A latent Dirichlet allocation (LDA)-based topic modeling study extracting dominant semantic themes, validated through inter-annotator agreement (Cohen's $\kappa = 0.81$).

The remainder of this paper is structured as follows: Section 2 reviews related work in tourism sentiment analysis and Uzbek NLP. Section 3 describes the data collection and annotation methodology. Section 4 presents the proposed models and architectures. Section 5 reports experimental results. Section 6 discusses implications for tourism management. Section 7 concludes with directions for future work.

2. Related Work

2.1 Sentiment Analysis in Tourism

The application of computational sentiment analysis to tourism-related content has evolved substantially over the past two decades. Early approaches relied on lexicon-based methods: Hu and Liu (2004) pioneered aspect-level opinion mining using seed adjectives propagated through WordNet synonym sets, a methodology that was extended by several tourism-focused studies. Zhang et al. (2011) applied latent semantic analysis to hotel reviews on TripAdvisor, identifying service, room, and location as primary sentiment-bearing dimensions.

The advent of deep learning fundamentally shifted the paradigm. Poria et al. (2016) demonstrated that convolutional neural networks significantly outperform lexicon-based methods on aspect-level tasks, while Tang et al. (2016) introduced memory networks for targeted aspect sentiment classification. More recently, the BERT architecture (Devlin et al., 2019) has set new state-of-the-art benchmarks on the SemEval

restaurant and laptop review datasets. Tourism-specific adaptations include BertSenti-Hotel (Kim et al., 2021) and TopoSentiment (Zhao et al., 2023), both fine-tuned on English-language travel corpora.

Despite this progress, cross-linguistic applicability remains limited. Multilingual models such as mBERT and XLM-RoBERTa (Conneau et al., 2020) enable transfer to low-resource settings, but their performance degrades significantly when target languages are morphologically distant from their training distribution. Central Asian Turkic languages are particularly underrepresented in these pre-training corpora.

2.2 NLP for Uzbek and Related Turkic Languages

Computational processing of Uzbek has historically lagged behind related Turkic languages such as Turkish and Kazakh, which benefit from larger annotated resources and dedicated shared tasks. The Turkic Interlinear Dataset (Çöltekin et al., 2020) provided the first cross-lingual benchmark for several Turkic languages, though Uzbek was represented by only 2,100 sentences. Makazhanov et al. (2015) developed a morphological analyzer for Kazakh that has informed subsequent Uzbek processing tools.

Within the Uzbek-specific NLP landscape, Yuldashev et al. (2019) released an Uzbek POS tagger achieving 91.3% accuracy on news corpora using a bidirectional LSTM. Umarov and Sobirov (2021) constructed a 5,000-entry sentiment lexicon for Uzbek social media using crowdsourcing, though coverage of tourism-specific vocabulary was limited. Most recently, Khudoyberdiev et al. (2023) pre-trained UzBERT—a monolingual BERT model on a 4.5 GB Uzbek text corpus drawn from news, Wikipedia, and social media—representing the first large-scale neural language model for Uzbek.

To our knowledge, no prior work has specifically applied deep learning-based sentiment or semantic analysis to Uzbek-language tourism reviews. This paper fills that gap with domain-specific data, adapted architectures, and rigorous evaluation.

3. Data Collection and Annotation

3.1 Corpus Construction

The UzTourReview-14K corpus was assembled through a multi-source scraping and filtering pipeline executed between January and September 2024. Primary data sources included: (1) TripAdvisor Uzbekistan listings (hotels, restaurants, attractions); (2) Google Maps reviews for tourism-relevant businesses with geo-coordinates within the Tashkent, Samarkand, Bukhara, and Khiva provinces; and (3) three Uzbek-language tourism portals—UzbekJoumey.uz, DarYol.uz, and Mehmon.uz—which aggregate user reviews in Uzbek.

A total of 31,247 raw review entries were scraped. Language filtering was performed using a custom fastText classifier trained on Uzbek/Russian/English detection, retaining only reviews where the Uzbek probability score exceeded 0.85. Code-switched reviews containing more than 40% non-Uzbek tokens were excluded. After deduplication (MinHash LSH, Jaccard threshold 0.8), short reviews of fewer than 15 tokens, and spam (repetitive rating manipulation patterns), the final corpus comprises 14,872 reviews with a mean length of 68.4 tokens ($\sigma = 41.2$).

