DETECTION AND CLASSIFICATION OF IDEOLOGICAL TEXTS IN THE KAZAKH LANGUAGE USING MACHINE LEARNING AND TRANSFORMERS

MILANA BOLATBEK
Al-Farabi Kazakh National University
bolatbek.milana@gmail.com

SHYNAR MUSSIRALIYEVA
Al-Farabi Kazakh National University
mussiraliyevash@gmail.com

KYMBAT BAISYLBAYEVA
Al-Farabi Kazakh National University
baisylbaeva.k@gmail.com

Abstract
Modern information technologies enable the automatic analysis of textual data to detect extremist and propagandistic content. This paper examines deep learning methods and transformers models for the automatic classification of ideologically charged texts in the Kazakh language. A comparison was conducted between neural network models (CNN, BiLSTM, GRU, Hybrid CNN+BiLSTM) and modern transformers (DistilBERT). The performance evaluation of the models was based on accuracy, recall, precision, and F1-score metrics, as well as error analysis. Experimental results showed that hybrid CNN+BiLSTM demonstrated the highest accuracy (95.11%), outperforming other models. CNN, BiLSTM and GRU also achieved high results (92-93%), making them effective for this task. Among transformers, DistilBERT proved to be the most balanced (85.74%). This study demonstrates that hybrid neural network models (CNN+BiLSTM) are the most effective solution, while DistilBERT performs best among transformer models. The findings can be utilized for developing automatic monitoring and filtering systems for Kazakh-language texts, capable of efficiently identifying ideologically charged content.

Keywords: ideological text classification, deep learning, transformers, propaganda, radicalization, recruitment

1. Introduction

With the development of digital technologies and the widespread use of social networks, the volume of textual content containing ideologically charged materials, including propaganda, radicalization, and recruitment, has increased. The automatic detection and classification of such texts play a crucial role in ensuring information security, preventing the spread of extremist materials, and monitoring ideological trends.

Traditional methods for identifying ideological content require manual analysis by experts, making the process labour-intensive and inefficient when dealing with large volumes of data. The application of deep learning methods and transformer models significantly improves the accuracy of automatic text classification and accelerates the analysis process.

This study examines four categories of texts that are most relevant for analyzing ideologically charged content:

Neutral texts (0) – Materials that do not contain propagandistic or extremist rhetoric.

Propaganda (1) – Texts aimed at spreading ideological views and shaping public opinion in a specific direction.

Recruitment (2) – Materials containing elements of recruitment, encouraging individuals to join certain groups or ideologies.

Radicalization (3) – Extremist texts that promote radical beliefs and justify violent actions (Mussiraliyeva et al., 2024).

Main objectives of the study:

Develop a model for the automatic classification of texts into four categories: radicalization, propaganda, recruitment, and neutral texts.

Compare neural network models (CNN, BiLSTM, GRU, Hybrid CNN+BiLSTM) with the transformer-based model (DistilBERT).

Identify the most effective model for analyzing Kazakh-language texts.

Assess the applicability of modern transformers for detecting ideologically charged content.

This research is focused on developing and evaluating automated methods that effectively analyze textual data in the Kazakh language and identify potentially dangerous or propagandistic content.

2. Literature Review

(Awan, 2017) notes that ISIS actively uses social networks to spread its ideology and extensively promote radical ideas among users. The process of radicalization does not occur instantly but gradually unfolds through social groups, chats, and video content, which predisposes individuals to extremist views. ISIS propaganda portrays terrorism as an attractive path for young people, influencing their perception of it as heroism and justice. The primary platforms for recruitment are Twitter and Facebook, although authorities attempt to restrict their activities. Combating cyber extremism requires actively removing extremist content, creating counter-narratives that expose terrorist misinformation, and implementing educational programs

that foster critical thinking and resistance to manipulation among young people. One of the studies in this field is (Awan, 2017), which analyses the characteristics of ISIS propaganda through social networks.

