ANNOTATION

Dissertation work of Sagynay Moldir on the topic "Development of deep learning models for the classification of destructive web- content in the Kazakh language", submitted for the degree of Doctor of Philosophy PhD according to the educational program "8D06301 – Information security systems"

The relevance of research. In the contemporary digital environment, the automatic detection of destructive web content has become an important area of research and practical application for ensuring information security and social stability. Cyberbullying, expressions of hatred, calls for aggression, and unlawful content disseminated online harm individuals and affect social relations. Developing intelligent systems capable of detecting and classifying such content in a timely manner is therefore a pressing task.

The relevance of this research is determined by three main factors: legislative requirements, technological limitations, and linguistic challenges.

From a legislative perspective, Article 174 of the Criminal Code of the Republic of Kazakhstan addresses the unlawful nature of acts involving incitement to social, national, ethnic, racial, or religious discord. In addition, the Law of the Republic of Kazakhstan "On Communications" contains provisions governing the dissemination of unlawful information through communication networks and the internet. Consequently, the timely detection of destructive web content is both a technical task and an essential condition for safeguarding national information security and maintaining law and order.

From a technological perspective, most research focuses on high-resource languages, including English and Russian. Ready-to-use classifiers, sufficiently large annotated corpora, and task-specific systems for detecting destructive web content in Kazakh are lacking. This gap necessitates the development of new methods and models for automated text analysis.

From a linguistic perspective, the task is complicated by the characteristics of Kazakh, an agglutinative language with a complex morphological system. Word meanings change through the addition of affixes to roots, while the meaning and tone of an utterance depend largely on context and semantic relationships. Furthermore, destructive content is often expressed implicitly through subtext, emotional connotations, and indirect linguistic constructions. Models for analysing web content in Kazakh must therefore account for morphological and semantic features alongside contextual information.

The relevance of this dissertation thus stems from the societal risks posed by destructive online content, the need for its timely detection in accordance with legal requirements, and the limited availability of technological solutions for Kazakh. These factors establish the scientific and practical significance of developing an improved model for detecting destructive web content in the Kazakh language.

The purpose of this study is to conduct a comprehensive investigation and develop deep learning models for the classification of destructive web content in the Kazakh language.

To achieve this goal, it is necessary to solve the following tasks:

1. Compile a dataset of destructive texts in Kazakh collected from social media platforms.
2. Investigate text embedding methods, classical machine learning methods, and deep learning methods for classifying destructive web content in Kazakh.
3. Develop a morphological feature module for Kazakh.
4. Develop ontology-based semantic features for classifying destructive web content in Kazakh.
5. Develop a hybrid architecture for classifying destructive web content according to its types and characteristics while accounting for the specific features of Kazakh.
6. Test the developed model in an experimental environment.

Research Object: Destructive textual content in Kazakh on social media and web resources.

Research Subject: Methods and models for integrating contextual, morphological, and ontology-based semantic features for the multiclass classification of destructive web content in Kazakh.

Research Methods: Machine learning and deep learning methods, text classification methods, natural language processing methods, social network analysis methods, and ontology-based classification methods.

Scientific novelty:

1. An annotated corpus of destructive web content in Kazakh, labelled across multiple classes, was compiled to support its classification.

2. Ontology-based semantic features identifying the type, severity, and target of destructive web content in Kazakh were developed for use in its classification.

3. A hybrid classification architecture integrating contextual embeddings, morphological features of Kazakh, and ontology-based semantic features was developed to classify destructive web content in Kazakh.

The following provisions are submitted for protection:

1. To address the task of multiclass classification of destructive web content in Kazakh, a corpus of destructive web content was developed and a method was proposed for integrating ontology-based semantic features characterizing the type, severity and target of the content with KazBERT embeddings and morphological features, and the application of the proposed method increased Accuracy from 88.015% to 90.492% and Macro F1-score from 88.086% to 90.567% compared with the baseline KazBERT model.

2. A hybrid architecture integrating KazBERT embeddings, morphological features, ontology-based semantic features, and a BiLSTM-based sequence analysis layer was developed for multiclass classification of destructive web content in Kazakh. Compared with the baseline KazBERT model, the developed architecture increased Accuracy from 88.015% to 91.045% and the Macro F1-score from 88.086% to 91.103%, achieving a 25.28% relative reduction in the misclassification rate calculated from Accuracy.

