AMERICAN JOURNALS
Open Access,
Peer Reviewed Journals

American Journal of Technology and Applied Sciences
ISSN (E): 2832-1766 Volume 31, December - 2024

ADVANCING OPINION MINING FOR LOW-RESOURCE LANGUAGES: A CASE STUDY OF UZBEK ABSA DATASET

Rajabov Jaloliddin Shamsuddin o‘g‘li
PhD Student at the National University of Uzbekistan named after Mirzo Ulugbek

ABSTRACT

The objective of enhancing the availability of natural language processing technologies for low-resource languages has significant importance in facilitating technological accessibility within the populations of speakers of these languages. This study addresses the gap in linguistic resources for the Uzbek language by introducing UzABSA, the first high-quality annotated aspect-based sentiment analysis (ABSA) dataset. The dataset comprises 3,500 document-level reviews and over 6,100 sentence-level instances, collected from Uzbek restaurant reviews. Evaluation of the annotation process using Cohen's kappa and Krippendorff's α demonstrates robust agreement levels. A classification model, K-Nearest Neighbour (KNN), was employed to validate the dataset, achieving accuracy rates of 72% to 88%. This pioneering work provides a foundational resource for advancing sentiment analysis for the Uzbek language.

KEYWORDS

Natural Language Processing, Uzbek Language, ABSA Dataset, Sentiment Analysis, Low-Resource Languages, UzABSA, Annotation Techniques, KNN Classification, Statistical Evaluation

Introduction

The increasing adoption of natural language processing (NLP) technologies in various applications underscores the need to extend these tools to low-resource languages, including Uzbek. Despite Uzbek being the native language for over 35 million speakers, the lack of linguistic resources impedes technological advancements tailored to this population. Aspect-based sentiment analysis (ABSA) is a critical NLP task that identifies opinions concerning specific aspects within texts. However, no open-source ABSA datasets exist for Uzbek. This study introduces UzABSA, a comprehensive dataset designed to address this gap and support further research in the domain.

Dataset Development

Data Collection

The dataset consists of online reviews sourced from Uzbek restaurant websites. It encompasses 3,500 reviews at the document level and over 6,100 sentences at the sentence level. Reviews were preprocessed to remove noise and ensure linguistic clarity.

Annotation Process

Annotations focused on four key attributes:

- Aspect Terms: Identifying specific components discussed in the text.
- Aspect Term Polarities: Determining sentiment polarity for each aspect term.
- Aspect Category Terms: Categorizing the broader topics.
- Aspect Category Polarities: Assigning sentiment polarity to each category.

Two independent annotators participated in the annotation process. Agreement levels were measured using Cohen's kappa coefficient and Krippendorff's α, yielding values of 0.81 and 0.79, respectively, indicating high reliability.

Model Implementation

To validate UzABSA, a K-Nearest Neighbour (KNN) classification model was implemented. Training and evaluation were conducted on the dataset, and metrics such as accuracy, precision, recall, and F1-score were computed. Results ranged from 72% to 88%, demonstrating the dataset's efficacy for sentiment analysis tasks.

Significance of UzABSA

UzABSA represents the first and largest ABSA dataset for the Uzbek language. It sets a benchmark for future studies and facilitates the development of NLP applications, such as recommendation systems and automated feedback analysis, tailored to Uzbek speakers.

Evaluation Metrics

To establish the quality and reliability of the annotated dataset, multiple evaluation metrics were employed. The Cohen's kappa coefficient of 0.81 and Krippendorff's alpha of 0.79 demonstrated robust inter-annotator agreement. The dataset also underwent a preliminary evaluation using a K-Nearest Neighbour (KNN) classifier. The classifier achieved an accuracy range between 72% to 88%, reflecting the dataset's practical applicability in real-world sentiment analysis scenarios. These outcomes underscore the dataset's robustness and potential to support advanced NLP tasks.

Applications of UzABSA

The dataset has significant implications for both academic and industrial applications. Academically, it lays the groundwork for further research in ABSA and sentiment analysis for low-resource languages. In industry, UzABSA can support the development of applications like review analytics, customer sentiment tracking, and opinion mining in Uzbek. Furthermore, the dataset can serve as a resource for developing recommendation systems tailored to the preferences and sentiments of Uzbek-speaking users.

Comparative Analysis

When compared to existing datasets for other languages, UzABSA exhibits unique characteristics. Unlike high-resource languages, where multiple ABSA datasets exist, UzABSA pioneers this effort for Uzbek. Additionally, its meticulous annotation process and high inter-annotator agreement metrics position it as a benchmark dataset. Future works can extend this dataset by incorporating multimodal data, such as images and audio reviews, to enrich the analysis.

Results and Discussions

The evaluation results underscore the robustness of UzABSA as a resource for ABSA tasks. High agreement among annotators ensures the dataset's reliability, while the KNN model's performance validates its applicability. The study highlights the need for further experimentation with advanced machine learning models, such as transformers, to enhance accuracy and scalability. Moreover, the dataset's potential applications in commercial and academic settings underline its significance.

Conclusion

UzABSA addresses a critical gap in linguistic resources for the Uzbek language. By providing a high-quality, annotated ABSA dataset, this study lays the groundwork for advancing NLP technologies for low-resource languages. Future work will explore integrating UzABSA with state-of-the-art NLP models to achieve greater precision and utility. The dataset's contribution will also extend to cross-lingual NLP research, fostering collaboration among researchers in the global community.

References

1. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46.
2. Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology. Sage Publications.
3. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.
4. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.
7. Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to Information Retrieval. Cambridge University Press.
8. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing. Pearson.
9. Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008.
10. Uzbek Restaurant Reviews Dataset. (2024). Collected from online review platforms for this study.