(Saifudeen, 2014) explains that online radicalization occurs in stages, starting with scepticism towards official sources and, in some cases, culminating in the acceptance of extremist violence. The internet facilitates the spread of radical ideas due to anonymity, echo chamber effects, and social media algorithms. Users may either remain cognitive radicals who support extremist ideas or become active extremists engaged in propaganda and recruitment. Informational content plays a crucial role in radicalization, utilizing emotional visual elements, slogans, and videos to reinforce extremist narratives. Extremist groups skilfully manipulate information to create an illusion of social support for radical views. Countering cyber extremism requires the early detection of radical content, the creation of alternative narratives that refute extremist ideology, and the use of modern formats for information dissemination. This aspect is explored in (Saifudeen, 2014), which examines how the internet influences radicalization, the role of social networks in spreading extremist ideologies, and strategies to counteract them.

(Mukhamedzhanova et al., 2019) argue that cyber extremism poses a significant threat to young people, as the internet has become the primary tool for spreading radical ideologies. Social media algorithms contribute to the dissemination of radical ideas by reinforcing echo chambers. Extremist groups use digital technologies, misinformation, intimidation, and psychological pressure to manipulate individuals. A comprehensive approach is necessary to combat cyber extremism, including legal regulation, education, and technological solutions. Blocking radical content and monitoring online spaces are essential steps in ensuring security. Additionally, developing critical thinking skills among young people helps reduce the influence of manipulation and misinformation.

(Rashid, 2023) shows that artificial intelligence plays a crucial role in combating cyber extremism by enabling the automatic detection, blocking, and prediction of the spread of radical content. Natural language processing, deep learning, and influence network analysis help identify extremist rhetoric in text, video, and multimedia formats. However, AI systems face challenges such as false positives, extremist tactics to bypass algorithms, and privacy concerns. Machine learning algorithms can predict the radicalization process by analyzing changes in users' vocabulary and social media activity. AI is particularly useful in high-risk regions, such as Pakistan, where radical groups use propaganda and misinformation to recruit individuals. However, clear legal regulations are needed to balance national security with freedom of speech. AI is a powerful tool in combating cyber extremism while emphasizing the need to consider technical, ethical, and legal aspects.

(Tahat et al., 2024) note that social media algorithms play a crucial role in countering cyber extremism by automatically detecting, restricting, and removing extremist content. Machine learning and AI analyse texts, images, and videos to identify and block dangerous content in a timely manner. However, extremist groups adapt by using coded symbols, slang, and encrypted messages to bypass

algorithmic restrictions, necessitating continuous updates and improvements in AI models. The study also talks about the importance of integrating technical, legal and social measures in the fight against cyber extremism. Despite the effectiveness of algorithms, they sometimes mistakenly block scientific, political, or public discussions, which in turn raises concerns about the balance between security and freedom of speech. Also, the use of artificial intelligence requires legal and ethical regulation, as excessive censorship may violate the rights of users.

The article (Lahnala et al., 2025) presents a unified model for analyzing extremism, which shows that extremist groups, regardless of their ideology, have similar cognitive and behavioral characteristics. The authors describe how text data analysis helps identify key indicators of radicalization. The study shows that radicalization often begins before people officially join extremist communities, and changes in language use can reveal radical trends months in advance. It has been shown that machine learning algorithms and psycholinguistic methods help to effectively identify individuals who are at risk; however, the study is limited to data in English, which does not consider political, economic or cultural factors.

(Berjawi et al., 2023) discuss the mechanisms of spreading radical ideology in social networks and the methods that are used to identify them. The study says that a comprehensive approach is needed in the fight against online radicalization, which includes technical, educational, and psychological strategies. It is also said that this task of detecting online radicalization and extremist statements is difficult due to the lack of multilingual datasets. Most of the existing research focuses on Islamic extremism, while right-wing radical movements and other forms of radicalization remain poorly understood. The article discusses the main methods for detecting extremist content, including network analysis, NLP methods, and hybrid approaches. Deep neural networks (such as LSTM, BERT, and RoBERTa) show high accuracy in detecting extremist content, although their effectiveness is limited by the lack of linguistic and ideological diversity in available datasets. Network methods help identify radical groups but are limited in analyzing individual users.

