INTERNATIONAL JOURNAL OF TECHNOLOGY AND ACADEMIC RESEARCH

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

DEVELOPING AN ARTIFICIAL INTELLIGENCE MODEL FOR DETECTING TOXIC (HARMFUL) CONTENT IN TEXTS

Toshpulatov Javoxir Xamidulla o‘g‘li

Tashkent International University of Education

Master’s student in Data Science

National Committee of the Republic of Uzbekistan for Ecology and Climate Change

Chief Specialist of the IT Department, “ECO-MEDIA CENTER”

toshpulatov100797@gmail.com

Doi: https://doi.org/10.5281/zenodo.23735945

Abstract. The article analyzes the theoretical and methodological foundations of building an artificial intelligence model for the automatic detection of toxic content in texts, namely insulting, degrading, threatening and hate-inciting statements. Taking into account the features of Uzbek as a low-resource language, the parallel use of the Latin and Cyrillic scripts and the phenomenon of code-switching with Russian, a conceptual model architecture for Uzbek-language texts is proposed.

Keywords: toxic content, artificial intelligence, natural language processing, machine learning, transformer, BERT, XLM-RoBERTa, low-resource languages, Uzbek language, content moderation.

Annotatsiya. Maqolada matnlardan toksik, ya’ni haqoratli, kamsituvchi, tahdidli va nafrat uyg‘otuvchi kontentni avtomatik aniqlash uchun sun’iy intellekt modelini yaratishning nazariy va metodologik asoslari tahlil qilingan. O‘zbek tilining kam resursli til sifatidagi xususiyatlari, lotin va kirill yozuvlarining parallel qo‘llanilishi hamda rus tili bilan aralash nutq hodisasi inobatga olingan holda o‘zbek tilidagi matnlar uchun konseptual model arxitekturasi taklif etilgan.

Kalit so‘zlar: toksik kontent, sun’iy intellekt, tabiiy tilni qayta ishlash, mashinali o‘qitish, transformer, BERT, XLM-RoBERTa, kam resursli tillar, o‘zbek tili, kontent moderatsiyasi.

Аннотация. В статье анализируются теоретические и методологические основы создания модели искусственного интеллекта для автоматического выявления токсичного контента в текстах, то есть оскорбительных, унижающих, угрожающих и разжигающих ненависть высказываний. С учётом особенностей узбекского языка как малоресурсного языка, параллельного использования латиницы и кириллицы, а также явления смешения с русским языком предложена концептуальная архитектура модели для текстов на узбекском языке.

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

Ключевые слова: токсичный контент, искусственный интеллект, обработка естественного языка, машинное обучение, трансформер, BERT, XLM-RoBERTa, малоресурсные языки, узбекский язык, модерация контента.

Introduction

Over the past decade, the widespread adoption of social networks, messengers, online media and forums has increased the speed of information exchange in society to an unprecedented degree. At the same time, the growing share of texts containing insults, discrimination, threats, hatred and incitement to violence in the digital environment poses a serious risk to the psychological safety of users, to social stability and to the reputation of the platforms themselves. Manual review of the millions of comments and messages created every day has become impossible both economically and in terms of time, which is why the development of intelligent systems capable of automatically detecting toxic content has turned into a pressing scientific and practical task. In the Republic of Uzbekistan, the introduction of artificial intelligence technologies into public administration and the economy has been declared a strategic priority, and the tasks of conducting research in this field, building national datasets and training specialists have been consolidated at the regulatory level [1]. The concept of toxic content is not interpreted uniformly in the scholarly literature. Some authors restrict it to hate speech, while others also include offensive, obscene, threatening and personally degrading expressions in this category. The study by T. Davidson and co-authors demonstrated that drawing the line between hate speech and merely offensive language is the most difficult problem of automatic classification, since lexically similar expressions acquire fundamentally different meanings in different contexts [2]. Furthermore, M. Sap and colleagues substantiated the risk of algorithmic bias, showing that toxicity detection models may wrongly label harmless texts typical of certain social groups or dialects as harmful [3].

Methodology and Literature Review

The methodological basis of the study consists of systematic literature review, comparative analysis, generalization and logical modeling. Davidson and co-authors, working with Twitter posts, applied algorithms such as logistic regression and support vector machines using word and character n-grams, TF-IDF weights and syntactic features [2]. Although this approach is easy to interpret and requires few computational resources, it ignores semantic relations between words and context, so that irony, veiled insults or vulgar words used in a positive context are misclassified. At the next stage, deep learning methods, in particular convolutional and recurrent neural networks together with distributed word vector representations, came into wide use. In their monograph, S.I. Nikolenko and co-authors explain in detail that such architectures make it possible to learn sequential and local patterns in text automatically, but that they require large volumes of labeled data [4]. The real leap in quality, however, came with the emergence of the transformer architecture based on the attention mechanism and of pre-trained language models. The BERT model proposed by J. Devlin and colleagues analyzes text in a bidirectional context and, after pre-training on a large corpus, can be fine-tuned for a target task on a relatively small labeled set [5]. The XLM-RoBERTa model developed by A. Conneau and co-authors was pre-

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

trained on more than one hundred languages and created the possibility of transferring knowledge from high-resource languages to low-resource ones [6].

