Comparative Evaluation of Traditional and Transformer-Based Models for Risk-Level Classification of Uzbek Telegram Messages

Feruzakhon A. Qoyliyeva[1], Ozod J.Babomuradov[2], Akmal A. Savurbayev[3]
Tashkent State Agrarian University, Tashkent, Uzbekistan[1]
Head of Department for Coordination of Digitalization and Artificial Intelligence Development,
Administration of the President of the Republic of Uzbekistan, Jizzakh, Uzbekistan[2]
Department of Financial Technologies, Digitalization and Artificial Intelligence,
Presidential Administration of the Republic of Uzbekistan, Tashkent, Uzbekistan[3]

***Abstract*—The rapid growth of Telegram-based communication has increased the dissemination of harmful and risky content, particularly in low-resource languages such as Uzbek. This study investigates the automatic classification of Uzbek Telegram messages according to risk level using both traditional machine learning and transformer-based models. A dataset consisting of 10,000 real Telegram messages was collected and manually annotated into two classes: Safe and Dangerous. To improve data quality and consistency, preprocessing techniques including URL removal, emoji normalization, stop-word filtering, and script unification were applied. The study compares the performance of TF-IDF + Logistic Regression, FastText, mBERT, and XLM-RoBERTa for harmful content detection in Uzbek Telegram texts. Experimental results show that transformer-based models significantly outperform traditional approaches. Among all evaluated models, XLM-RoBERTa achieved the highest performance, with an Accuracy of 91.2%, a precision of 90.8%, a recall of 91.5%, and an F1-score of 91.1%, while mBERT achieved an Accuracy of 84.9% and an F1-score of 84.6%. The results demonstrate the effectiveness of contextual transformer architectures for identifying harmful content in low-resource language environments. The findings confirm that transformer-based models provide a reliable solution for automatic risk-level classification of Uzbek social media texts and can support practical applications in content moderation, information security, and social media monitoring systems.**

***Keywords*—Uzbek language; telegram messages; risk-level classification; harmful content detection; machine learning; transformer models; mBERT; XLM-RoBERTa**

I. Introduction

In recent years, the rapid growth of textual data generated through social media platforms has led to emerging challenges related to information security, public order, and digital culture. In particular, short, informal, and often unmoderated content shared on platforms such as Telegram and YouTube has made the timely detection of harmful content a critical issue [1–2].

Traditional rule-based approaches and classical machine learning methods such as Naive Bayes, Support Vector Machine, and Logistic Regression often fail to capture the deep semantic meaning of such complex textual data [3]. This limitation becomes even more pronounced for the Uzbek language due to the coexistence of Latin and Cyrillic scripts, widespread use of slang, abbreviations, and emojis [4].

The primary objective of this study is to perform a comparative analysis of traditional machine learning and transformer-based models for classifying Uzbek social media texts according to risk levels and to identify the most effective approach.

The exponential growth of textual content on social media has intensified the need for automated detection of hate speech, toxic expressions, user sentiment, and content safety. Recent studies indicate that deep learning techniques, particularly transformer-based architectures, significantly outperform traditional approaches in these tasks [5–6].

Several studies have demonstrated that multimodal approaches (text + image) can further improve hate speech detection performance. For instance, models based on RoBERTa and Swin Transformer have achieved high F1 scores in meme classification tasks, showing that combining textual and visual features yields better performance than single-modality approaches [7–8].

Transformer models such as BERT, RoBERTa, mBERT, and XLM-R have consistently demonstrated superior performance compared to classical machine learning methods (e.g., SVM, Logistic Regression, Random Forest), even in multilingual and imbalanced dataset scenarios. In particular, RoBERTa and XLM-R show stable results in low-resource language settings [9].

Research conducted for the Uzbek language indicates that monolingual pretrained models such as BERTbek can outperform multilingual models in certain NLP tasks. However, challenges such as script variation (Latin/Cyrillic) and limited corpus size may negatively affect model accuracy [10].

In the context of social media and user-generated content analysis, sentiment analysis and aspect-based classification tasks have also been widely studied. While classical machine learning models achieve moderate performance, transformer-based approaches have demonstrated significantly higher accuracy in sentiment detection and categorization tasks [11–

14]. Furthermore, incorporating non-standard elements such as emojis has been shown to improve model performance [15].

Overall, existing studies suggest that transformer-based models represent the most effective approach for analyzing social media text. However, challenges such as the scarcity of annotated datasets in Uzbek, the complexity of manual labeling, and the difficulty of capturing context-dependent expressions remain unresolved [12,16]. Therefore, this study aims to explore the effectiveness of modern transformer models for the automatic analysis of Uzbek social media texts.

