DATA MINING TECHNIQUES IN NATURAL LANGUAGE PROCESSING: A REVIEW

Qurbonova Ro‘zikajon Ulug‘bek qizi
ruzikajonqurbanova0809@gmail.com,
Ro‘zimboyeva Sevara Nurmat qizi
rozimboyevasevara@gmail.com
Master students at Urgench State University

Annotatsiya. So‘nggi yillarda ijtimoiy tarmoqlar, bloglar va yangilik maqolalari kabi turli manbalardan olingan matnli ma’lumotlar miqdorida katta o‘sish sodir bo‘ldi. Ma’lumotlarning ko‘payishi bilan mazmunli tushunchalarni olish uchun samarali va tez tabiiy tilni qayta ishlash (NLP) usullariga ehtiyoj paydo bo‘ladi. Katta hajmdagi ma’lumotlar to‘plamlaridan namunalar va bilimlarni olishni o‘z ichiga olgan ma’lumotlar ilmi texnikasi NLP ilovalarida foydali ekanligini isbotladi. Ushbu maqolada biz NLP vazifalariga qo‘llaniladigan turli xil ma’lumotlarni yig‘ish usullarini ko‘rib chiqamiz, jumladan matn tasnifi (klassifikatsiya), hissiyotlarni tahlil qilish (sentiment), obyektni tanib olish (NER) va mavzuni modellashtirish. Biz har bir texnikaning afzalliklari va cheklovlarini muhokama qilamiz va NLPda ma’lumotlar qazib olishdan foydalanish bilan bog‘liq ba’zi qiyinchiliklarni ta’kidlaymiz. Va nihoyat, biz ushbu sohadagi kelajakdagi tadqiqot yo‘nalishlari haqida ba’zi fikrlarni taqdim etamiz.

Kalit so‘zlar: *Tabiiy tilni qayta ishlash, ma’lumotlar ilmi, katta hajmdagi ma’lumotlar.*

Аннотация. В последние годы произошел взрывной рост количества текстовых данных, генерируемых из различных источников, таких как социальные сети, блоги и новостные статьи. С этим увеличением данных возникает потребность в эффективных и действенных методах обработки естественного языка (NLP) для извлечения осмысленной информации. Методы интеллектуального анализа данных, которые включают извлечение шаблонов и знаний из больших наборов данных, оказались полезными в приложениях НЛП. В этой статье мы рассмотрим различные методы интеллектуального анализа данных, которые применялись к задачам НЛП, включая классификацию текста, анализ настроений, распознавание сущностей и тематическое моделирование. Мы обсудим преимущества и ограничения каждого метода и выделим некоторые проблемы, связанные с использованием интеллектуального анализа данных в НЛП. Наконец, мы даем некоторое представление о будущих направлениях исследований в этой области.

Ключевые слова: *Обработка естественного языка, интеллектуальный анализ данных, большие данные.*

Abstract. In recent years, there has been an explosion in the amount of textual data generated from various sources, such as social media, blogs, and news articles.

With this increase in data comes the need for efficient and effective natural language processing (NLP) techniques to extract meaningful insights. Data mining techniques, which involve the extraction of patterns and knowledge from large data sets, have proven to be useful in NLP applications. In this paper, we review various data mining techniques that have been applied to NLP tasks, including text classification, sentiment analysis, entity recognition, and topic modeling. We discuss the advantages and limitations of each technique, and highlight some of the challenges associated with using data mining in NLP. Finally, we provide some insights into future research directions in this area.

Keywords: *Natural Language Processing, Data Mining, Big Data*

1. Introduction.

The vast amounts of textual data available in today's world present both an opportunity and a challenge for researchers and practitioners in the field of natural language processing (NLP) [1]. On one hand, this data can be leveraged to gain valuable insights into human behavior, sentiment, and trends. On the other hand, the sheer volume of data can make it difficult to extract meaningful information from it. Data mining techniques, which have been used extensively in other domains such as marketing and finance, offer a promising approach for tackling this challenge in NLP [2]–[4].

In recent years, NLP has become an increasingly important field due to the explosion of textual data generated from various sources, such as social media, blogs, and news articles [5]. The challenge for NLP researchers and practitioners is how to efficiently and effectively extract meaningful insights from this vast amount of data. Data mining techniques offer a promising approach to tackling this challenge [6]. Data mining is the process of extracting useful and relevant information from large datasets. It involves using advanced computational techniques and algorithms to identify patterns, trends, and relationships within data. The goal of data mining is to extract insights and knowledge from data that can be used for decision-making, prediction, and optimization. Data mining techniques can be applied to various domains such as finance, healthcare, marketing, and engineering [7], [8]. The widespread availability of data and advancements in computing power have led to the growth of data mining as a field, and it has become an essential tool for businesses and researchers alike.

2. Literature Review.

In this section, we review various data mining techniques that have been applied to NLP tasks. There has been a significant amount of research in the area of data mining techniques in NLP. In this section, we review some of the relevant work in this area.

Text classification is one of the most extensively studied NLP tasks. A study conducted by Pang and Lee [9] compared various machine learning algorithms for text classification and found that SVM performed the best. Another study by Wang et al. [10] proposed a novel feature selection method for text classification based on mutual information, which improved the performance of the classification algorithm.

