LINGUISTIC CHALLENGES IN PERFORMING ASPECT-BASED SENTIMENT ANALYSIS IN THE UZBEK LANGUAGE

Malika Suyunova
PhD Student at Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, Uzbekistan

ABSTRACT

This paper analyzes the linguistic, technical, and methodological challenges encountered in performing Aspect-Based Sentiment Analysis (ABSA) for the Uzbek language. The study demonstrates that the agglutinative nature of the Uzbek language, its morphological richness, flexible word order, and context-dependent emotional expressions significantly affect the sentiment analysis process. In addition, the lack of annotated corpora and sentiment lexicons, as well as issues related to encoding, spelling, and mixed writing systems, limit the effectiveness of existing models. To address these challenges, a comprehensive approach is proposed, including the expansion of language resources, the application of preprocessing techniques that account for morphological and semantic features, the use of context-aware deep learning models, and the development of aspect-based analytical methods. The results of this study provide both theoretical and practical foundations for advancing sentiment analysis in the Uzbek language and serve as a methodological basis for future research.

KEYWORDS: Aspect-Based Sentiment Analysis, sentiment analysis, Uzbek language, NLP, linguistic features, agglutinative language, morphological analysis, contextual models, dataset, preprocessing.

INTRODUCTION

In the modern era of rapidly advancing digital technologies, millions of textual data are generated every day through social networks, electronic platforms, and online reviews. Extracting users' opinions and sentiments from this data has become highly important for business, politics, and social research. Sentiment analysis is the process of automatically identifying and classifying emotions and opinions expressed in texts, typically categorizing them as positive, negative, or neutral. One of the main reasons for the growing popularity of sentiment analysis is that people tend to rely on others' opinions when making rational decisions. As a result, consumers are more likely to purchase products or services that receive the highest ratings.

Aspect-Based Sentiment Analysis (ABSA) is a more advanced form of sentiment analysis that aims to identify sentiments separately for different aspects (features or attributes) mentioned in a text. For example, in the sentence "The food was very delicious, but the service was slow" from a restaurant review, there is a positive sentiment toward the food aspect and a negative sentiment toward the service aspect.

When performing sentiment analysis on Uzbek language texts, the unique linguistic characteristics of the language play a crucial role. In particular, the agglutinative nature of Uzbek, its rich morphology, and the large number of word forms significantly complicate the process of automatic text processing. In Uzbek, words are formed by sequentially adding affixes

to a root, which leads to a dramatic increase in the number of word forms. This, in turn, results in the problem of data sparsity during text processing [1]. Moreover, the relatively free syntactic structure of the Uzbek language and the fact that emotional expressions often depend on context create additional challenges for sentiment analysis systems. These factors limit the effectiveness of traditional approaches and necessitate the use of modern models capable of deep contextual understanding. Therefore, thoroughly studying the linguistic features of the Uzbek language and adapting them to sentiment analysis processes is one of the key directions of this research [2].

Implementing aspect-based sentiment analysis for the Uzbek language presents a number of specific challenges. These challenges are associated with the language's morphological, syntactic, and semantic features, as well as cultural context, technical limitations, and the lack of available resources.

Literature Review

The task of performing aspect-based sentiment analysis on Uzbek language texts is one of the important directions in the field of Natural Language Processing (NLP), and extensive research has been conducted worldwide in this area. In the early stages of sentiment analysis, texts were classified at a general level into positive, negative, and neutral categories. However, in order to achieve more precise results, aspect-based approaches were later developed. Within the framework of the SemEval international evaluation platform organized by Pontiki et al. (2014), ABSA tasks were divided into several stages, highlighting key problems such as aspect detection, categorization, and sentiment classification [3]. These studies played a significant role in the formation and development of the ABSA field.

In recent years, significant progress has been achieved in sentiment analysis with the development of deep learning models. In particular, the BERT model proposed by Devlin et al. (2019) marked a new stage in the development of context-sensitive language models [4]. This model enables a deeper understanding of context by learning text bidirectionally and demonstrates high performance in sentiment analysis tasks, including ABSA. Additionally, models based on LSTM and attention mechanisms have also been proven effective in capturing long-range dependencies in text. However, these models have mainly been developed for English and other resource-rich languages, and their effectiveness may be relatively limited for low-resource languages.