In the article (Govers et al., 2023), the authors emphasize that it is necessary to create unified models that will be able to detect ERH by including multimodal data (text, images, video), multilingual contexts, and network interactions. The study also emphasizes that it is important to balance measures to combat extremism with the protection of freedom of speech. The authors say that online extremism is a growing threat that requires effective automated methods to detect and prevent the spread of radical content. According to their analyses, modern approaches such as machine learning, NLP and network methods have noticeable limitations. Machine learning classifiers (SVM, Random Forest, LSTM) can classify extremist texts with sufficient accuracy, but they also often give false positives. Deep neural networks, including CNN and LSTM architectures, allow for the analysis of multimodal content, however, they require large datasets and significant computing resources. Also, the multilingual nature and cultural specificity of extremist rhetoric can create additional problems for automatic detection.

In the article (Aldera et al., 2021), the authors emphasize that it is necessary to develop unified multilingual models for detecting extremist content. They say that future research should include text analysis, network interactions, and multimedia data in universal algorithms. The study also emphasizes that it is important to consider ethical aspects so that measures to combat extremism do not infringe on freedom of speech. This paper also examines the advantages and limitations of modern detection methods and examines the problems associated with multilingualism.

3. Materials and methods

3.1. Data

The dataset is a textual collection designed for the classification of ideologically charged content. It includes both original and pre-processed versions of messages (corrected and stemmed). The primary purpose of this dataset is to train machine learning models for the automatic detection of propaganda, radicalization, and recruitment in Kazakh-language texts.

The dataset was compiled using real-world texts, including articles from open sources, comments from social networks, and analytical publications. Expert evaluation was employed to ensure accurate labelling of the texts, minimizing classification errors.

The dataset contains four categories of texts:

- propaganda – propagandistic texts,
- recruitment – recruitment (calls for joining),
- radicalization – radicalization (extremist ideas),
- neutral – neutral texts.

3.2 Data Preprocessing

This study focuses on the classification of texts into four categories: radicalization, propaganda, recruitment, and neutral, utilizing Machine Learning and Transformer-based models. The implementation of these models required careful preparation of the dataset, including appropriate preprocessing and feature extraction procedures.

The dataset was compiled from publicly available sources, including social media platforms such as Telegram, VK, and YouTube. It consists of approximately 8,000 texts, divided into four distinct classes: radicalization (2,200 texts), propaganda (1,800 texts), recruitment (1,700 texts), and neutral (1,700 texts).

Before being processed by the models, the dataset underwent several preprocessing steps, including tokenization, stop-word removal, and punctuation cleaning. Additionally, text vectorization methods such as TF-IDF (Term Frequency-Inverse Document Frequency) were applied to enhance feature representation. Transformer-based tokenizers were also utilized to ensure compatibility with deep learning architectures.

By combining traditional machine learning techniques with transformer-based models, this study aims to improve the accuracy and efficiency of ideological content classification.

3.3. Classification Methods

In this study, two groups of models were used for the classification of ideologically charged texts in the Kazakh language:

1. Deep Neural Networks

CNN (Convolutional Neural Networks) – uses convolutional layers to extract local features from text, making it highly effective for short-text classification.

BiLSTM (Bidirectional Long Short-Term Memory) – a bidirectional recurrent neural network capable of capturing both preceding and succeeding context in a text.

GRU (Gated Recurrent Unit) – a simplified version of LSTM that reduces computational complexity while maintaining efficiency, enabling faster model training.

Hybrid CNN+BiLSTM – combines the strengths of CNN and BiLSTM, allowing for better recognition of structural text features while capturing a broader context.

2. Transformers

DistilBERT – a lightweight version of BERT, trained on the same data but optimized for reduced computational costs, making it more efficient for high-speed processing tasks.

Based on the experimental results, Hybrid CNN+BiLSTM demonstrated the best balance between accuracy and performance, achieving the highest classification scores. DistilBERT was the most effective among the transformer models but still performed slightly lower compared to deep neural network models.