Although previous studies have shown competitive results for BERTbek, this study focuses on comparing traditional machine learning approaches with multilingual transformer models. Evaluation of monolingual Uzbek transformer models is left for future work.

II. Theoretical Background and Proposed Approach

In this study, a real-world dataset consisting of Uzbek social media texts was collected from the Telegram platform. Telegram is characterized by rapid information dissemination, high user engagement, and diverse content types, making it a suitable source for studying both harmful and safe content.

The dataset was compiled from public Telegram channels, discussion groups, and comments under posts covering a variety of topics, including social, everyday, political, religious, and informational content. In total, 10,000 textual messages were collected and analyzed from an information security perspective.

Within the scope of this research, the texts were categorized into two classes based on their semantic content:

- 0 – Safe: Neutral or positive messages that do not pose a threat to public safety.
- 1 – Dangerous: Messages containing hate speech, violence, provocation, or content that may threaten societal safety.

To prepare the dataset for model training, a multi-stage preprocessing pipeline was applied. In the first stage, irrelevant elements such as URLs, HTML tags, excessive punctuation, and special characters were removed. This step helped eliminate noise and reduce non-informative features.

In the next stage, emojis and emoticons were normalized into standard tokens or corresponding textual representations to preserve emotional context within the text. Additionally, due to the mixed usage of Latin and Cyrillic scripts in Uzbek, all texts were converted into a unified format.

Subsequently, stop words (low-information auxiliary words) were removed to enhance the importance of semantically meaningful keywords. After that, tokenization was performed, and the processed texts were transformed into input representations compatible with transformer-based models.

The labeling process was conducted manually through expert annotation. Each message was carefully analyzed based on its semantic meaning and assigned an appropriate risk label. This approach ensured higher reliability compared to automatic labeling methods and improved the quality of model training.

As a result, a cleaned and annotated dataset was prepared for training and evaluating transformer-based models.

III. Method

In this study, various machine learning and deep learning models with different architectures were comparatively analyzed to evaluate their effectiveness in classifying Uzbek social media texts by risk level. The selected models range from traditional approaches to modern transformer-based architectures.

A. Traditional and Neural-Based Models

In the initial stage, a TF-IDF + Logistic Regression model was applied. Texts were transformed into vector representations using TF-IDF, followed by classification using Logistic Regression. This approach served as a baseline model for comparison with transformer-based methods.

Next, the FastText model was employed. FastText operates at the subword level, making it particularly effective for morphologically rich languages. This allowed the model to capture suffixes and various word forms in the Uzbek language.

Due to computational and time constraints, each model was evaluated using a single train-test split. Future work will include multiple runs with different random seeds to further assess model stability and reproducibility.

B. Transformer-Based Models

The main focus of this study was on transformer-based architectures. In particular, the mBERT model was utilized due to its training on large multilingual corpora, enabling contextual understanding of Uzbek text. mBERT evaluates each word within its surrounding context, achieving higher accuracy compared to traditional embedding methods.

Additionally, the XLM-RoBERTa (XLM-R) model was applied. This model is trained on massive multilingual datasets and is known for its strong performance in low-resource languages. XLM-R demonstrated superior ability in capturing the semantic meaning of Uzbek texts, including slang, abbreviations, and informal expressions, outperforming other models in this study.

C. Training Procedure and Hyperparameters

All transformer-based models were fine-tuned using pre-trained weights and adapted to the specific dataset used in this study. The fine-tuning process was conducted using the following hyperparameters:

- Learning rate: 2e-5
- Batch size: 16
- Number of epochs: 4
- Loss function: Cross-Entropy

These hyperparameters were selected based on experimental validation to ensure model stability and generalization capability.

D. Dataset Split and Evaluation Strategy

The dataset was randomly divided into 80% training and 20% testing subsets. This approach enabled an unbiased evaluation of model performance on unseen data.

All models were trained under identical conditions, and their performance was evaluated using standard classification metrics, including:

- Accuracy
- Precision
- Recall
- F1-score

This ensured a fair and consistent comparison of model effectiveness[17-19].

As a result, the comparative analysis clearly demonstrated the superiority of transformer-based models over traditional approaches and provided a solid methodological foundation for selecting the most effective model.

IV. Results and Discussion

This section presents the experimental results of the models applied for classifying Uzbek social media texts according to risk levels. All models were trained under the same conditions and evaluated on the test dataset. The results provide an objective comparison of overall performance and class-wise effectiveness.

A. Overall Model Performance

Table I summarizes the evaluation results of the applied models based on the key performance metrics: Accuracy, Precision, Recall, and F1-score.

TABLE I. Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
TF-IDF + Logistic Regression	71.4	70.2	69.8	69.9
FastText	78.6	77.9	78.1	78.0
mBERT	84.9	84.2	85.1	84.6
XLM-R	91.2	90.8	91.5	91.1

All Precision, Recall, and F1-score values are reported as weighted averages calculated on the test dataset.