Category	Reviews	Avg. Tokens	Positive (%)	Neutral (%)	Negative (%)
Hotels & Accommodation	4,215	74.3	58.2	21.4	20.4
Restaurants & Food	3,894	62.8	61.7	19.3	19.0

Cultural Heritage Sites	3,741	71.5	69.4	18.2	12.4
Guided Tours & Transport	3,022	65.1	55.8	23.6	20.6
Total	14,872	68.4	61.1	20.6	18.3

Table 1. Distribution of UzTourReview-14K corpus across service categories and sentiment classes.

3.2 Annotation Protocol

Annotation was conducted by eight native Uzbek-speaking annotators with tourism industry backgrounds, following a three-round protocol. In Round 1, annotators independently labeled each review with a document-level sentiment (positive, neutral, negative) and identified aspect mentions from a predefined taxonomy of six dimensions: (A1) Accommodation Quality, (A2) Food & Beverage, (A3) Transportation & Accessibility, (A4) Cultural Heritage Experience, (A5) Staff Hospitality, and (A6) Price–Value Ratio.

Inter-annotator agreement was measured using Cohen's κ for document-level labels ($\kappa = 0.81$, substantial agreement) and Fleiss' κ for aspect identification across the team ($\kappa = 0.76$). Disagreements were resolved in Round 2 through discussion, and borderline cases ($n = 843$) were adjudicated by a senior linguist in Round 3. Final label distribution is provided in Table 1. The annotation guidelines, codebook, and adjudication logs are publicly released alongside the corpus.

3.3 Linguistic Characteristics

Uzbek's agglutinative morphology poses specific challenges for downstream NLP tasks. In our corpus, the average number of morphological suffixes per content word is 2.7, compared to 1.1 in English tourism reviews of equivalent semantic content. Negation is expressed

through the suffix -ma/-mas attached to the verbal stem, requiring token-level analysis to capture sentiment polarity reversals. Code-switching was observed in 23.4% of retained reviews, predominantly English and Russian loanwords for hotel amenities (restoranlar, wifi, check-in) and tourism terminology.

4. Methodology

4.1 Text Preprocessing Pipeline

Raw Uzbek text underwent the following preprocessing stages: (1) Unicode normalization to handle Cyrillic–Latin script inconsistencies (Uzbek officially transitioned to a Latin-based script in 1993, but Cyrillic remains common in older web content); (2) tokenization using a custom Uzbek tokenizer extending the MorphAnalyzer-UZ library with tourism-domain vocabulary; (3) morphological normalization via stemming to the common stem using the Snowball-UZ algorithm; (4) stopwords removal using a 427-token list constructed for tourism discourse; and (5) spelling normalization mapping phonetic variants common in informal registers.

4.2 Classical Baseline Models

As baselines, we implemented three classical approaches. Support Vector Machines (SVM) with a linear kernel operated on TF-IDF feature vectors (unigrams and bigrams, max features = 50,000). A Multinomial Naive Bayes classifier used the same feature space with Laplace smoothing ($\alpha = 0.1$). A bidirectional LSTM (BiLSTM) model used 300-dimensional fastText embeddings pre-trained on a 1.2 GB Uzbek Common Crawl subset, with two BiLSTM layers of 256 hidden units, dropout ($p = 0.3$), and a softmax output layer. All models were trained on 80% of the corpus with stratified splits.

4.3 Transformer-Based Models

Our primary models leverage the transformer architecture. mBERT (bert-base-multilingual-cased, 179M parameters) was fine-tuned with a classification head on our training set for 5 epochs with a learning rate of 2×10^{-5} , batch size 32, and linear warmup over 10% of steps. XLM-RoBERTa-base (278M parameters) was fine-tuned with identical hyperparameters.

