ANNOTATION

dissertation for the degree "Doctor of Philosophy" (Ph.D) in the specialty 6D070400 - Computer Science and Software Engineering

Kirill Yakunin

“Development of models and methods for collection, analysis and classification of media-publications based on natural language processing methods”

In the dissertation work, methods of classification of natural language texts were proposed for solving media monitoring problems. The proposed methods are based on the idea of using large volumes of heterogeneous data (Big Data) to extract knowledge from them about the hidden structures of the text corpus, which allow constructing effective vector representations of texts, as well as solving classification problems. At the same time, the proposed methods have low computational complexity and allow you to create text classifiers according to an arbitrary evaluation criterion with a minimum amount of manual labelling, or in some cases automatically (without labelling).

Purpose of work. The purpose of the work is the development of models and methods for automatic multicriteria assessment of textual information from media sources and social networks within the distributed information system.

The task of research

To achieve the goals of the work the following tasks need to be solved:

- to identify the main factors limiting the possibility of constructing effective (in terms of computational difficulty and the required expert labelling) classification models for texts;
- develop a method of effective vectorization of texts on the basis of unlabeled data (Big Data);
- to develop a method of self-learning of the classification model on the basis of objective explicit indicators for the recognition of hidden (implicit) parameters of documents;
- to develop a method of multi-criteria evaluation of documents and media sources;
- to develop a distributed information system for the collection, storage, processing and classification of textual information from the media and social networks and to verify the accuracy of the results of the work of the information system.

The relevance of research. Media monitoring tasks are very popular in the world market, primarily in the form of reputation management products. However, higher-level tasks related to decision-making based on data from the media space at the state level are currently practically not solved automatically.

Despite the significant development of the topic of text classification, modern SoTA (state of the art are the best at the moment) models require large amounts of markup for classification according to a given criterion, as well as high performance (hundreds and thousands of processors for training). Consequently, there is a need for more efficient models for the classification and analysis of textual information in terms of the required expert markup and computational performance.

The object of the research is the methods of classification of text documents**.**

The subject of the research is methods of multi-criteria assessment of text documents based on thematic modeling.

The purpose of the dissertation research is the development of models and methods for automatic multi-criteria assessment of text information from media sources and social networks within a distributed information system.

Research methods. The following research methods are used in the dissertation work: classification methods, methods of vectorization of text documents, methods of multicriteria analysis for decision-making, technologies for the design and development of information systems.

Scientific novelty

1. A method for vectorizing text documents using the BigARTM thematic model is proposed
2. A method for assessing thematic interbody imbalance is proposed for self-learning of the classification model
3. A method of multifactorial assessment of the social significance of a publication is proposed.
4. The proposed methodology for multi-criteria assessment of mass media MMA based on the Bayesian aggregation system, the analysis of hierarchies (AHP) and thematic modeling

The main results of the research are: development of a methodology for multi-criteria assessment of text documents and media sources and related methods for self-learning and supervised learning of relevant models; development of a distributed information system for collecting, storing, processing and classifying data from mass media and social networks, which is a software implementation of the proposed methods; a multifactorial methodology for assessing the social significance of publications in the media and social networks. The following practical conclusions and recommendations were obtained during the implementation of the dissertation work can be distinguished:

1 The analysis of the market of systems providing media monitoring services, as well as the analysis of the regulatory framework and technical features, was carried out. Weaknesses of existing solutions are revealed, recommendations are formed;

2 The question of the influence of open information sources on society has been investigated, the main directions of influence have been identified, a list of informative indicators has been formed, on the basis of which this influence can be assessed;

3 Existing approaches to document classification and text vectorization were investigated, problems and weaknesses of current solutions were identified, recommendations were formed;

4 An approach to vectorization of texts based on a thematic model has been developed;

5 A method for assessing inter-corpus thematic imbalance has been developed, which makes it possible to automatically or semi-automatically obtain topic weights in relation to a given attribute;

6 Developed a method for multi-criteria assessment of MMA media sources based on the Bayesian model of aggregation;

7 Developed a distributed information system based on Open-Source solutions, which allows collecting (scraping), storing, processing textual information, as well as building thematic models and classifiers with the ability to visualize the results;

8 Collected a corpus of more than 6 million publications from Kazakhstani and Russian sources, including both publication texts and metadata;

9 The proposed models and methods were validated. The main series of computational experiments was associated with the definition of so-called dangerous news (socially significant, resonant negative publications). A comparison was made with the deep learning BERT model, the quality metrics of the proposed models are comparable, with less computational complexity and less requirements for the volume of manual expert markup.

Scientific results made it possible to develop models, methods and software tools that allow classifying text documents according to a number of features (criteria) with a minimum amount of manual labelling, using the so-called high-level markup of topics or article metadata. The information system developed within the framework of the research is easily scalable (both functionally and in terms of performance) and allows you to store, process, aggregate and visualize large amounts of data texts.

Recommendations and baseline data for the specific use of the results

The results of scientific research, in particular, the developed information system can be used for a number of purposes:

1. Use by researchers and scientists. As shown in section 5.3.1 of the work, there is great potential for using the developed information system for a variety of humanities research.

2. Use by large companies and government agencies to support decision-making.

3. Use by large companies to solve the problem of reputation management.

4. Use by ordinary users for intelligence search of data of interest.

At the same time, the universality of the system should be noted - it can be applied to a wide variety of text data corpora, for example, to internal documents of organizations, scientific publications, personal correspondence, etc.

Provisions for Defense:

1. A methodology for multifactorial assessment of social significance has been developed on the basis of a multi-criteria methodology for assessing mass media MMA using thematic vectorization of text documents.

2. An architecture and software implementation of a distributed information system for collecting, processing, evaluating and visualizing textual data from mass media and social networks Media Analytics has been developed, including the proposed evaluation methodology, as well as its evaluation and verification.

Connection of the topic with the plans of research programs. The presented results were obtained during the implementation of the project IIVT KN MES RK (source of funding Science Committee MES RK): program-targeted funding (PCF) SC MES RK BR05236839 "Development of information technologies and systems to stimulate sustainable development of the individual as one of the foundations of the development of digital Kazakhstan" in 2018-2020 years;

Approbation of research results. The main provisions and results of the research were reported at: Procedia 9th International Young Scientist Conference (Scopus, 2020), International Conference on Digital Transformation and Global Society (Springer, 2019, 2020) and other conferences. The main results of the study were published in the journal Symmetry, which has an impact factor of 2.51 (Q1). The developed information system is implemented in the Ministry of Education and Science of the Republic of Kazakhstan (Appendix A of the Dissertation).

On the topic of the dissertation, 16 papers have been published, of which 6 articles were published in journals included in the Scopus and Thomson Reuters databases (2 articles - Q1, 2 article - Q2, 2 articles - Q3), one of them was published in publications recommended by the Control Committee in education and science MES RK, 3 articles in journals with a CiteScore (2 articles - CiteScore 17%, 1 article - CiteScore 69%) in Scopus without an assigned quartile, 2 articles in journals included in the RSCI (impact factor 0.482 and 1.385) , 6 articles have been published in collections of international scientific and practical conferences. 4 copyright certificates were issued for the results of the work.

The work consists of an introduction, five sections, a conclusion and a list of used literature. The total volume of the dissertation is 130 pages.