ANNOTATION

of the dissertation work of Ybytayeva Galiya Seitkalievna on the topic: "Development of an information and analytical system for monitoring illegal text information based on an ontological approach", submitted for the degree of Doctor of Philosophy (PhD) in the educational program 8D06103 – "Management information systems"

The modern information space, saturated with numerous Internet content, provides the user with extensive opportunities for obtaining information. However, along with the increase in the amount of available knowledge, the problem of identifying unreliable, malicious or illegal content that can harm both individual users and society as a whole is also increasing. The importance and relevance of this research work are due to the increased activity of criminal structures and the spread of false information on the Internet.

Currently, law enforcement and government agencies in many different countries tend to focus more on preventing crimes and terrorism before they are committed than on combating crime after it has been committed. In order to follow this crime prevention paradigm, it is necessary to analyze a huge amount of information, including text and voice information, use advanced data mining and text analysis technologies, as well as NLP tools and approaches.

One of the main problems of modern society related to the expansion of the potential of new information technologies is the emerging possibility of using the Internet as a tool and means of crime with destructive and antisocial goals. On the other hand, emerging new technologies for analyzing unstructured information contained in computer-mediated communication (CMC), such as Facebook, Twitter, Instagram, YouTube, make it possible to carry out preventive processing of text data and prevent potential crimes. However, due to the inability to manually find and track the contents of all sites that may indicate intent or preparation for a crime, it is necessary to automate the identification of illegal Internet information. Such systems of automatic or automated search and analysis of multilingual illegal content should determine with a high degree of probability whether any person is planning to commit or has already committed a criminal act.

All of the above defines **the main problem** that this study is aimed at solving, concerning the need to identify digital traces of potential terrorism, extremism and other illegal and violent actions in Internet content. At the same time, approaches based on statistical methods and machine learning methods do not give good results due to the lack of trained corpora and the blurriness of the subject area of the illegal Internet, which has weak "linguistic" markers. Therefore, **the second problem** considered in the framework of this study is the need to take into account the semantic (semantic) component of the text when determining its belonging to illegal content. This problem should be solved by using an ontological approach, which consists in adding semantic differentiating features to the applied machine learning methods. **The third problem** is the lack of an ontology in the subject area of "Illegal Internet content" for the Kazakh language, as well as the lack of such ontologies in

the public domain for Russian and English. In addition, the problem related to the previous one remains the lack of effective algorithms and software for automatic generation of ontologies based on text corpora.

Based on the above, it can be concluded that at the moment the need for effective methods, models and software tools for automatic search and analysis of multilingual illegal Internet content is especially **relevant**.

The purpose of the study is to develop an information model of the automatic identification system of illegal texts of the Kazakh and Russian languages in Internet networks.

Research objectives. In order to achieve the set goals of the study, the following issues are being solved:

1) Analysis of modern methods and models for monitoring illegal textual information based on an ontological approach.

2) Development of corpus crime related Internet texts in Kazakh, Russian, Ukrainian and English languages.

3) Creation of a multilingual (Kazakh, Russian and English languages) terminological thesaurus.

4) Creation of the ontology "Illegal Internet content".

5) Development of a method and tools for automatic semantic markup of specialized corpus of criminally colored texts.

6) Development of an information and analytical system for monitoring illegal textual information based on an ontological approach.

The information and analytical system includes the ontology "Illegal Internet content", specialized text corpora, software tools for automatic semantic markup of specialized corpora of criminally colored texts and integrated technology for analyzing and monitoring illegal content in social networks and other Internet sources.

The use of the developed system makes it possible to increase the efficiency of law enforcement and special government organizations, by increasing the likelihood of solving crimes and preventing illegal actions. The social effect of this study is to improve the legal and criminal situation and improve the quality of life of society as a whole.

The object of the study is the automatic search and analysis of illegal multilingual text information.

The subject of the study is models and methods, software tools for searching and analyzing illegal multilingual text information based on an ontological approach.

Research methods. Statistical probabilistic methods, machine learning methods, corpus linguistics methods, theory of intelligence, as well as methods of semantic and grammatical analysis of natural language texts and methods of expert assessments.

The scientific novelty of the research

– Corpus of criminally colored multilingual Internet texts and tools for automatic semantic markup of corpus of criminally colored texts based on an ontological approach have been developed.

– A multilingual terminological thesaurus, an ontology "Illegal Internet content" and an information model have been created.

– An information and analytical system for monitoring illegal textual information based on an ontological approach has been developed.

The work is important in the field of information security and combating illegal content, and also opens up prospects for further research in the field of text processing and semantic analysis. The economic and industrial interest from the implementation of the system lies in the possibility of using the results obtained for operational use by behavioral analysis specialists, law enforcement agencies and security services when performing or improving procedures for assessing possible illegal threats.