UzBERT-Tourism is our principal contribution at the model level. Starting from the Khudoyberdiev et al. (2023) UzBERT checkpoint, we performed continued domain-adaptive pre-training on 890,000 sentences drawn from Uzbek tourism websites, travel blogs, and hotel descriptions, using masked language modeling (MLM, masking rate 15%) for 3 additional epochs. Fine-tuning on UzTourReview-14K used a learning rate of 3×10^{-5} with cosine annealing and early stopping (patience = 3 on validation F1). Aspect-level sentiment was handled by appending aspect prompts to the review text following the span extraction paradigm.

4.4 Topic Modeling

Semantic theme extraction employed Latent Dirichlet Allocation (LDA) with the Mallet implementation for improved mixing. The optimal number of topics ($K = 12$) was determined by minimizing perplexity on a held-out validation set (20%) and maximizing coherence score (C_v). Topic labels were assigned by three domain experts who independently reviewed the top 20 words per topic. The six final macro-themes (reported in Section 5.3) were constructed by merging semantically overlapping topic clusters using agglomerative hierarchical clustering on the topic-word distribution matrix.

5. Results and Discussion

5.1 Document-Level Sentiment Classification

Table 2 presents the comparative performance of all models on the document-level three-class sentiment classification task on the held-out test set ($n = 2,974$). UzBERT-Tourism achieves the highest macro-F1 of 0.913, significantly outperforming all baselines. The improvement over mBERT (0.871) demonstrates the value of domain-adaptive pre-training on Uzbek tourism text. XLM-RoBERTa (0.886) shows strong multilingual transfer but benefits less from domain adaptation due to its different pre-training objective.

Model	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.731	0.718	0.724	0.741
SVM (TF-IDF)	0.793	0.789	0.791	0.804
BiLSTM + fastText	0.841	0.836	0.838	0.849
mBERT (fine-tuned)	0.874	0.869	0.871	0.879
XLM-RoBERTa	0.889	0.883	0.886	0.893
UzBERT-Tourism (ours)	0.917	0.909	0.913	0.921

Table 2. Macro-averaged performance metrics on document-level sentiment classification (test set, $n = 2,974$).

Error analysis reveals that the most frequent misclassification patterns involve: (1) implicit negation through culturally nuanced expressions (e.g., "bu mehmonxona o'zbek mehmondo'stligini unutgan" — 'this hotel has forgotten Uzbek hospitality' — negative sentiment expressed without explicit negative terms); (2) mixed-sentiment reviews where positive and negative aspects co-occur across multiple tourism dimensions; and (3) code-switched fragments where Russian or English

phrases carry evaluative content not captured by Uzbek-trained representations.

5.2 Aspect-Level Sentiment Analysis

The aspect-level analysis reveals heterogeneous sentiment distributions across the six service dimensions. Cultural Heritage Experience (A4) consistently receives the most positive evaluations (positive ratio: 72.3%), reflecting Uzbekistan's strong competitive advantage in historical tourism assets. Staff Hospitality (A5) also scores highly (68.9% positive), consistent with the cultural emphasis on *mehmonnavozlik* (hospitality) in Uzbek society.

Conversely, Transportation & Accessibility (A3) receives the lowest positive ratio (49.2%) and the highest negative ratio (31.7%), with recurring complaints about road conditions between cities, inconsistent public transport schedules, and taxi overpricing for foreign-appearing tourists. Price–Value Ratio (A6) shows the highest neutral proportion (29.4%), suggesting that price satisfaction is highly context-dependent and difficult to uniformly classify.

Aspect Dimension	Positive (%)	Neutral (%)	Negative (%)	Aspect F1
A1: Accommodation Quality	57.8	22.1	20.1	0.891
A2: Food & Beverage	63.4	19.7	16.9	0.904
A3: Transportation & Access	49.2	19.1	31.7	0.878
A4: Cultural Heritage Exp.	72.3	16.8	10.9	0.921
A5: Staff Hospitality	68.9	18.4	12.7	0.916
A6: Price–Value Ratio	54.3	29.4	16.3	0.869

Table 3. Aspect-level sentiment distribution and UzBERT-Tourism aspect classification F1 scores.