To achieve effective model training and attain high classification accuracy, it is crucial to define clear criteria for text categorization. In the absence of precise annotation instructions, models may find it difficult to distinguish between closely related classes, which can lead to a higher rate of errors. To mitigate this, an Ideology Dataset Annotation guideline was created to organize the data labeling process and reduce subjectivity in classification.

This guideline helps standardize annotation procedures and improves the quality of machine learning by clearly delineating the boundaries between categories. It offers detailed descriptions of key characteristics associated with each type of extremist content, enabling algorithms to better differentiate between recruitment, propaganda, and radicalization. This clarity is especially important for ensuring accurate model performance, as overlapping rhetorical methods and stylistic elements in texts can make automatic classification more challenging.

The Ideology Dataset Annotation guideline is intended to classify extremist content into three main categories: recruitment, propaganda, and radicalization. It supports machine learning algorithms in detecting extremist materials, reducing false positives and improving understanding of the different threat levels.

The three principal categories of extremist content classification are Recruitment, Propaganda, and Radicalization. Below is an overview of each category along with examples.

Recruitment refers to the encouragement of individuals to become members of extremist organizations. This category covers both direct and indirect appeals to join extremist groups, offers of social, financial, or personal advantages, and depictions of the group as a defender of justice or security. It does not include explicit calls for violence or action.

Examples:

"Join us – together we will restore justice!" (Recruitment)
"Our group offers security and employment." (Recruitment)

Propaganda is defined as the spreading of extremist ideology without direct calls to action.

Propaganda seeks to promote hostility toward certain groups, spread misinformation, distort facts, and present radical solutions without explicitly urging individuals to act. It often creates a division between "us" and "them."

Examples:

"They have always oppressed us, we will not tolerate it anymore!" (Propaganda)
"This nation has never been trustworthy." (Propaganda)

Radicalization refers to open calls for violent actions.

Radicalization involves explicit encouragement of violence, calls for attacks against specific groups, justification of violent acts, and demands for armed resistance.

Examples:

"We can't wait any longer – we must act now!" (Radicalization)
"They are our enemies, we must eliminate them!" (Radicalization).

4. Results and Discussion

4.1. Comparative Analysis of Model Accuracy

To evaluate the effectiveness of the models, we used the metrics Accuracy, Precision, Recall, and F1-Score, which assess their ability to accurately classify texts. During the experiments, we tested neural network architectures (CNN, BiLSTM, GRU, Hybrid CNN+BiLSTM) and transformers (DistilBERT).

The best performance was achieved by the hybrid CNN+BiLSTM model, reaching an accuracy of 95.11% and demonstrating high Precision and Recall, making it the most effective model for automatic text classification. CNN and

BiLSTM also showed high results (93-94%), but slightly lagged behind the hybrid architecture (Table 1).

Among transformers, DistilBERT demonstrated balanced accuracy (85.74%), providing acceptable Precision and Recall values.

Table 1: Classification results

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.9357	0.9999	0.9476	0.9511
BiLSTM	0.9315	0.9996	0.9494	0.9494
GRU	0.928	0.9983	0.9442	0.9442
DistilBERT	0.8574	0.8573	0.8574	0.8568
Hybrid CNN+BiLSTM	0.9511	0.9999	0.9511	0.9494