Theoretical significance of the study the theoretical significance of this research lies in advancing the theoretical and methodological foundations of destructive web content classification in the Kazakh language. During the study, a system of ontology-based semantic features describing the type, severity and target of destructive web content was developed, and its role in improving text classification was theoretically substantiated. Furthermore, the scientific foundations of a hybrid approach integrating the morphological characteristics of the Kazakh language with contextual semantic representations were established.

The proposed approach contributes to the advancement of research in Natural Language Processing (NLP), automated destructive content detection, and semantic text analysis for the Kazakh language. Moreover, the findings of this study provide a theoretical foundation for improving content analysis models for the Kazakh language and for developing intelligent text processing methods that can be adapted to other Turkic languages.

Practical value.

The practical value lies in the potential application of the model in systems for the automatic detection and classification of destructive web content in Kazakh. The findings can be used to monitor social media, web platforms, and information systems, detect unlawful or destructive content at an early stage, and improve the efficiency of content moderation. The proposed methods and the developed dataset provide a basis for further applied research in Kazakh-language NLP, information security, and artificial intelligence. Based on the research findings, State Registration Certificate No. 69171 was obtained on 27 March 2026 for the computer program entitled “Improving the effectiveness of BERT-based deep learning models for classifying destructive web content in Kazakh using Supervised Contrastive Learning (SupCon)”.

Personal contribution of the applicant.

All the results presented in this dissertation were obtained by the doctoral student. The main contributions include the development of a system of ontology-based semantic features describing

the type, severity and target of destructive web content for destructive web content classification in the Kazakh language; the development of a hybrid classification architecture integrating the morphological features of the Kazakh language with contextual KazBERT embeddings; the development of an enhanced deep learning model for destructive web content classification by combining KazBERT embeddings with morphological and ontology-based semantic features; and the implementation of software for evaluating the proposed model, together with the experimental results and their statistical analysis.

The doctoral student independently developed the proposed algorithms, implemented the software, trained and evaluated the models, optimized the hyperparameters, conducted the experimental studies, performed the comparative and statistical analyses of the obtained results, formulated the scientific conclusions, and prepared the dissertation manuscript. The scientific publications and the state registration certificate for the computer program were prepared in collaboration with the scientific supervisors.

The relationship of the thesis topic with the plans of scientific research. This research was carried out within the framework of the project entitled "Development of Models and Methods for Detecting Youth Extremism and Ensuring Youth Safety in the Modern Information Space," funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19576868).

Approbation of the Dissertation Results. The research findings of this dissertation were presented and discussed on several occasions at research seminars of the Department of Cybersecurity and Cryptology, Faculty of Information Technology and Artificial Intelligence, at al-Farabi Kazakh National University. The findings were also presented at international and national scientific conferences and were positively evaluated by the scientific community.

The Author's Contribution to Each Publication. The dissertation findings were reported in 10 scientific publications. One article was published in a Scopus-indexed publication, three articles in the proceedings of Scopus-indexed conferences, and two articles in journals recommended by the Committee for Control in the Sphere of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan. Four articles were published in international conference proceedings. One certificate of entry into the State Register of Rights to Objects Protected by Copyright in the Republic of Kazakhstan was obtained.

In journals recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

1. Bolatbek M.A., Sagynay M., Mussiraliyeva Sh.Zh. Application of machine learning methods to detect destructive web content in Kazakh. News of the National Academy of Sciences of the Republic of Kazakhstan. Physical and Mathematical Series. 2024; 4: 99–111. https://doi.org/10.32014/2024.2518-1726.310

2. Sagynay M., Bolatbek M. A morphologically and semantically enhanced model for classifying destructive web content in Kazakh. Bulletin of KazATC. 2026; 144(3): 151–166.

Scientific articles in journals indexed in Scopus

1. Bolatbek M., Sagynay M., Mussiraliyeva S., Yeltay Z. Detection of offensive content in the Kazakh language using machine learning and deep learning approaches. PeerJ Computer Science. 2025; 11: e3027. https://doi.org/10.7717/peerj-cs.3027 (indexed in Web of Science: Q2; indexed in Scopus; 2024 SJR: 0.719; 80th percentile).

Scientific articles in conference proceedings indexed in Scopus

1. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Yeltay Zhastay. Machine learning methods for identifying destructive texts in the Kazakh language. 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology (ADMIT 2024). https://doi.org/10.1145/3701100.3701121

2. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Kymbat Baisylbayeva, Zhastay Yeltay. Kazakh language dataset for hate speech detection on social media text. 2024 9th International Conference on Computational Intelligence and Applications (ICCIA 2024). https://doi.org/10.1109/ICCIA62557.2024.10719327

3. Moldir S., Bolatbek M., Mussiraliyeva S., Baisylbayeva K. Natural Language Processing and Machine Learning for Detecting Harmful Speech in Kazakh Social Media. In: 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI). March 2025. IEEE; pp. 347–351.