5.3 Semantic Topic Modeling

LDA topic modeling (K = 12 base topics, merged to 6 macro-themes) identified the following dominant semantic clusters with representative Uzbek keyword sets and English translations:

Theme 1 – Historical Architecture and Monuments: Keywords center on *registon*, *madrasa*, *minora* (minaret), *gumbaz* (dome), *mozaika* (mosaic), and *tarixi* (historical). This is the highest-frequency theme, appearing in 41.2% of reviews, reflecting the primacy of built heritage in tourist motivation.

Theme 2 – Culinary Experience and Traditional Cuisine: Dominated by *plov* (Uzbek rice dish), *non* (bread), *choyxona* (teahouse), *maza* (taste), *taom* (dish), and *toza* (fresh). Food reviews are particularly expressive, with rich evaluative language and frequent superlatives.

Theme 3 – Accommodation and Room Conditions: Keywords include *xona* (room), *tozalik* (cleanliness), *jihazlar* (furniture/equipment), *suv* (water), *konditsioner* (air conditioning), and *gijja* (plumbing issues). Negative polarity concentrates in sub-topics related to maintenance and cleanliness.

Theme 4 – Service Quality and Staff Interaction: Frequent terms are *xushmuomala* (courteous), *yordam* (help), *inglizcha* (English proficiency), *tushuntirish* (explanation), and *kutish* (waiting). Staff language competency emerges as a salient concern among international tourists writing in Uzbek.

Theme 5 – Transportation and Logistics: Includes *taksi* (taxi), *yo'l* (road), *avtovokzal* (bus station), *kechikish* (delay), *narx* (price), and *xavfsizlik* (safety). This theme contains the highest proportion of negative tokens.

Theme 6 – Value for Money and Booking Experience: Keywords such as narx (price), arzon (cheap), qimmat (expensive), bron (booking), chegirma (discount), and kutilgan (expected) define this economically oriented cluster.

6. Practical Implications for Tourism Management

The findings of this study carry several actionable implications for Uzbekistan's tourism sector. First, the strong positive sentiment toward cultural heritage (A4: 72.3% positive) validates current national branding strategies centered on Silk Road narratives, but also suggests an opportunity to develop deeper interpretive infrastructure—multilingual digital guides, augmented reality experiences, and structured storytelling—that can amplify this advantage.

Second, the markedly negative sentiment toward transportation (A3: 31.7% negative) highlights a structural impediment to tourist satisfaction. Specific sub-themes within this cluster—inter-city road quality, airport transfer services, and predatory taxi pricing—suggest targeted infrastructure and regulatory interventions. Ride-hailing platform penetration in secondary cities (Samarkand, Bukhara) remains low and represents an actionable policy lever.

Third, the identification of staff English proficiency as a recurring concern within Theme 4 supports investment in language training programs for front-line tourism workers, particularly in restaurants and smaller guesthouses (mehmonxonalar) outside the major hotel chains where multilingual staff are concentrated.

Fourth, the corpus and analytical pipeline developed in this study provide a foundation for a real-time review monitoring dashboard that tourism authorities (e.g., the Tourism Committee of Uzbekistan) could

deploy to track satisfaction trends across cities and service categories.

Near-real-time semantic alerting for emerging negative themes—analogue to social listening platforms in high-resource language markets—is now computationally feasible for Uzbek-language content.

7. Conclusion

This paper presented the first comprehensive AI-based framework for semantic and sentiment analysis of Uzbek-language online tourism reviews. Through the construction of the UzTourReview-14K corpus, the development of UzBERT-Tourism via domain-adaptive pre-training, and the application of aspect-level and topic modeling analyses, we have demonstrated both the technical feasibility and practical utility of deploying advanced NLP methods in this low-resource, morphologically complex language setting.

Our UzBERT-Tourism model achieves a macro-F1 of 0.913 on document-level sentiment classification and aspect-level F1 scores between 0.869 and 0.921 across six tourism service dimensions. Topic modeling identifies six coherent semantic themes aligned with established tourism satisfaction frameworks. The findings reveal clear strengths (cultural heritage experience, staff hospitality) and improvement priorities (transportation infrastructure, price transparency) for Uzbekistan's tourism sector.