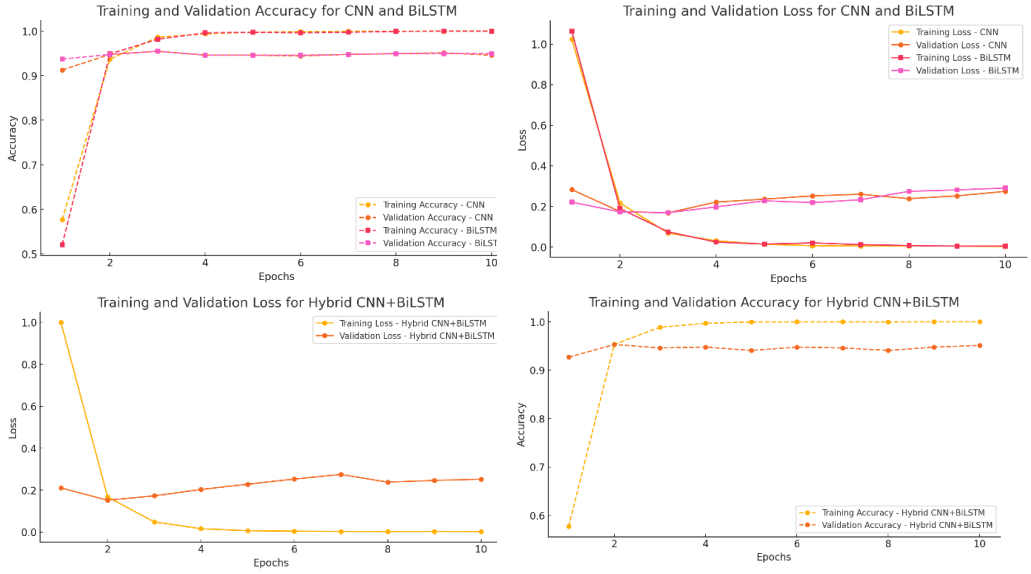
4.2. Training Graphs

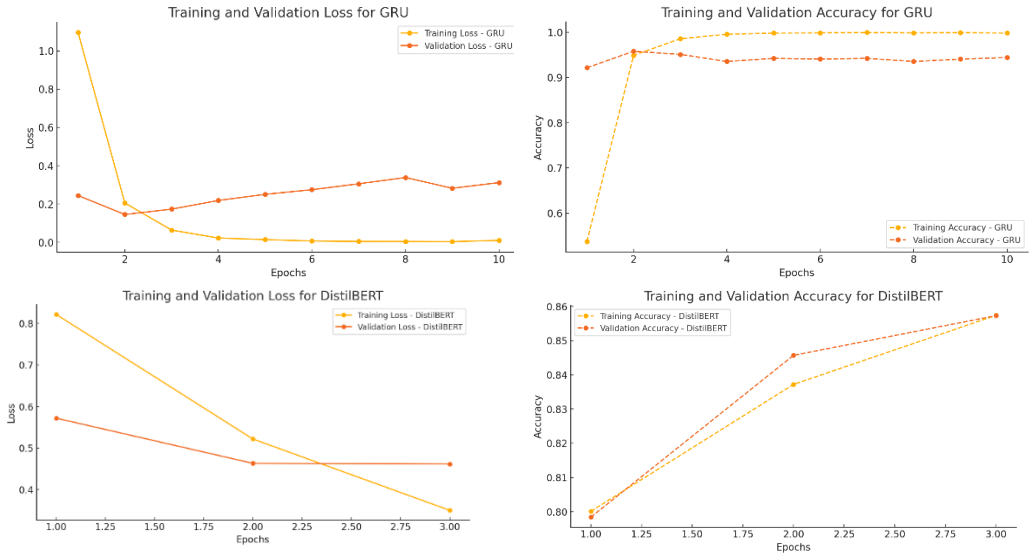
The training graphs illustrate the process of loss function reduction and accuracy improvement during model training. They provide a visual representation of the learning dynamics and help identify potential issues such as overfitting or underfitting.

During the experiments, it was observed that CNN, BiLSTM, and GRU exhibit stable learning, gradually reducing the loss function while increasing accuracy. The hybrid CNN+BiLSTM model trains most efficiently, as evidenced by its superior performance.

DistilBERT also demonstrates balanced learning, but its final accuracy is lower compared to neural network models.