The theoretical significance of the results obtained lies in the adaptation of existing and the development of new models and methods for processing and analyzing multilingual textual information.

The practical value of the results obtained lies in the development of an information and analytical system based on the provisions submitted for protection.

The main provision to be defended.

1) Two corpora and a multilingual terminological thesaurus have been developed:

– Russian Russian, Ukrainian and English texts, comprising 3,147 texts in Ukrainian, 5,506 texts in Russian, 300 texts in English;

– Russian Russian parallel corpus, comprising 3,000 texts in Russian and 3,000 texts in Kazakh, including 2,000 texts containing Kazakh-Russian sentences aligned in meaning;

– A thesaurus containing more than 600 basic words (330 nouns, 107 adjectives and about 170 verbs) and more than 2500 synonyms of basic words.

2) The ontology "Illegal Internet content" has been created.

3) Methods and tools for automatic semantic markup of specialized corpus of criminally colored texts have been developed.

4) An information model and software tools have been developed for the automatic search and analysis of multilingual illegal Internet content based on an ontological approach.

The degree of reliability and approbation of the results. The main results of the dissertation were reported and discussed at international and foreign scientific conferences and scientific seminars:

1) Nina Khairova, Anastasiia Kolesnyk, Orken Mamyrbayev, Galiya Ybytayeva, Yuliia Lytvynenko. Automatic Multilingual Ontology Generation Based on Texts Focused on Criminal Topic // Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). – 2021. – Vol. 1. – P. 108-117.

2) Г.С. Ыбытаева, О.Ж. Мамырбаев, Н.Ф. Хайрова, Б.Ж. Жумажанов. Қазақ тіліндегі мәтіндерде коллокацияларды анықтаудың статистикалық әдістерін талдау // «Информатика және қолданбалы математика» VI Халықаралық ғылыми конференциясының материалдары. – Алматы, Қазақстан, 2021. – Б. 256-262.

3) Г.С. Ыбытаева, О.Ж. Мамырбаев, Н.Ф. Хайрова, Б.Ж. Жумажанов, К.Ж. Мухсина. Параллель корпусты әзірлеу мәселелері // «Информатика және қолданбалы математика» VII Халықаралық ғылыми конференциясының материалдары. – Алматы, Қазақстан, 2022. – Б. 175-182.

4) Ybytayeva, Galiya & Orken, Mamyrbayev & Khairova, Nina & Rizun, Nina & Berdali, Sanzharsultan & Kuralai, Mukhsina. (2023). Creating a Thesaurus "Crime-Related Web Content" Based on a Multilingual Corpus // Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023). – 2023. – Vol. 3396. – P. 77-87.

5) Ybytayeva, G.; Khairova, N.; Mamyrbayev, O.; Mukhsina, K. and Zhumazhanov, B. Experimental Verification of Collocation Detection Methods // Proceedings of the 5th Workshop for Young Scientists in Computer Science and Software Engineering - CS&SE@SW, SciTePress. – 2023. – P. 13-18. DOI: 10.5220/0012008900003561.

6) Мамырбаев О.Ж., Хайрова Н.Ф., Ыбытаева Г.С**.** «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасын құру // «Информатика және қолданбалы математика» VIII Халықаралық ғылыми-практикалық конференциясы. – Алматы, Қазақстан, 2023. – Б. 107-114.

Personal contribution of the researcher. The doctoral student independently completed and solved the tasks of the dissertation work. He has developed multilingual corpus of criminally colored Internet texts. He created the ontology "Illegal Internet content". He developed an information and analytical system for monitoring illegal textual information based on an ontological approach. Performed an experimental evaluation of the developed models and technology.

The connection of the thesis topic with the plans of research work. The dissertation work was carried out within the framework of the project on grant research of the Ministry of Internal Affairs of the Republic of Kazakhstan "Information model and software tools for automatic search and analysis of multilingual illegal web content based on an ontological approach" - AR09259309 (2021-2023).

Publication of the main results of the dissertation research.

On the topic of the dissertation, 2 author's certificates were obtained and 6 works were published, of which 3 articles were published in journals recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Internal Affairs of the Republic of Kazakhstan, 2 articles were published in publications indexed by the Scopus and Web of Science database, 1 monograph:

1. N. Khairova, O. Mamyrbayev, N. Rizun, M. Razno and **G. Ybytayeva**, "A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages," in IEEE Access, vol. 11, pp. 54093-54111, 2023, doi:

10.1109/ACCESS.2023.3281680. **(Scopus: Процентиль – 89, Q1; Web of science: IF – 0.89, Q2);**

2. Kartbayev, A., Mamyrbayev, O., Khairova, N., **Ybytayeva, G.,** Abilkaiyr, N., Mussayeva, D. Correction of Kazakh synthetic text using finite state automata (2021) Journal of Theoretical and Applied Information Technology, 99 (22), pp. 5559-5570. **(Scopus: Процентиль – 30, Q3**);