B. Results Analysis and Discussion

As shown in Table I, the traditional TF-IDF + Logistic Regression model provides baseline performance; however, its ability to capture deep semantic relationships remains limited. Although the FastText model achieves improved results due to its subword-level representation, its overall performance is still inferior compared to transformer-based models.

The mBERT model demonstrates significantly better performance by leveraging multilingual contextual representations. Nevertheless, the highest performance across all evaluation metrics is achieved by the XLM-RoBERTa model, which consistently outperforms all other models.

C. Threshold Analysis

The effect of classification threshold selection was evaluated using the XLM-RoBERTa model. A threshold value of 0.7 was chosen to increase prediction reliability and reduce false-positive classifications. Under this setting, the model classifies a message as dangerous only when a higher confidence level is achieved[20-21].

Experiments conducted on the dataset of 10,000 messages demonstrated that the threshold of 0.7 improves classification stability and reliability. Although fewer messages are classified as dangerous, the precision of harmful content detection increases, making the system more suitable for practical applications such as content moderation and information security.

For threshold validation, a balanced subset of 200 messages (100 safe and 100 dangerous) was selected from the main dataset. This subset was used to evaluate the effect of threshold adjustment on model behavior and prediction confidence.

The results indicate that a higher threshold reduces false positives while maintaining strong classification performance. In particular, XLM-RoBERTa demonstrated stable performance in distinguishing between Safe and Dangerous messages. The high Recall observed for the Dangerous class suggests that the model effectively minimizes the risk of missing harmful content, which is essential for real-world monitoring and security systems.

Overall, the combination of a 10,000-message dataset and a threshold of 0.7 provides a reliable balance between precision and recall, enabling accurate identification of harmful content in Uzbek social media texts.

D. Error Analysis

The analysis of misclassified instances revealed that most errors occurred in texts containing sarcasm, irony, or implicit meanings. Additionally, very short or context-limited messages posed challenges for all models in determining the correct risk level.

Despite these challenges, the XLM-R model maintained relatively stable performance even in complex cases. This can be attributed to its training on large-scale multilingual corpora and its ability to capture deep contextual relationships.

E. Summary of Results

Overall, the experimental results clearly demonstrate that transformer-based models—particularly XLM-RoBERTa—achieve high effectiveness in classifying Uzbek social media texts by risk level. The model exhibits high accuracy, robustness, and generalization capability, making it a suitable solution for practical deployment [21-22].

F. Discussion

The findings of this study highlight the superior performance of transformer-based models in classifying Uzbek social media texts by risk level. The experimental results confirm that deep learning approaches significantly outperform traditional machine learning methods.

The baseline TF-IDF + Logistic Regression approach relies on surface-level textual features and fails to adequately capture contextual meaning. This limitation becomes particularly evident in emotionally rich or informal texts. The FastText model partially addresses this issue through subword representations; however, it still lacks the ability to fully capture long-range semantic dependencies.

The mBERT model demonstrates improved performance due to its multilingual contextual understanding. However, architectural limitations and the presence of cross-lingual noise slightly restrict its effectiveness.

The best performance is achieved by the XLM-RoBERTa model. This can be explained by its training on large-scale multilingual datasets and its ability to capture deep semantic relationships, even in low-resource languages. Furthermore, XLM-R effectively handles Uzbek-specific challenges such as slang, abbreviations, emojis, and mixed Latin-Cyrillic scripts.

Class-wise analysis indicates that high Recall in the Dangerous category is particularly important in practice, as it minimizes the risk of missing harmful content. Misclassifications between the Safe and Dangerous classes are mainly caused by ambiguous, context-dependent, or borderline messages that are difficult to interpret correctly. Despite these challenges, the transformer-based models demonstrated strong performance in distinguishing between the two classes.

At the same time, the study reveals several limitations. Even advanced transformer models struggle with sarcasm, irony, and implicit meanings in short texts. This suggests the need for future research focusing on deeper contextual modeling and multimodal approaches [23-25].

Overall, the results confirm that dataset expansion, improved preprocessing, and appropriate model selection play a crucial role in achieving high-performance text classification systems. The study successfully achieves its objective of developing an accurate and robust risk-level classification approach for Uzbek social media texts.

V. Conclusion

This study presents a comprehensive analysis of risk-level classification of Uzbek social media texts using various machine learning and deep learning models. The experimental results clearly demonstrate that transformer-based models—especially XLM-RoBERTa—significantly outperform traditional approaches.