Figure 1: Classification results



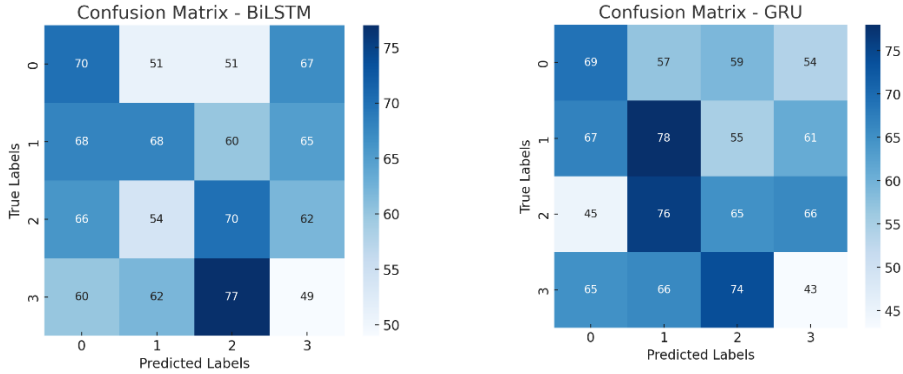


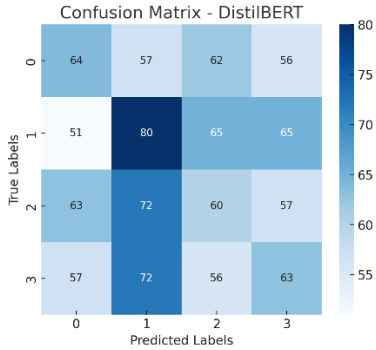
4.3. Error Analysis

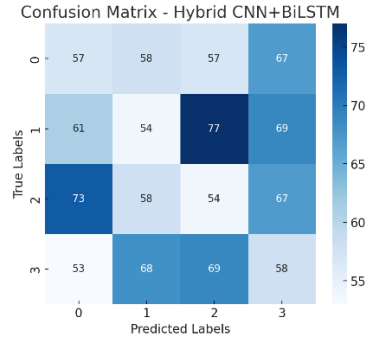
Error analysis was conducted using the confusion matrix, which helps identify the classes where models most frequently make mistakes.

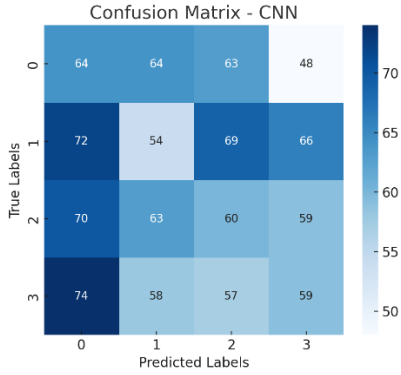
The most challenging categories for classification were "Propaganda" and "Recruitment", as they share similar lexical structures and rhetorical techniques. CNN, BiLSTM, and the hybrid model demonstrated the lowest number of errors. DistilBERT exhibited balanced errors but performed worse compared to CNN and BiLSTM (Figure 2).