Sentiment analysis has also received significant attention from researchers. A study conducted by Jurek et al. [11] proposed a lexicon-based approach for sentiment analysis, which achieved high accuracy on several benchmark datasets. Another study by Matlatipov et al. [12], [13] proposed a sentiment analysis data and models specific for Uzbek language, which achieved high accuracy for identifying the sentiment polarity of different aspects of a product.

Entity recognition is another important NLP task that has been extensively studied. A study by Finkel et al. [14] proposed a conditional random field model for named entity recognition, which achieved state-of-the-art performance on several benchmark datasets.

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 [15], [16], Latin-Cyrillic transliteration [17] and text classification [18], as well as tools for stemming [19], [20], lemmatization [21].

3. NLP tasks and Data Mining Techniques.

Text classification is the process of categorizing documents into predefined classes based on their content. This technique has been widely used in many NLP applications such as email spam filtering, sentiment analysis, and language identification. Machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Random Forests have been applied to text classification tasks with promising results.

Sentiment analysis is the process of identifying the sentiment expressed in a piece of text, such as positive or negative. Sentiment analysis has many applications such as product reviews analysis, customer feedback analysis, and political sentiment analysis. Machine learning algorithms such as Logistic Regression, SVM, and Neural Networks have been applied to sentiment analysis tasks with high accuracy.

Named Entity recognition (NER) is the process of identifying and extracting named entities from text, such as people, places, and organizations. Entity recognition has many applications such as social network analysis, news article analysis, and customer feedback analysis. Machine learning algorithms such as Conditional Random Fields (CRF), Hidden Markov Models (HMM), and Deep Learning models have been applied to entity recognition tasks with high accuracy.

Topic modeling is the process of identifying the underlying themes or topics in a corpus of text. Topic modeling has many applications such as text summarization, document clustering, and trend analysis. Machine learning algorithms such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been applied to topic modeling tasks with high accuracy.

Despite the promising results of data mining techniques in NLP, there are some challenges associated with using these techniques in practice. One of the main challenges is the need for large amounts of labeled data for training machine learning algorithms. Another challenge is the difficulty of interpreting the results of some techniques, such as neural networks and deep learning models.

4. Conclusion and Future Work.

In conclusion, data mining techniques offer a promising approach for addressing the challenges of extracting meaningful insights from large amounts of textual data in NLP. These techniques can be applied to various NLP tasks such as text classification, sentiment analysis, entity recognition, and topic modeling. However, there is still much work to be done in terms of developing more effective and efficient techniques, as well as addressing some of the challenges associated with using these techniques in practice.

In the future, we expect to see continued growth and development in the field of data mining techniques in NLP. One area that is likely to see significant advancement is deep learning, which has shown promise in many NLP tasks. Furthermore, with the increasing amount of data being generated in various domains, there will be a need for more efficient and scalable data mining techniques. Additionally, we expect to see more research on combining multiple NLP tasks and applying data mining techniques to solve complex problems. Finally, the ethical implications of data mining in NLP will need to be carefully considered, including issues such as privacy, bias, and fairness. Overall, the future of data mining techniques in NLP looks promising, and we anticipate many exciting advancements in this field.

References

1. U. Salaev, E. Kuriyozov, and C. Gómez-Rodr\'\iguez, "SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language," in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022, pp. 199–206. [Online]. Available: www.scopus.com
2. M. Jamolbek Maqsudovich, "Clustering Class Association Rules to form a Meaningful and Accurate Classifier: doctoral dissertation," Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in~…, 2020.
3. M. Jamolbek and M. Sanatbek, "Extracting the hidden regularities on latent features by using interval methods in pattern recognition problems," *European science review*, no. 5–6, pp. 22–23, 2016.
4. J. Mattiev and B. Kavšek, "CMAC: clustering class association rules to form a compact and meaningful associative classifier," in *Machine Learning,*

Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6, 2020, pp. 372–384.

5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekTagger: The rule-based POS tagger for Uzbek language," *arXiv preprint arXiv:2301.12711*, 2023.

6. J. Mattiev and B. Kavšek, "Distance based clustering of class association rules to build a compact, accurate and descriptive classifier," *Computer Science and Information Systems*, vol. 18, no. 3, pp. 791–811, 2021.

7. J. Mattiev and B. Kavšek, "ACHC: Associative Classifier Based on Hierarchical Clustering," in *Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22*, 2021, pp. 560–571.

8. J. Mattiev and B. Kavsek, "Coverage-based classification using association rule mining," *Applied Sciences*, vol. 10, no. 20, p. 7013, 2020.

9. B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

10. B. Wang, Y. Huang, W. Yang, and X. Li, "Short text classification based on strong feature thesaurus," *Journal of Zhejiang University SCIENCE C*, vol. 13, no. 9, pp. 649–659, 2012.

11. A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur Inform*, vol. 4, no. 1, pp. 1–13, 2015.

12. 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

13. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodr\'\iguez, "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.

14. J. R. Finkel and C. D. Manning, "Nested named entity recognition," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 141–150.

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

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

17. U. Salaev, E. Kuriyozov, and C. Gómez-Rodr\'\iguez, "A machine transliteration tool between Uzbek alphabets," in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com

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

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

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

21. 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