UDK: 811.111.111`11

DATA SCIENCE IN CORPUS LINGUISTICS

Babayev Saidmuxammadjon Saidkamolovich
Master student at Urgench State University
saidmuhammadbabayev@gmail.com
Ruzimboyev Xusniddin Raximberganovich
Master student at Urgench State University
khusniddindatascientistn1@gmail.com

Annotatsiya. Ma’lumotlar ilmi (Data science) va korpus lingvistikasi so‘nggi yillarda tobora o‘zaro bog‘langan ikkita sohadir. Ushbu maqola til ma’lumotlarining tahlil qilish uchun ma’lumotlar ilmi va tabiiy tillarni qayta ishlash (NLP) texnnologiyasidan foydalangan so‘nggi tadqiqotlarni ko‘rib chiqish orqali ushbu sohalarning kesishishini o‘rganadi. Biz ushbu usullardan Korpus lingvistikasida an’anaviy ravishda qo‘lda bajariladigan ko‘plab vazifalarni avtomatlashtirish uchun qanday foydalanish mumkinligini va ulardan ma’lumotlardagi qolip va tendensiyalarni aniqlash uchun qanday foydalanish mumkinligini muhokama qilamiz. Shuningdek, his-tuyg‘ularni vaqt o‘tishi bilan til o‘zgarishini tahlil qilish va til ma’lumotlariga asoslangan bashoratli modellarni yaratish uchun ma’lumotlar ilmi va NLP texnikasining imkoniyatlarini o‘rganamiz. Va nihoyat, biz ushbu sohalarning kelajagi va NLP va Data Science texnikasidan foydalanish bilan bog‘liq axloqiy va maxfiylik muammolarini hal qilish zarurligini muhokama qilamiz.

Kalit so‘zlar: *NLP, ma’lumotlar ilmi, Korpus lingvistikasi, Data mining.*

Abstract. Data Science and Corpus Linguistics are two fields that have become increasingly intertwined in recent years. This paper explores the intersection of these fields by reviewing recent studies that have used Data Science and NLP techniques to analyze large corpora of language data. We discuss how these techniques can be used to automate many of the tasks traditionally done manually in Corpus Linguistics, and how they can be used to identify patterns and trends in the data. We also explore the potential of Data Science and NLP techniques for analyzing sentiment, language change over time, and for building predictive models based on language data. Finally, we discuss the future of these fields and the need to address ethical and privacy concerns related to the use of NLP and Data Science techniques.

Keywords: *NLP, Data Science, Corpus Linguistics, Data Mining.*

Аннотация. Наука данных и корпусная лингвистика — две области, которые в последние годы все больше переплетаются. В этой статье рассматривается пересечение этих областей путем обзора недавних исследований, в которых использовались методы наука данных и NLP для анализа больших корпусов языковых данных. Мы обсудим, как эти методы

можно использовать для автоматизации многих задач, традиционно выполняемых вручную в корпусной лингвистике, и как их можно использовать для выявления закономерностей и тенденций в данных. Мы также изучаем потенциал методов наука данных и NLP для анализа настроений, языковых изменений с течением времени и для построения прогностических моделей на основе языковых данных. Наконец, мы обсуждаем будущее этих областей и необходимость решения проблем этики и конфиденциальности, связанных с использованием методов NLP и наука данных.

Ключевые слова: *NLP, наука данных, корпусная лингвистика, анализ данных.*

Introduction.

Corpus Linguistics is a subfield of linguistics that involves the analysis of large collections of language data, known as corpora. Data Science, on the other hand, involves using statistical and computational methods to extract insights and knowledge from data. Together, these two fields have the potential to revolutionize the way that we analyze and understand language. Corpus Linguistics has traditionally relied on manual methods for analyzing language data. However, with the advent of Data Science techniques, it is now possible to automate many of these tasks, which can save time and improve accuracy. For example, Data Science techniques can be used to analyze the frequency of words and phrases in a corpus, which can help researchers to identify patterns and trends in the data.

One area where Data Science has been particularly impactful in Corpus Linguistics is in the analysis of sentiment. Sentiment analysis involves the use of computational methods to identify the emotional tone of a piece of text. This can be useful for researchers who are interested in analyzing the attitudes and opinions expressed in a corpus. Data Science techniques can be used to analyze large corpora of text and identify patterns of sentiment, which can then be used to draw conclusions about the attitudes and opinions of the speakers or writers.

Another area where Data Science has been used in Corpus Linguistics is in the analysis of language change over time. By analyzing large corpora of text from different time periods, researchers can identify changes in the frequency of words and phrases, as well as changes in the structure and grammar of the language. Data Science techniques can be used to automate many of these tasks, which can save time and improve the accuracy of the analysis.

Data Science can also be used to build predictive models based on language data. For example, machine learning algorithms can be used to identify patterns in a corpus of text and make predictions about future language use. This has the potential to be useful in a variety of fields, such as marketing, where predictive models can be used to anticipate changes in consumer behavior.