The findings highlight that dataset expansion, optimized preprocessing, and proper model selection have a substantial impact on performance. The XLM-R model achieved high Accuracy, Precision, Recall, and F1-score, enabling more accurate semantic classification of informal, short, and emotionally rich Uzbek texts. In particular, the high Recall in detecting harmful content makes this model highly suitable for practical applications in information security systems.

The results indicate that even for low-resource languages such as Uzbek, it is possible to develop high-quality automated text analysis systems by applying appropriate methodologies and modern transformer architectures. This has important implications for social media monitoring, early warning systems, and ensuring information security in digital environments.

However, the study also identifies certain limitations. Detecting sarcasm, irony, and implicit meanings remains a challenging task. Additionally, very short and context-limited texts continue to pose difficulties for classification models.

Future research directions include:

- Developing real-time harmful content monitoring systems.
- Applying multimodal analysis (text + image + video) for risk detection.
- Expanding the dataset and improving class balance.
- Incorporating contextual and dialogue-based models for detecting sarcasm and implicit meanings.

The results of this study provide a strong foundation for further research in intelligent analysis of Uzbek social media content and open opportunities for real-world implementation.

References

[1] E. Kuriyozov et al., “BERTbek: A pretrained language model for Uzbek,” in Proc. SIGUL @ LREC-COLING, 2024.

[2] B. Mansurov and A. Mansurov, “UzBERT: Pretraining a BERT model for Uzbek,” arXiv preprint arXiv:2108.09814, 2021.

[3] E. Kuriyozov et al., “Text classification dataset and analysis for Uzbek language,” arXiv preprint arXiv:2302.14494, 2023.

[4] S. G. Matlatipov et al., “Aspect-based sentiment analysis for the Uzbek language (UzABSA dataset),” in Proc. SIGUL @ LREC-COLING, 2024.

[5] S. Allanazarova and D. Elova, “Uzbek texts sentiment analysis: Database development,” in Proc. ScitePress, 2023.

[6] K. Madatov et al., “TF-IDF-based classification of Uzbek educational texts,” Applied Sciences, vol. 15, no. 19, Art. no. 10808, 2025.

[7] X. Wang, “CLTL @ Multimodal hate speech event detection,” in Proc. CASE @ EACL, 2024.

[8] S. Thapa, “Extended multimodal hate speech event detection during Russia–Ukraine war,” in Proc. CASE, 2024.

[9] T. Mandl et al., “HASOC 2024: Hate speech identification in English and Bengali,” in Proc. CEUR Workshop, 2024.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.

[11] S. Matlatipov et al., “UzABSA: Aspect-based sentiment analysis for Uzbek,” in Proc. SIGUL, 2024.

[12] A. Taniberdiyev, “Ijtimoiy tarmoq matnli yozishmalarini tasniflash texnologiyalari,” CyberLeninka, 2024.

[13] Y. Yusufu et al., “A case study of the Uzbek language Google Play reviews,” Information Processing & Management, 2025.

[14] HuggingFace, “Uzum Market sentiment dataset,” 2024–2025.

[15] I. Rabbimov, I. Mporas, V. Simaki, and S. Kobilov, “Investigating the effect of emoji in opinion classification of Uzbek movie review comments,” 2020.

[16] S. Matlatipov, J. Rajabov, E. Kuriyozov, and G. Matlatipov, “Uzbek sentiment analysis based on local restaurant reviews,” 2022.

[17] O. J. Babomuradov and F. A. Qo‘yliyeva, “Analysis and comparison of methods for evaluating social media messages,” in Proc. Int. Sci. Pract. Conf. Scientific-Technical and Sociocultural Problems of Modern Society, Jizzakh, Uzbekistan, 2025, pp. 114–118.

[18] J. Sosa et al., “Multimodal pipeline for collection of misinformation data from Telegram,” in Proc. LREC, 2022.

[19] P. Kapil et al., “A survey on combating hate speech through detection and countermeasures,” in Proc. ICON (ACL), 2024.

[20] S. Al-Saqqa et al., “A survey of hate speech detection for Arabic social media,” Procedia Computer Science, vol. 238, 2024.

[21] J. Opitz et al., “A closer look at classification evaluation metrics and a guide for practitioners,” Trans. ACL, 2024.

[22] A. Abdullayev et al., “Dialect-sensitive sentiment analysis for Uzbek news,” in Proc. Conf., 2025.

[23] J. Rajabov, “Analysis and retraining of the BERT model for the Uzbek language: Methods and results,” in Proc. Int. Conf. Educational Discoveries and Humanities, 2024.

[24] UnitaryAI, “Detoxify: Trained models and code to predict toxic comments,” GitHub repository, 2020–2025.

[25] A. Banerjee et al., “A comprehensive survey on transformer-based machine translation,” ACM Computing Surveys, 2025.