Abstract for the Dissertation of Aigerim Bastarbekkyzy Toktarova, submitted for the degree of Doctor of Philosophy (PhD) in the field of Information Systems (8D06115), titled «Hate Speech Detection Using Natural Language Processing and Machine Learning Techniques in Online User Content»

ABSTRACT

Relevance of the Research: In today's digital landscape, machine learning algorithms make it possible to automate and efficiently compile contextual databases of hate speech, hostility, and veiled threats on social media. Hate speech includes verbal or written expressions that are intended to insult or discriminate against individuals or groups, covering topics such as religion, ethnicity, gender, and some others.

With the growing number of internet users, the spread of harmful and offensive comments on social media is becoming an increasingly serious problem. Hate speech, including threats and derogatory statements, significantly affects psychological well-being, leading to increased risks for victims of cybercrime on social media and other online platforms.

Hate speech manifests itself in various ways, such as offensive remarks, spreading harmful rumors, posting derogatory images or videos, and exclusion from social media. The consequences of exposure to such content can be severe, leading to anxiety, depression, low self-esteem, and even suicidal thoughts. A 2022 Facebook online survey of students attending summer camps found that among social media users in Kazakhstan, 75% of youth had experienced some form of cyberbullying. Psychological consequences include insomnia, loss of appetite, neglect of self-care, and decreased academic or social performance. In extreme cases, constant exposure to online hate can lead to tragic consequences. In 2024, 165 teenagers in Kazakhstan committed suicide, and 378 minors attempted suicide due to offensive or humiliating comments online. Approximately 5% of teenagers experience online bullying two to three times a month, and 12% report at least one incident. The social and professional consequences of hate speech damage reputations and create difficulties not only in building future relationships but also in finding employment. Statistics show that more than a third of online bullying victims have received private messages containing discriminatory or offensive language, 24% have experienced humiliating messages, and 31% have experienced negative comments under their photos.

Educational activities in this area play a crucial role in reducing online risks. Webinars, tutorials, and awareness-raising activities for both educators and parents can help reduce the prevalence of hate speech and the harm it causes. Creating a culture of respect and responsibility in online spaces can help create a safer digital environment for all users.

As of 2024, the internet penetration rate in Kazakhstan is 92.3%, which is 18.9 million regular users out of a total population of 20.07 million people. Social networks

are used by 14 million users, or 71.5% of the population. Given the constant availability of online information and the anonymity provided by digital platforms, people are increasingly vulnerable to cyberbullying and attacks at any time. Psychologists emphasize that emotional abuse, including online hate speech, can have a deeper psychological impact than physical violence. Therefore, identifying and mitigating harmful online content is a top priority for information technology research. This dissertation addressed exactly this problem and its solution by implementing automated detection of such text and words using machine learning.

Research Goal. The primary goal of this research is to develop a deep neural network model capable of automatically detecting hate speech in textual data.

Research Objectives:

To achieve this goal, the following objectives were established:

1. Conduct a comprehensive analysis of machine learning algorithms for binary and multi-class classification of offensive comments.
2. Collect and preprocess Kazakh-language data for training machine learning (ML) and deep learning (DL) algorithms.
3. Examine deep learning architectures and explore different DL techniques, including:
    a) Convolutional Neural Networks (CNNs)
    b) Deep Neural Networks (DNNs)
    c) Recurrent Neural Networks (RNNs)
    d) Long Short-Term Memory Networks (LSTMs)
    e) Multilayer Perceptrons (MLPs)
    f) Artificial Intelligence-based deep learning models
4. Conduct experimental studies, compare models, and optimize hyperparameters to enhance classification performance.

Research Object: Social media platforms (VKontakte, Instagram, YouTube, Twitter) and news portals (Nur.kz, Tengrinews).

Research Topic: The application of machine learning and deep learning algorithms for detecting hate speech in text data.

Theoretical Significance: The study examines existing research on the detection of ambiguous linguistic expressions in textual data and evaluates natural language processing (NLP) tools for automated classification.

Practical Significance: The research contributes to the development and training of deep neural networks for hate speech detection. The findings support advancements in NLP applications for filtering and moderating harmful content on social media.

Main results of the dissertation work:

1. Creation, preprocessing, and manual classification of a Kazakh-language dataset for machine and deep learning tasks.

2. Development and training of a deep neural network incorporating an attention mechanism for binary classification of hate speech.
3. Comparative analysis of machine learning and deep learning algorithms for hate speech detection in Kazakh-language texts.

The main results of the dissertation:

1. Creation of the Kazakh language dataset: A thorough Kazakh language dataset was assembled, annotated, and pre-processed, establishing a crucial resource for the training and assessment of machine learning and deep learning algorithms. Manual classification and linguistic standardization were employed to guarantee uniformity and precision in data representation.

2.Development of a hybrid deep neural network model: A distinctive hybrid deep learning architecture was developed and executed, using BERT alongside an attention mechanism. The model was designed for the binary classification of hate speech in online text and shown competence in recognizing the contextual and semantic nuances of abusive language.

3. A thorough comparative evaluation of machine learning and deep learning algorithms, including CNN, RNN, LSTM, MLP, and BiLSTM, was performed in relation to the suggested hybrid model. The research shown enhanced contextual comprehension and adaptability of transformer-based models in managing intricate linguistic patterns in multilingual environments.

4. Effective preprocessing techniques for noisy online data: The research established resilient preprocessing frameworks adept at managing informal language, code-switching, emoticons, and special characters prevalent in social media interactions. This greatly enhanced the model's resilience in real-world application scenarios.