An analysis of Russian-language research shows that the rich morphology and complex word formation of Russian create additional difficulties for detecting toxic content. S. Smetanin compared classical and neural network methods for detecting toxic comments in Russian and showed that multilingual and Russian-adapted transformer models have a considerable advantage over traditional approaches [7]. D. Dementieva and co-authors approached the problem from a different angle, proposing methods not only for detecting toxic texts but also for rewriting them into a neutral form while preserving their meaning, that is, detoxification, which defines the next direction in the evolution of moderation systems [8]. The review of sources on the Uzbek language indicates that this field is still taking its first steps. E. Kuriyozov and co-authors created labeled datasets for sentiment analysis in Uzbek and evaluated the effectiveness of classical and neural network models under low-resource conditions [9]. The UzBERT model developed by B. Mansurov and A. Mansurov is one of the first transformer models pre-trained specifically for Uzbek texts, and it outperformed multilingual models on Uzbek-specific tasks [10]. Nevertheless, the lack of open labeled corpora and dedicated models devoted directly to toxic content detection in Uzbek confirms the existence of a research gap.

Results and Discussion

The generalization of the reviewed literature made it possible to identify three main patterns in the development of toxic content detection models. First, there is a shift from methods relying on lexical features towards models that take context into account, since toxicity is often manifested not in an individual word but in the overall meaning of the utterance and the communicative situation [2; 5]. Second, instead of training a model from scratch for each language, the approach of fine-tuning multilingual pre-trained models for the target task prevails, and for low-resource languages in particular it is the only practical solution [6; 7]. Third, alongside accuracy, there is a growing demand for fairness, transparency and explainability of models, since a biased model may unjustifiably restrict the freedom of expression of members of certain social groups [3]. These patterns define the key principles that must be taken into account when designing a toxic content detection system for the Uzbek language.

According to the author's analysis, an effective model for Uzbek-language texts should be built on a conceptual architecture consisting of several sequential stages. The first stage is text preprocessing, in which problems specific to Uzbek must be resolved: transliteration of texts written in Latin and Cyrillic scripts into a single script, standardization of the various spellings of the apostrophe and of the letters o‘ and g‘, and normalization of abbreviations and deliberately distorted words that are widespread in social networks. At the second stage, the text is converted into a contextual vector representation by a pre-trained transformer model such as XLM-RoBERTa or UzBERT [6; 10]. Given that Russian words and phrases are frequently mixed into Uzbek online communication, a multilingual model appears preferable, as it also retains knowledge of Russian toxic vocabulary [7]. At the third stage, toxicity is classified not as a single binary label but in a multi-label manner across non-mutually-exclusive categories such as insult, threat, hate speech, obscene language and personal degradation. Such an approach shows the moderator not only that a text is harmful but also the type of harm, which facilitates decision-

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

making. As a fourth stage, a module for automatically rewriting detected toxic texts into a neutral form may be added, which makes it possible to preserve the user's opinion in a civil form instead of deleting it completely [8].

The implementation of the proposed architecture raises a number of issues that require discussion. The main obstacle is the absence of a high-quality labeled corpus of toxic content in Uzbek. To address this problem, it is recommended to draw on the experience of building sentiment analysis datasets, to develop unified annotation guidelines and to follow the principle of having each text assessed by several independent annotators [9]. In addition, regional dialects and the national cultural context affect the very notion of toxicity: an expression perceived as ordinary in one region may be regarded as an insult in another, which increases the risk of algorithmic bias [3]. Introducing mechanisms for visualizing attention weights and highlighting influential words in order to explain the model's decisions would ensure the transparency of the moderation process. Finally, the practical deployment of such systems must be balanced with personal data protection and guarantees of freedom of expression, and in significant cases the final decision should be made by a human [1].

Conclusion

The results of the literature review and the author's own analysis show that the field of detecting toxic content in texts is steadily evolving from methods based on lexicons and statistical features towards transformer models with a deep understanding of context. The results achieved in international and Russian research confirm that fine-tuning multilingual pre-trained models is the most promising path for low-resource languages. The Uzbek language, however, requires a specific approach because of its dual script, code-switching with Russian, the diversity of its dialects and the scarcity of labeled data. The conceptual architecture proposed in this article, comprising the stages of preprocessing, contextual vectorization, multi-label classification and detoxification, has been developed with these features in mind. Future research should focus on creating an open national corpus of toxic content in Uzbek, developing criteria for assessing model fairness, and improving the legal and ethical foundations for deploying such systems on public and private information platforms.

References

1. O‘zbekiston Respublikasi Prezidentining 2021-yil 17-fevraldagi «Sun'iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish chora-tadbirlari to‘g‘risida»gi PQ-4996-son qarori // Qonunchilik ma'lumotlari milliy bazasi. URL: https://lex.uz (accessed: 20.10.2026).

2. Davidson T., Warmsley D., Macy M., Weber I. Automated hate speech detection and the problem of offensive language // Proceedings of the International AAAI Conference on Web and Social Media. 2017. Vol. 11, No. 1. P. 512-515.

3. Sap M., Card D., Gabriel S., Choi Y., Smith N.A. The risk of racial bias in hate speech detection // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019. P. 1668-1678.

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

4. Николенко С.И., Кадурин А.А., Архангельская Е.О. Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2018. 480 с.

5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT 2019. Minneapolis, 2019. P. 4171-4186.

6. Conneau A., Khandelwal K., Goyal N. et al. Unsupervised cross-lingual representation learning at scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440-8451.

7. Smetanin S. Toxic comments detection in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2020». Moscow, 2020. Vol. 19. P. 1149-1159.

8. Dementieva D., Moskovskiy D., Logacheva V. et al. Methods for detoxification of texts for the Russian language // Multimodal Technologies and Interaction. 2021. Vol. 5, No. 9. Art. 54.

9. Kuriyozov E., Matlatipov S., Alonso M.A., Gómez-Rodríguez C. Construction and evaluation of sentiment datasets for low-resource languages: the case of Uzbek // Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science. Vol. 13212. Cham: Springer, 2022. P. 232-243.

10. Mansurov B., Mansurov A. UzBERT: pretraining a BERT model for Uzbek // arXiv preprint arXiv:2108.09814. 2021.