Oleksandr Mazurets
Ph.D in Engineering Science, Associate Professor
Roman Vit
Postgraduate student
Khmelnytskyi National University, Ukraine

PRACTICAL APPLICATION OF METHOD OF THEMATIC CLASSIFICATION OF TEXT INFORMATION USING LDA

Abstract. Method of thematic classification of textual information has been developed, examples of the analysis of the effectiveness of the created method on an English-language data set using the corresponding developed software are given. From the thematic modeling in the dataset, as a result of applying the method, a cross-validation check was carried out, which gave a result of 0.86, which is an improvement of 0.15 in comparison with the use of LDA in its pure form for classification.

With the growth of the use of the Internet and digital platforms, the amount of unstructured data has increased significantly, which has increased the need for efficient methods of processing it. Thematic classification of text information allows automating the process of organizing and structuring large volumes of text, which greatly facilitates their further analysis. Thanks to the development of natural language processing and machine learning technologies, new methods have appeared that provide high accuracy and efficiency in the classification of text data.

An approach using machine learning to automate the thematic classification of text information was developed, which emphasizes the possibility of flexible definition of topics and consideration of target objects of the subject area [1]. This method differs from the existing ones by the possibility of flexible definition of topics due to the application of thematic modeling, the formation of an expanded set of keywords due to the inclusion of both keywords found using LDA and target objects of the subject area in the form of a collection of NER and keywords that are not included to the list found by the LDA algorithm.

The purpose of the work is the analysis of applied application of the method of thematic classification of textual information.

The English-language "fake-and-real-news-dataset" divided into two files: "Fake.csv" (23,502 fake articles) and "True.csv" (21,417 true news) was chosen as the research data set. As part of the study, only the "True.csv" part will be used. To validate the proposed approach, a software implementation was created in the form of a Jupyter Notebook, using the Google Colab environment. After carrying out thematic modeling without specifying the number of topics, based on the coherence index, it was determined that the optimal number of dataset topics is 14. From the thematic modeling, a presentation of keywords was obtained, a visual presentation of examples for some topics is shown in Figure 1.

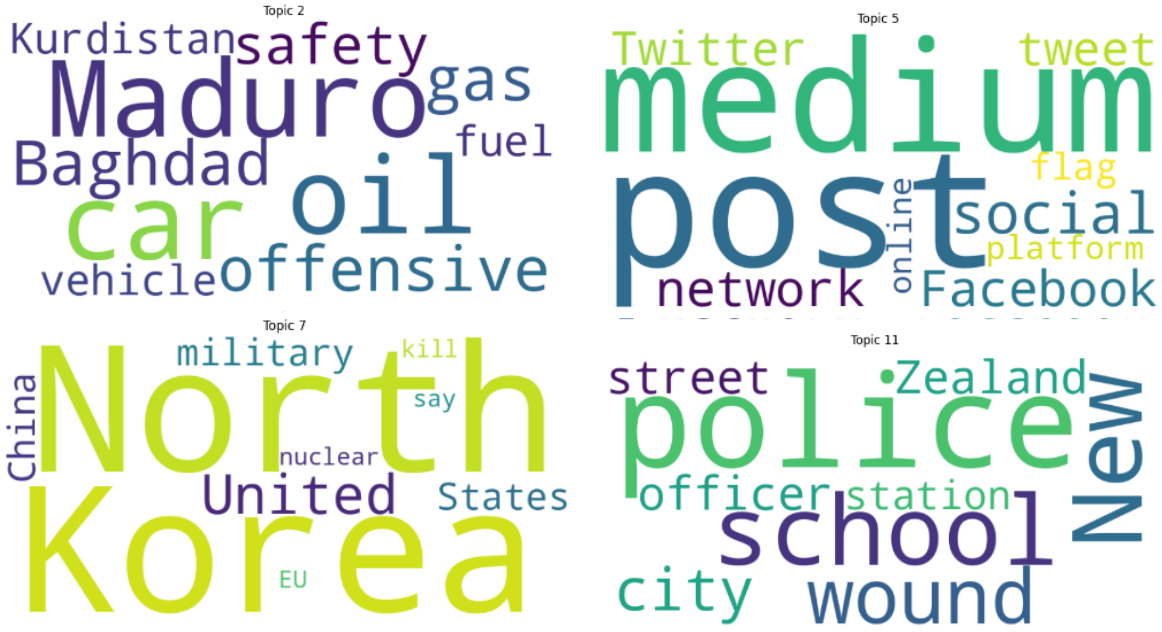

Figure 1 – Visual presentation of examples of sets of keywords for some automatically detected text categories in the dataset

The field of developed method application is thematic classification of text information, which is necessary, for example, when solving tasks of determining sentiment tone of texts in socially-oriented services, analyzing reviews in electronic commerce systems, detecting disinformation in news aggregators, etc. Improving the quality of thematic classification of textual information directly ensures increase in efficiency of solving these applied problems. This approach allows to efficiently organize text data large volumes, providing structured access to information. Thus, applying the proposed method to social media data containing many opinions, comments, and discussions can help in understanding the main topics discussed by users and identifying sentiments in the community.

The method is implemented by software and investigated on English-language dataset. From thematic modeling in dataset, the optimal topics number was determined to be 14, and the cross-validation test yielded result of 0.86, which is improvement of 0.15 compared to use of LDA in its pure form for classification. Future scientific research will be focused on increasing topic classification percentage under conditions of uneven distribution of data in classes and researching topic modeling algorithms to improve proposed method.

References:

1. O. Mazurets, O. Sobko, R. Vit, V. Pasternak, Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database, in: Modern Scientific Challenges are the Driving Force of the Development of Scientific Research, Proceedings of XXIV International Scientific and Practical Conference, 2024, pp. 91-96.