Related Work.

Several researchers have explored the intersection of Data Science and Corpus Linguistics in recent years. For example, Li et. al. [1] used a combination of machine learning algorithms and natural language processing techniques to analyze a corpus of Chinese microblog data. Their study focused on identifying patterns of sentiment and emotion in the data, and they found that their approach was able to accurately predict the emotional tone of the microblogs with high accuracy.

In another study, Tadesse et. al. [2] used machine learning algorithms to analyze a corpus of English-language tweets. They focused on identifying patterns of opinion and emotion in the data, and found that their approach was able to accurately classify tweets into different sentiment categories.

These studies demonstrate the potential of Data Science and NLP techniques for analyzing language data in new and innovative ways. They also highlight the importance of developing accurate and efficient algorithms for processing and analyzing large corpora of text. As the field of Data Science and NLP continues to evolve, we can expect to see even more exciting applications of these techniques in the field of Corpus Linguistics.

Regarding the advances in Uzbek NLP resources that make use of Data science algorithms, a number of works have been done in such manner such as sentiment analysis [3], [4], morphological tagged corpus [5], as well as text summarization [6], [7]. Furthermore, there has been a recent rapid increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for removal of stopwords [8], [9], Latin-Cyrillic transliteration [10] and text classification [11], as well as tools for stemming [12], [13], lemmatization [14].

Methodology.

NLP tasks that can benefit from Data Science techniques include many of the traditional tasks of NLP, such as sentiment analysis, named entity recognition, and part-of-speech tagging. Machine learning algorithms can be used to automate many of these tasks, making it possible to process large volumes of text data quickly and efficiently. In addition, deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown great promise for tasks such as sentiment analysis and text generation. Clustering algorithms can be used to identify patterns and groups within large corpora of text, while topic modeling algorithms such as Latent Dirichlet Allocation (LDA) can be used to identify latent themes and topics within a corpus. Finally, word embedding techniques such as Word2Vec and GloVe can be used to represent words as vectors in a high-dimensional space, enabling semantic similarity and distance calculations, which are essential for many NLP tasks.

Following are the algorithms in the field of Data science that can be beneficial for NLP tasks:

- Naive Bayes: A probabilistic algorithm commonly used for text classification tasks, such as sentiment analysis or spam detection.

- Support Vector Machines (SVM): A supervised learning algorithm that can be used for text classification, named entity recognition, and other tasks.
- Random Forests: An ensemble learning algorithm that can be used for text classification, topic modeling, and other tasks.
- Convolutional Neural Networks (CNNs): A deep learning algorithm that can be used for sentiment analysis, text classification, and other tasks.
- Recurrent Neural Networks (RNNs): A deep learning algorithm that can be used for language modeling, text generation, and other tasks.
- Latent Dirichlet Allocation (LDA): A topic modeling algorithm that can be used to discover latent themes or topics in a corpus of text.
- Hierarchical clustering: An unsupervised learning algorithm that can be used for text clustering, topic modeling, and other tasks.
- k-means clustering: An unsupervised learning algorithm that can be used for text clustering and topic modeling.
- Principal Component Analysis (PCA): A dimensionality reduction technique that can be used for text visualization and clustering.
- t-SNE: A dimensionality reduction technique that can be used for text visualization and clustering.

Conclusion and Future work.

In conclusion, Data Science has the potential to revolutionize the field of Corpus Linguistics by providing new methods for analyzing language data. By automating many of the tasks that were traditionally done manually, researchers can save time and improve the accuracy of their analysis. Additionally, Data Science techniques such as sentiment analysis and predictive modeling can provide new insights into language use and change over time. As these fields continue to evolve, we can expect to see even more exciting applications of Data Science in Corpus Linguistics.

The future work in the field of Data Science and NLP will likely focus on advancing the accuracy and efficiency of these techniques. As the amount of unstructured data continues to grow, it will become increasingly important to develop methods for processing and analyzing this data in real-time. This will require the development of more sophisticated machine learning algorithms and the use of distributed computing frameworks to process large datasets. Additionally, there will be a need to address ethical and privacy concerns related to the use of NLP and Data Science techniques, such as the potential for bias and the need to protect sensitive information. As these fields continue to evolve, it will be important to stay up-to-date with the latest research and developments in order to take advantage of the full potential of NLP and Data Science.

References

1. X. Li, J. Zhang, Y. Du, J. Zhu, Y. Fan, and X. Chen, “A Novel Deep Learning-based Sentiment Analysis Method Enhanced with Emojis in Microblog Social Networks,” *Enterp Inf Syst*, pp. 1–22, 2022.

2. M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44883–44893, 2019.

3. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com

4. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.

5. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.

6. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.

7. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.

8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.

9. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” Univerza na Primorskem, Inštitut Andrej Marušič, 2021.

10. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com

11. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

12. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.

13. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.

14. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022