Articles in international conference proceedings

1. Sagynay M., Bolatbek M. Development of deep learning models for classifying destructive web content in Kazakh. Proceedings of the International Conference of Students and Young Scientists "Farabi Alemi".

2. Bolatbek M., Sagynay M., Mussiraliyeva Sh. The problem of detecting harmful web content on web resources and its relevance. Proceedings of the 9th International Scientific and Practical Conference "Computer Science and Applied Mathematics", dedicated to the 90th anniversary of Al-Farabi Kazakh National University.

3. Sagynay M., Bolatbek M. Development of machine learning models for detecting and classifying harmful web content in Kazakh. "Farabi Alemi". 2025; p. 265.

4. Bolatbek M., Mussiraliyeva S., Sagynay M. Machine Learning-Based Hate Speech Detection in the Kazakh Language // Research in Language. – 2025. – Vol. 23. – P. 311–319.

Certificates of entry into the State Register of Rights to Objects Protected by Copyright

1. Bolatbek M.A., Sagynay M. "Improving the effectiveness of BERT-based deep learning models for classifying destructive web content in Kazakh using Supervised Contrastive Learning (SupCon)". Certificate No. 69171, dated 27 March 2026.

Length and structure of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a list of references, and an appendix. Its total length is 98 pages of typescript, including 26 figures, 21 tables, 107 references, and one appendix.

The introduction establishes the relevance of the dissertation topic and defines the research aim, objectives, object, and subject. It describes the research methods, scientific novelty, and practical significance of the findings. The main propositions submitted for defence are formulated, and information is provided on the presentation and discussion of the findings and the resulting publications. The length and structure of the dissertation are also specified.

The first chapter of the dissertation examines the theoretical and methodological foundations of detecting destructive web content in Kazakh. It defines destructive web content, describes its main types and impact on information security, and analyses the linguistic characteristics of Kazakh-language web texts. A comparative review is presented of natural language processing approaches, traditional machine learning methods, and deep learning methods used for the automatic detection of destructive texts. The chapter identifies research challenges associated with the agglutinative structure of Kazakh, the context-dependent nature of destructive meaning, and the limited availability of annotated data. Based on this analysis, the research problem is formulated and the need to integrate contextual, morphological, and ontology-based semantic features is substantiated.

The second chapter of the dissertation develops the foundations of a method for classifying destructive web content in Kazakh. It describes the collection, filtering, preprocessing, and expert annotation of texts from Telegram, YouTube, and VKontakte. A corpus of 12,059 texts was compiled and assigned to four classes: Neutral, Bullying, Nazism, and Violent Extremism. Its statistical and linguistic characteristics were analysed. A method was proposed for constructing morphological features covering part of speech, tense, person, and mood. An ontological knowledge model was developed to describe destructive web content in terms of its type, severity, and target, and a procedure was established for extracting semantic features using a lexicon and a graph structure. The chapter also describes the methodological basis for integrating these features with the contextual representations generated by KazBERT.

The third chapter of the dissertation develops deep learning models and a hybrid architecture for classifying destructive web content in Kazakh. KazBERT is adopted as the baseline contextual model, and methods are described for augmenting it with a morphological feature module and an ontology graph module. A mathematical model is proposed for integrating contextual, morphological, and ontology-based semantic representations. The chapter considers a text classification configuration using the [CLS] representation and alternative configurations that process token sequences using LSTM, BiGRU, or BiLSTM. It describes the structure of the hybrid architecture comprising KazBERT, MFA, ontology-based semantic features, and BiLSTM, together with the feature processing procedure and the calculation of class probabilities.

The fourth chapter of the dissertation evaluates the effectiveness of the proposed models through experimental studies. It describes the division of the data into training, validation, and test sets, the experimental environment, and the evaluation metrics. The models were run with five different random seeds under a common evaluation protocol, and the mean results and standard deviations were calculated. The contributions of the morphological, ontological, and sequence processing components to classification performance were assessed, and misclassifications were analysed using confusion matrices. Differences between the predictions of the proposed model and the baseline KazBERT model were evaluated using McNemar's test. Among the models examined, the BiLSTM-based hybrid architecture achieved the best performance, with an accuracy of 91.045% and a macro F1-score of 91.103%. The potential practical application of the model was demonstrated through a web interface prototype.

The conclusion presents the main results obtained in the dissertation.