Significant progress has also been made in performing sentiment analysis in Turkic languages. For instance, in the case of the Turkish language alone, there are 78 relevant studies and more than 80 datasets available for sentiment analysis. Nevertheless, a number of linguistic challenges arise in this process. For example, Akylbek Maxutov et al. conducted a study to evaluate the performance of existing large language models (LLMs) in the Kazakh language. They highlighted that the presence of eight types of possessive agreement and the rules of vowel and consonant harmony further complicate the morphological structure, thereby affecting the accuracy of sentiment analysis [5].

The issue of sentiment analysis in low-resource languages requires special attention. In this regard, Kuriyozov et al. (2019) conducted one of the first studies on creating and evaluating a sentiment dataset for the Uzbek language [6]. In their research, a corpus of Uzbek texts was compiled, annotated with sentiment categories, and tested using various classification models.

The results showed that the lack of data and the complex morphological structure of the Uzbek language are among the main challenges in performing sentiment analysis. Furthermore, Mengliev et al. (2021) developed a morphological analysis system for the Uzbek language, demonstrating the complexity of affixation processes and their impact on NLP tasks [6]. They also proposed a rule-based approach for sentiment analysis in Uzbek and noted that factors such as word order, stress, expressions of politeness, and idiomatic expressions can create difficulties in accurately identifying sentiment [7].

The linguistic characteristics of the Uzbek language have a direct impact on the sentiment analysis process. In agglutinative languages, the large number of word forms and the semantic load carried by affixes lead to the problem of data sparsity. This reduces the effectiveness of traditional approaches such as the "bag-of-words" model. At the same time, the widespread use of idiomatic expressions, metaphors, and context-dependent constructions in Uzbek complicates sentiment detection. Some studies (Salaev, 2024) have emphasized that applying morphological normalization and lemmatization techniques can improve model performance [8].

In general, the analysis of existing literature shows that although significant achievements have been made in the field of ABSA worldwide, this area remains insufficiently developed for low-resource languages such as Uzbek. In particular, there is a need to develop specialized models that take into account the linguistic characteristics of the Uzbek language, create annotated datasets, and apply context-sensitive approaches. Therefore, this study aims to address these gaps and contribute to the improvement of ABSA methodologies for the Uzbek language.

Linguistic and Technical Factors Affecting Sentiment Analysis in the Uzbek Language

The analysis of factors influencing the sentiment analysis process in the Uzbek language shows that NLP tasks for this language, particularly ABSA, involve multifaceted complexities. First of all, the lack of annotated corpora and sentiment lexicons in Uzbek is one of the main challenges. Due to the limited availability of resources, the ability to effectively train models is reduced, which negatively affects the overall accuracy of the results. In addition, the agglutinative nature of the Uzbek language leads to morphological complexity. At the same time, phenomena such as lemmatization and polysemy make it difficult to accurately identify sentiment. For example, the same word may carry different sentiment values depending on the context.

Syntactic and semantic factors also play an important role. The relatively free word order in Uzbek and its strong dependence on context limit the effectiveness of traditional models. In particular, negation constructions (e.g., "not bad") and implicit aspects create serious challenges in sentiment detection. Furthermore, metaphorical and culturally specific expressions introduce additional complexity in accurately identifying sentiment.

Table 1. Linguistic and Technical Factors Affecting Sentiment Analysis in the Uzbek Language

Problem Type	Description	Example	Impact on Sentiment Analysis
Lack of language resources	Annotated datasets and corpora are scarce	"The food was delicious, but the service was slow"	Reduces model training quality
Limited sentiment lexicons	Sentiment dictionaries for words are not fully available	"great", "regrettable"	Limits lexical-based approaches

Absence of aspect lexicons	Domain-specific terms are not systematized and compiled into lexicons	"service quality", "battery capacity"	Decreases ABSA accuracy
Morphological complexity	Words take multiple affixes	good, better	Creates difficulty in sentiment detection
Polysemy	A word can have multiple meanings	"terrible"	May lead to incorrect sentiment classification
Context dependency	Meaning changes depending on context	"This is very good"	Simple models may produce more errors
Negation words	Sentiment polarity is altered	"not bad"	Causes reversed interpretation issues
Implicit aspects	Aspects are not explicitly mentioned	"we waited a long time"	Leads to implicit sentiment challenges
Idiomatic expressions	Everyday expressions are widely used	"no words", "it's unbelievable"	Causes semantic misinterpretation
Encoding issues	Special characters may be incorrectly rendered	"O'zbek" → "Õzbek"	Leads to preprocessing errors
Latin–Cyrillic mixing	Two scripts are used together	"Бу kitob yaxshi"	Disrupts tokenization
Spelling errors	Informal writing variations	"yakhshi"	Reduces model accuracy

Overall, these factors indicate the necessity of developing specialized approaches, adapted models, and rich language resources in order to effectively perform sentiment analysis in the Uzbek language.