Figure 2: Error analysis









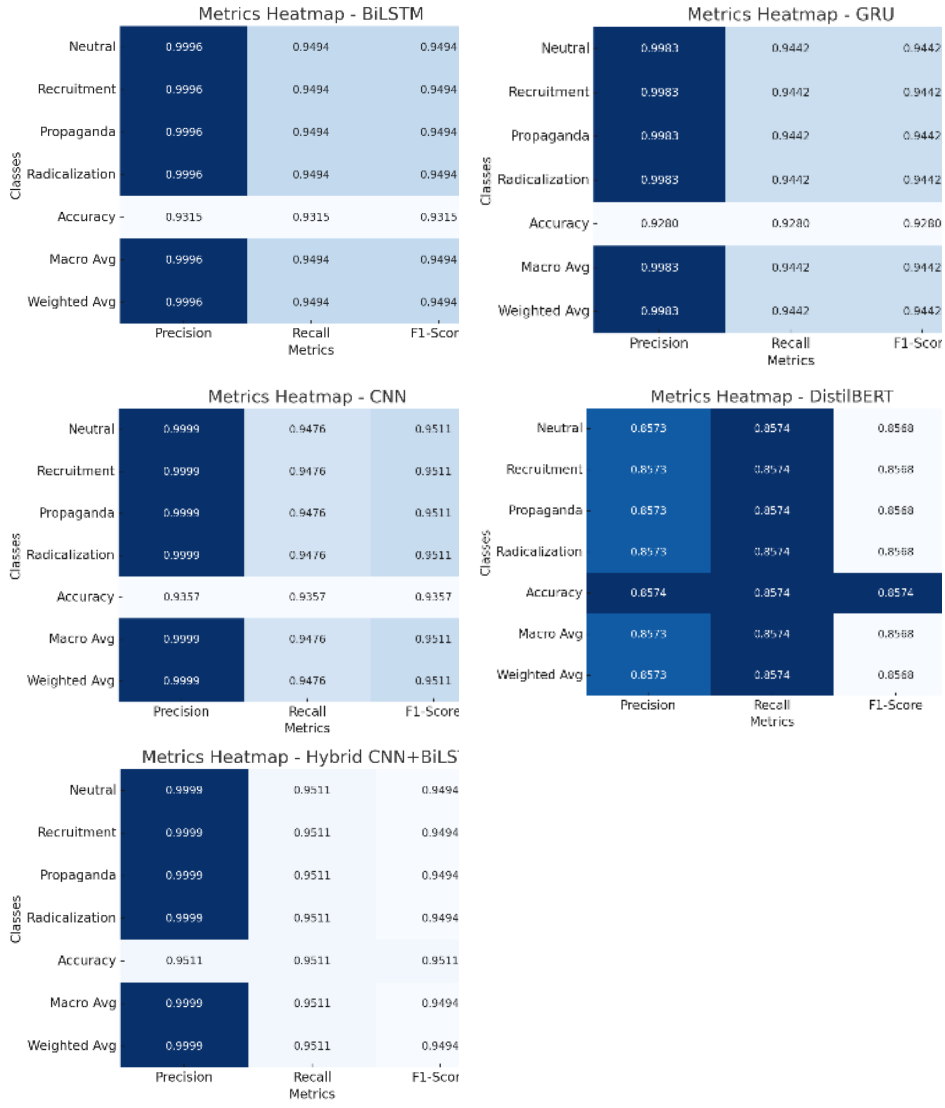
4.4. Comparison of Classification Metrics for Different Models

To visually compare the effectiveness of the models, heatmaps of metrics were constructed, displaying Precision, Recall, and F1-Score for each text category. These visualizations help identify which models perform better for specific classes.

The hybrid CNN+BiLSTM model achieved the best scores across all metrics, making it the optimal choice for this task.

DistilBERT occupies an intermediate position between neural network models, providing relatively high performance but still lagging behind CNN and BiLSTM (Figure 3).

Figure 3: Comparison of Classification Metrics for Different Models



Thus, the study confirmed the advantage of hybrid models over traditional neural networks and transformers, as well as the potential of DistilBERT for automatic analysis of Kazakh text.

5. Conclusion

This study conducted a comparative analysis of various machine learning and deep learning methods for the automatic classification of ideologically charged texts in the Kazakh language. The experiments tested traditional methods

(TF-IDF + Logistic Regression), neural network models (CNN, BiLSTM, GRU), and transformers (DistilBERT).

The experimental results demonstrated that the hybrid CNN+BiLSTM model achieved the highest accuracy (95.11%), outperforming both traditional methods and individual neural network architectures. CNN and BiLSTM also showed strong results, confirming their effectiveness in text classification tasks. Among transformers, DistilBERT proved to be the most balanced option, providing 85.74% accuracy while maintaining moderate computational costs.

Practical Significance. The developed classification system can be applied to:

- Automated monitoring of the information space, including media and social network analysis.
- Content filtering on online platforms, including comment and forum moderation.
- Text analysis in law enforcement agencies to detect threats and prevent the spread of extremist materials.
- Research tasks related to the study of information warfare and propaganda influence.

