PRINCIPLES OF USING LINGUISTIC RESOURCES IN THE SENTIMENT ANALYSIS PROCESS OF UZBEK TEXTS

Rakhimov Hasanboy Komiljonovich
PhD Student (3rd year), Andijan State University (ADU) Trainee Lecturer,
Namangan State Institute of Foreign Languages

Abstract: This article deeply explores the principles of utilizing linguistic resources in the sentiment analysis process of Uzbek texts. Considering the agglutinative nature of the Uzbek language, vowel harmony, and cultural context, the role of lexical databases, corpora, and morphological analyzers is analyzed. The article examines the stages of data preparation, feature extraction, classification, and aspect-based sentiment analysis (ABSA). It also discusses challenges in low-resource languages and future prospects, including hybrid approaches and the development of language models. These principles, grounded in empirical and theoretical methods of linguistics, serve to enhance the accuracy and applicability of the analysis.

Keywords: Sentiment analysis, Uzbek language, linguistic resources, natural language processing (NLP), agglutinative language, corpora, lexical databases, morphological analysis, aspect-based sentiment analysis (ABSA), machine learning models.

The principles of using linguistic resources in the sentiment analysis of Uzbek texts represent an important aspect of modern language technologies, aimed at determining the emotional orientation of texts based on the theoretical foundations of linguistics. Sentiment analysis, or text sentiment analysis, is one of the main tasks in the field of natural language processing (NLP), applied in analyzing data from social networks, customer reviews, and mass media by identifying positive, negative, or neutral tones in texts. In low-resource languages like Uzbek, this process is more complex, because the language's agglutinative nature, vowel harmony, and the influence of borrowed words from Persian, Arabic, and Russian require linguistic resources. Linguistic resources, such as lexical databases, corpora, and morphological analyzers, play a central role in increasing the accuracy of sentiment analysis, as they account for the syntactic, semantic, and pragmatic layers of the language. These principles, based on empirical and theoretical methods of linguistics, are applied considering the cultural context, idiomatic expressions, and syntactic structures of Uzbek texts, ensuring the depth of analysis.

The linguistic features of the Uzbek language define the principles of using linguistic resources in sentiment analysis. Uzbek belongs to the Turkic language family and has

an agglutinative structure, meaning words are formed through numerous affixes (suffixes), which can alter the emotional orientation of words. For example, the word “yaxshi” expresses positive sentiment, while the negation form “yaxshi emas” turns it negative, or comparative degrees like “zo‘rroq” add nuances. Vowel harmony and dialectal differences (for example, Andijan and Qashqadaryo dialects) complicate the interpretation of texts, so linguistic resources are necessary for modeling these features. From a linguistic perspective, these resources cover the phonetic, morphological, and semantic levels of the language, for example, morphological analyzers like MorphUz separate affixes and identify word roots. Such an approach in low-resource languages, including Uzbek, when combined with pre-trained models (for example, mBERT or XLM-RoBERTa), can achieve analysis accuracy of 72–88%.

One of the main principles of using linguistic resources lies in the data preparation and cleaning stage, where corpora and lexical databases take center stage. For sentiment analysis in Uzbek, corpora, such as those collected from social networks and Google Play reviews, provide a natural distribution of texts. For example, corpora like UzSentiment include 2500 positive and 1800 negative reviews, manually annotated to define emotional polarity. In creating these corpora, linguistic resources, including stop-word lists (731 156 terms) and tokenization tools, are used to break texts into word and character-level n-grams. In preprocessing, stemming and lemmatization, considering the morphological complexity of Uzbek, reduce words to their base forms, making TF-IDF vectorization effective in feature extraction. According to linguistic principles, this stage solves problems related to the agglutinative nature of the language, for example, by separating affixes and matching word roots with emotional dictionaries.

In the feature extraction and classification stages, the principles of linguistic resources become more profoundly evident, where lexical databases and semantic models are integrated. For sentiment classification in Uzbek, three main dictionaries are used: the first includes a dictionary for forming word shapes with over 300 affixes; the second is a catalog of exception words that do not conform to standard morphological rules; the third is a dictionary of word roots consisting of 88 879 entries, some of which are marked with positive or negative polarity. These resources are applied in a rule-based algorithm: the sentence is divided into word forms, compared with the exception dictionary, if no match, stemmed to the root, and matched with the emotional dictionary. If no match is found after three stemmings, the word is considered unknown. Sentiment polarity is determined based on the number of positive and negative roots, identifying nuanced emotions considering the complex morphological structures of Uzbek. Such an approach, when combined with machine learning models

(for example, Naive Bayes, SVM, or LSTM), effectively handles dialectal differences and negation forms.

In aspect-based sentiment analysis (ABSA), the principles of linguistic resources are applied in a more advanced form, where aspects (for example, food, service, price) and their polarities are analyzed separately. The UzABSA database is the first annotated resource in Uzbek, including 6500 document-level and 6175 sentence-level instances from restaurant reviews. Annotation is based on SemEval-2014 guidelines, covering four tasks: extracting aspect terms, determining their polarities, identifying aspect categories, and assigning polarities to categories. Linguistic resources are applied here through the BRAT tool, with inter-annotator agreement evaluated by Cohen's kappa (0.72–0.83) and Krippendorff's alpha (0.55–0.88). These principles, considering the low-resource nature of Uzbek, model semantic relations by integrating graph convolutional networks (GCN) and mT5 models, allowing for deeper analysis of grammatical structures.

The problems and future prospects of the principles of using linguistic resources determine the development of sentiment analysis in Uzbek. The main problems are the language's low-resource status, limited databases, and dialectal differences, which reduce the generalization ability of models. For example, cultural context and pragmatic nuances (for example, sarcasm) are not fully covered by standard dictionaries, so hybrid approaches – combining lexicon-based and deep learning models – are proposed. In the future, through developing language models like UzLM, pre-training on corpora of 80 million words is possible, integrating sentiment analysis with other NLP tasks (for example, speech recognition). These principles, based on empirical linguistic research, play an important role in adapting Uzbek to global NLP standards, resulting in increased accuracy and applicability of analysis.

References

1. Matlatipov, S. G., Rajabov, J., Kuriyozov, E., & Aripov, M. (2024). UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language. In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024 (pp. 394–403). Torino, Italia: ELRA and ICCL.
2. Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek Sentiment Analysis based on local Restaurant Reviews. In Proceedings of the ALTNLP: The International Conference and workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (pp. 126–136). CEUR Workshop Proceedings, Vol. 3315.
3. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2022). Construction and evaluation of sentiment datasets for low-resource languages: The

case of Uzbek. Natural Language Engineering and Computational Linguistics (Springer series).

4. Kuriyozov, E., & Matlatipov, S. (2019). Building a New Sentiment Analysis Dataset for Uzbek Language and Creating Baseline Models. Proceedings, 21(1), 37.

5. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2019). Deep Learning vs. Classic Models on a New Uzbek Sentiment Analysis Dataset. In Human Language Technologies as a Challenge for Computer Science and Linguistics.