Solutions and Recommendations

In order to effectively perform sentiment analysis on Uzbek language texts, it is necessary to address the aforementioned linguistic, technical, and methodological challenges through a comprehensive approach. In this regard, a number of scientific and practical solutions are proposed:

Table 2. Solutions and Recommendations for ABSA in the Uzbek Language

Problem Area	Proposed Solutions and Approaches
Lack of language resources	Create large-scale annotated datasets in Uzbek, develop open corpora, collect data across various domains (e-commerce, education, healthcare), and perform sentiment and aspect annotation
Limited sentiment and aspect lexicons	Develop a national sentiment lexicon, include synonyms, phraseological units, and emotional expressions, and build an aspect-oriented terminological base
Syntactic flexibility and context issues	Apply context-sensitive models (LSTM, BiLSTM, Transformer, BERT), particularly fine-tuned mBERT or UzBERT models
Negation and contextual sentiment	Develop negation handling mechanisms and implement attention-based models that consider context

Multi-word and implicit aspects	Use sequence labeling methods (CRF, BiLSTM-CRF), and identify aspects through attention mechanisms and dependency parsing
Encoding and script issues	Normalize text, develop Latin–Cyrillic conversion modules, and standardize special characters
Spelling errors and informal texts	Introduce automatic spell-checkers and text normalization systems, and develop preprocessing techniques suitable for social media texts
Improving model performance	Apply ensemble approaches (combining multiple models), and optimize using hyperparameter tuning and cross-validation

Conclusion

This paper provides a comprehensive analysis of the linguistic, technical, and methodological challenges that arise in performing aspect-based sentiment analysis in the Uzbek language. The analysis shows that the agglutinative structure of Uzbek, its rich morphology, flexible word order, and the presence of context-dependent emotional expressions are among the key factors that complicate the sentiment analysis process. In particular, negation constructions, polysemy, idiomatic expressions, and implicit aspects introduce ambiguities in sentiment detection. Furthermore, the lack of annotated corpora, sentiment and aspect lexicons, the coexistence of different writing systems, and issues related to encoding and spelling have been identified as significant factors that reduce the effectiveness of sentiment analysis models. These findings highlight the limitations of existing traditional approaches and emphasize the need to develop specialized methodologies tailored to the Uzbek language.

Based on the research findings, several solutions and recommendations have been proposed to advance sentiment analysis in the Uzbek language. These include the creation of large-scale annotated datasets, the implementation of preprocessing techniques that consider morphological and semantic features, the use of context-sensitive deep learning models, and the development of aspect-based approaches. Additionally, incorporating cultural and pragmatic factors, standardizing technical issues, and developing domain-specific models have been identified as important directions for improving the performance of sentiment analysis systems.

Overall, this study establishes both theoretical and practical foundations for the development of aspect-based sentiment analysis in the Uzbek language and serves as a methodological basis for future research in this field.

REFERENCES

1. N. Abdurakhmonova, “Morphological Analysis and Generation for the Uzbek Language,” Applied Sciences, vol. 11, no. 21, Art. no. 10267, 2021.
2. E. Kuriyozov, S. Matlatipov, M. A. Alonso, C. Gómez-Rodríguez, “Construction and Evaluation of Sentiment Datasets for Low-Resource Languages: The Case of Uzbek,” in Lecture Notes in Computer Science, Springer, 2019, pp. 232–243.
3. M. Pontiki et al., “SemEval-2014 Task 4: Aspect Based Sentiment Analysis,” in Proc. COLING 2014, 2014, pp. 27–35.

4. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

5. A. Maxutov, A. Myrzakhmet, P. Braslavski, "Do LLMs Speak Kazakh? A Pilot Evaluation of Seven Models," in Proc. SIGTURK Workshop, 2024, pp. 81–91.

6. D. Mengliev et al., "Morphological Analysis for the Uzbek Language," Applied Sciences, vol. 11, no. 21, Art. no. 10267, 2021.

7. D. B. Mengliev, E. Y. Akhmedov, V. B. Barakhnin, Z. A. Hakimov, O. M. Alloyorov, "Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language," in Proc. IEEE APEIE, 2023, pp. 1720–1724.

8. U. Salaev, "Morphological Analysis Model for Uzbek Language," arXiv preprint, arXiv:2405.14179, 2024.