3. Картбаев, А., **Ыбытаева, Г.,** Мамырбаев, О., Мухсина, К., & Жумажанов, Б. (2022). Методы формального представления сущностей в криминальных новостях для автоматического построения онтологии преступлений. Известия НАН РК. Серия информатики, (3), 136-152. https://doi.org/10.32014_2518-1726_2022_343_3_136-152;

4. **Ыбытаева, Г**., Н.Ф. Хайрова, К.Ж. Мухсина, & Б.Ж. Жумажанов. (2022). Лингвистикалық онтологияны қолдану және қалыптастыру мәселелеріне шолу. ҚР ҰҒА жаңалықтары. Информатика сериясы, (1), 96-106. https://doi.org/10.32014/2022.2518-1726.121;

5. **Г.С. Ыбытаева,** О.Ж. Мамырбаев, Н.Ф. Хайрова, К.Ж. Мухсина, Б.Ж. Жұмажанов. Сарапшылар пікірлерінің келісім өлшемі ретінде Коэннің каппа коэффициентінің ерекшеліктері. ҚР ҰИА жаңалықтары. Ақпараттық технологиялар сериясы, № 3 (89), 2023, 139-151. https://doi.org/10.47533/2023.1606-146X.26;

6. Мамырбаев О., **Ыбытаева Г.,** Бердали С. Веб-приложение многоязычной базовой онтологии «Противоправный веб-контент». № 32055 от 26.01.2023 г.;

7. **Ыбытаева Г.,** Мамырбаев О. «Құқыққа қайшы интернет-контент» көптілді онтология қосымшасы. № 38766, 29.08.2023 ж.

8. O. Mamyrbayev, N. Khairova, W. Wójcik, **G. Ybytayeva**. Automatic identification of illegal texts in Internet. Almaty. Institute of Information and Computational Technologies. 2023. 151 p. (монография).

The structure and scope of the thesis. The dissertation research work consists of an introduction, 4 sections, a conclusion, a list of references from 172 titles and 3 appendices. The work is presented on 93 pages and contains 19 figures and 13 tables.

Brief description of the thesis

In the introduction, the general characteristics of the dissertation work are given, the relevance is determined, the purpose and objectives of the dissertation, the object and subject of research are formulated, the main position to be defended is determined, the scientific novelty, theoretical significance, practical value of the research performed are substantiated, the personal contribution of the researcher, the degree of reliability and approbation of the results, the connection of the topic of the dissertation with the plans of scientific research are given.research work, publication of the main results of the dissertation research, the structure and scope of the dissertation work.

The first chapter provides an analytical overview of existing problems in the field of automatic search and analysis of multilingual illegal Internet content. The

main problems in the field of semantic analysis and semantic markup of corpora are described. An overview of the problems of using and formatting linguistic ontologies is also carried out. The state and prospects of development of methods for extracting basic concepts from texts used in automatic ontology generation are considered. A comparative analysis of data-driven and knowledge-driven approaches to extracting events from unstructured texts is presented. An overview of the existing possibilities of using Text Mining methods and tools for the analysis and semantic markup of corpora is carried out.

The second chapter examines the main sources of filling the ontology of illegal content, which include the developed multilingual terminological thesaurus with criminal vocabulary and the created text corpus of criminally significant SMS information. The chapter also includes a description of the approach developed for the automated filling and expansion of the thesaurus, based on the existing model of extracting facts from unstructured texts. The development of a multilingual ontology "Illegal Internet content", visualization and limitations of its use are presented. The ontology is based on a synonymous dictionary of terms related to illegal actions and crime.

The third chapter discusses the development of a method and tools for automatic semantic markup of specialized corpus of criminally colored texts. It is proposed to use linguistic corpora to extract instances of the classes of the ontology "Illegal Internet content" and introduces a method for identifying linguistic and lexical markers of illegal content in corpus texts. A model for determining the role of arguments of events in the text is presented, based on a logical-linguistic model of extracting facts. The chapter describes methods for automatically generating an ontology that allows you to define events by trigger type, and extract event participants, event attributes, and participant roles.

The fourth chapter discusses an integrated technology for searching and analyzing illegal content on social networks and other Internet sources, including machine learning methods and an ontological approach. This chapter shows the use of Cohen's kappa metric to evaluate the accuracy of the results of automatic extraction of concepts from the corpus of criminally significant texts of the modern Internet. An experimental proof of the effectiveness of the developed technology is also described. The efficiency of the developed technology is compared with other approaches to event monitoring.

The conclusion reflects the scientific and practical results of the dissertation work on the development of an information and analytical system for monitoring illegal textual information based on an ontological approach.