The study demonstrated that neural network models and transformers significantly outperform traditional machine learning algorithms in the classification of ideological texts. Hybrid CNN+BiLSTM proved to be the most effective approach, while DistilBERT emerged as the best transformer in terms of balancing accuracy and computational efficiency. Future work in this direction will focus on adapting transformers to the specifics of the Kazakh language, expanding the dataset, and developing ensemble methods to improve overall classification accuracy.

Funding statement

This research was carried out within the framework of the project funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.AP19676342).

References

Mussiraliyeva, S, Baisylbayeva, K., Bolatbek, M., Yeltay, Z., Decoding Ideology: Machine learning-based Detection of Extremist Content. 2024 International Conference on Intelligent Computing, Communication, Networking and Services, ICCNS 2024. doi 10.1109/ICCNS62192.2024.10776480

Imran Awan, Cyber-Extremism: Isis and the Power of Social Media, Social Science and Public Policy, Volume 54, pages 138–149, (2017), https://link.springer.com/article/10.1007/s12115-017-0114-0

Saifudeen, O. A. (2014). The Cyber Extremism Orbital Pathways Model. RSIS Working Paper No. 283, S. Rajaratnam School of International Studies, Nanyang Technological University, Singapore

Mukhamedzhanova, L. A., Kadirova, D. S., Agzamova, N. S., Tulaev, A. I., Rajabov, S. S., Alimov, S. K. (2019). Formation of Cyber Space, Protecting Youth From the Danger of Cyber Extremism. International Journal of Recent Technology and Engineering (IJRTE), 8(2 S4): 612-616. DOI:10.35940/ijrte.B1121.0782S419

Rashid, W. (2023). Using Artificial Intelligence to Combat Extremism. Pakistan Journal of Terrorism Research (PJTR), 5(2)

Tahat, K., Habes, M., Mansoori, A., Naqbi, N., Al Ketbi, N., Maysari, I., Tahat, D., & Altawil, A. (2024). Social media algorithms in countering cyber extremism: A systematic review. Journal of Infrastructure, Policy and Development, 8(8), 6632. https://doi.org/10.24294/jipd.v8i8.6632

Lahnala, A., Varadarajan, V., Flek, L., Schwartz, H. A., & Boyd, R. L. (2025). Unifying the Extremes: Developing a Unified Model for Detecting and Predicting Extremist Traits and Radicalization. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 19(1). https://doi.org/10.1609/icwsm.v19i1.35860

Berjawi, O., Fenza, G., & Loia, V. (2023). A Comprehensive Survey of Detection and Prevention Approaches for Online Radicalization: Identifying Gaps and Future Directions. IEEE Access, 11, 1-1. https://doi.org/10.1109/ACCESS.2023.3326995

Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. ACM Computing Surveys, 55(14s). https://doi.org/10.1145/3583067

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., & Alothaim, A. (2021). Online Extremism Detection in Textual Content: A Systematic Literature Review. IEEE Access, 9, 42384-42396. https://doi.org/10.1109/ACCESS.2021.3064178.

Milana Bolatbek is a researcher specializing in Artificial Intelligence and Natural Language Processing. She holds PhD degree in Information Security Systems and focuses on the development of intelligent systems for text analysis, hate speech detection, and digital content monitoring. Her academic interests include deep learning, computational linguistics, and social media analytics. Milana Bolatbek has contributed to several interdisciplinary projects integrating linguistics, psychology, and AI for cybersecurity applications. She has co-authored papers published in peer-reviewed and Scopus-indexed journals and actively participates in international conferences on artificial intelligence and data science.

Shynar Mussiraliyeva is a researcher in the field of Cyber Security and Data Analytics. She is a professor of the department of Cybersecuirty and Cryptology at al-Farabi Kazakh National University and has extensive experience in machine learning, natural language processing, and intelligent information systems. Her research focuses on applying AI technologies to solve problems in cybersecurity, social media analysis, and digital communication. Shynar Mussiraliyeva has published numerous papers in international peer-reviewed and Scopus-indexed journals. She is actively involved in academic collaborations and has supervised several research projects related to AI applications in language and behavior analysis.