5. The suggested model's performance was assessed utilizing data from the Kazakh news websites Contur.kz and Serke.org. These platforms were selected for their unique emphasis on socio-political commentary and editorial material, frequently mirroring robust public sentiment and possibly contentious discussions. The assessment demonstrated the model's efficacy in precisely identifying hate speech in in test mode, emotionally charged news stories.

Doctoral Candidate's Personal Contribution: The doctoral candidate conducted an extensive literature review, analyzed patents relevant to the dissertation topic, selected appropriate research methods, and conducted theoretical and experimental studies.

Validation of Research Results: the main findings of the dissertation were presented in the following scientific publications and at international conferences**:**

Approbation of research results. The main results of the dissertation were reported at seminars and meetings of the Department of Computer Engineering, at the International Kazakh-Turkish University named after Khoja Ahmed Yasawi, at international conferences held in Astana, India:

1. Extended meeting of the Department of Computer Engineering, No. 6, 19.02.2025.

2. 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), Astana

3. The11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, India

Web of Science and Scopus Indexed Publications:

1. Toktarova, A., Abushakhma, A., Adylbekova, E., Manapova, A., Kaldarova, B., Atayev, Y., ... & Aidarkhanova, A. (2023). Offensive language identification in low resource languages using bidirectional long-short-term memory network. International Journal of Advanced Computer Science and Applications, 14(6). DOI: http://dx.doi.org/10.14569/IJACSA.2023.0140687

2. Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., ... & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5).

DOI: http://dx.doi.org/10.14569/IJACSA.2023.0140542

In Scientific Journals Recommended by the Committee on Science and Higher Education of the Republic of Kazakhstan:

1. B. Toktarova, B.S.Omarov, G.N. Kazbekova, S.AMamikov, F.E. Temirbekova Collecting hate speech database on social network in Kazakh language by using machine learning/ Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Series of Physics and Mathematics, 1991-346X, No. 1, pp. 191-203, 2023, DOI:https://doi.org/10.32014/2023.2518-1726.177

2. A.B. Toktarova, Zh.Zh. Azhibekova, D.R. Sultan, M.A. Kerimbekov “Collecting hate speech in Kazakh language in online content using machine learning”/ Bulletin of Abai KazNPU, series “Physics and Mathematics Sciences”, vol. 81, No. 1, 2023

3. A.B. Toktarova, B.S. Omarov, Zh.Zh. Azhibekova, S. A. Mamikov “The importance of artificial intelligence in identifying offensive words in online content” Bulletin of Toraigyrov University. ISSN 2710-3420 Energy series. No. 1, p. 311-322, 2023

4. A.B. Toktarova, B.S. Omarov, Zh.Zh. Azhibekova, Automated classification of offensive words using the “emotional” opinions of network users / Bulletin of the National Academy of Sciences of the Republic of Kazakhstan, No. 2 (88), DOI:https://doi.org/10.47533/2023.1606-146X.9

5. A.B. Toktarova, B.S. Omarov, B.A. Kaldarova, Using bilstm in identifying offensive words from low-resource languages Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Physics and Mathematics Series, 1991-346X, No. 3, pp. 174-189, 2024, DOI: 10.32014/2024.2518-1726.299

6. A.B. Toktarova, B.S.Omarov, G.I. Beissenova, R.B. Abdrakhmanov Analysis of hate speech words in online content by using data mining Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. Series: Physics and Mathematics, No. 2 (346), pp. 237-251, 2023, https://doi.org/10.32014/2023.2518-1726.196

7. A.B. Toktarova, B.S. Omarov, F.S. Temirbekova "Classification of Kazakh language obscene words and adaptation of machine learning methods to their detection" / Bulletin of KazNPU named after Abai, series "Physical and Mathematical Sciences", Vol. 82, No. 2 (2023)

8. A.Toktarova, Zh.Azhibekova, A.Aliyeva, N.Sarsenbieva "Bidirectional long short-term memory in hate speech detection problem on networks" / Bulletin of KazNPU named after Abai, series "Physical and Mathematical Sciences" Vol. 87, No. 3 (2024), DOI: 10.51889/29595894.2024.87.3.010

In Proceedings of International Scientific Conferences:

1. Toktarova, A., Sultan, D., & Azhibekova, Z. (2024, May). Review of Machine Learning Models in Cyberbullying Detection Problem. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 233-238). IEEE.

2. Sultan, D., Suliman, A., Toktarova, A., Omarov, B., Mamikov, S., & Beissenova, G. (2021, January). Cyberbullying detection and prevention: Data mining in social media. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 338-342). IEEE.

Certificate of state registration of rights to a copyright object. Database "Detection of hate speech using natural language processing and machine learning methods in online user content". Entry in the register under No. 54674 dated "14" February 2025.

Implementation certificate. The developed model was tested on the information sites contur.kz and serke.org in test mode on 31.01.2025.

Volume and structure of the dissertation.

The dissertation work consists of an introduction, four chapters, a conclusion. The work is completed in printed form on 121 pages, using computer capabilities for focusing attention in the form of illustrations, diagrams and tables. The list of references consists of titles.

The author expresses deep gratitude to the scientific supervisor, Doctor of Philosophy (PhD), Associate Professor of the International University of Information Technologies, Batyrkhan Sultanovich Omarov and foreign consultant, Professor of the Istanbul Technical University, Yeshref Adali (Turkey, Istanbul) for their invaluable work and consultations during the research.