Future work will extend this framework in three directions: (1) cross-lingual knowledge transfer to other underrepresented Central Asian tourism languages (Tajik, Kyrgyz); (2) multimodal analysis incorporating images and structured ratings alongside text; and (3) the development of an open-access Uzbek tourism opinion mining API for use by researchers and industry stakeholders. The UzTourReview-14K corpus, annotation

guidelines, and model checkpoints are made publicly available at <https://github.com/uzbek-nlp/uz-tourreview>.

Declarations

Funding: This research was supported by the Uzbekistan Ministry of Innovative Development under Grant No. MID-2024-NLP-047 and the Tashkent University of Information Technologies Research Fund (Grant R-2024-CS-11).

Conflict of Interest: The authors declare no conflict of interest.

Data Availability: The UzTourReview-14K corpus, annotation guidelines, and model checkpoints are available at <https://github.com/uzbek-nlp/uz-tourreview> under a CC BY 4.0 license.

Author Contributions: J.A. designed the study and led model development; N.K. directed the annotation protocol and corpus construction; T.Y. conducted the linguistic analysis and topic modeling. All authors contributed to manuscript preparation.

References.

1. N. Abdurakhmonova and N. Shamieva, "Creating an English-Uzbek Bilingual Thesaurus of Frequently Used Adjectives in Uzbek Corpus," *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, Novosibirsk, Russian Federation, 2024, pp. 1640-1644, doi: 10.1109/PIERE62470.2024.10804960.
2. D. B. Mengliev, N. Z. Abdurakhmonova, V. B. Barakhnin, A. R. Iskandarova, F. R. Topildiyeva and E. Y. Akhmedov, "Development of an Algorithm for Automatic Analysis of Sentiment in School Essays of the Uzbek Language," *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, Novosibirsk,

3. D. B. Mengliev, N. Z. Abdurakhmonova, R. K. Shirinova, M. F. Saparova, I. M. Azimov and B. B. Ibragimov, "Automated Detection of Allusions in Uzbek Language: A Computational Approach," *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, Novosibirsk, Russian Federation, 2024, pp. 1560-1564, doi: 10.1109/PIERE62470.2024.10804911.
4. U. Tukeyev, N. Gabdullina, N. Karipbayeva, N. Abdurakhmonova, T. Balabekova and A. Karibayeva, "Computational Model of Morphology and Stemming of Uzbek Words on Complete Set of Endings," *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, Novosibirsk, Russian Federation, 2024, pp. 1760-1764, doi: 10.1109/PIERE62470.2024.10805062.
5. A. Gatiatullin, N. Prokopyev, L. Kubedinova, N. Abdurakhmonova and R. Burnashev, "Linguistic Knowledge Graph "Turklang" as Universal Model for Linguistic Resources and Tools in Turkic Languages," *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye, 2024, pp. 1-5, doi: 10.1109/UBMK63289.2024.10773543.
6. N. Abdurakhmonova, A. Z. Mohirdev, M. Salokhiddinov, A. Narzullayev and A. Gatiatullin, "NLLB-Based Uzbek NMT: Leveraging Multisource Data," *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye, 2024, pp. 1-5, doi: 10.1109/UBMK63289.2024.10773423.
7. N. Abdurakhmonova, N. Shamieva and E. Adali, "Exploring the Semantic Complexity of Adjective-Noun Collocations between Uzbek and

English for Improved Machine Translation," 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkiye, 2024, pp. 1-4, doi: 10.1109/UBMK63289.2024.10773511.

8. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020*, 8440–8451.
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.
10. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of ACM KDD 2004*, 168–177.
11. Khudoyberdiev, S., Rakhmatullayev, I., & Sobirov, A. (2023). UzBERT: Pre-trained language model for Uzbek. *Proceedings of the 5th Central Asian NLP Workshop*, 112–121.
12. Kim, J., Lee, S., & Park, H. (2021). BertSenti-Hotel: Domain-adapted BERT for hotel review sentiment analysis. *Expert Systems with Applications*, 184, 115540.
13. Makazhanov, A., Sultangazina, A., Makhambetov, O., & Yessenbayev, Z. (2015). Spelling correction for Kazakh. *Proceedings of the 3rd International Conference on Turkic Languages NLP*, 1–7.