MUHANDISLIK & IQTISODIYOT

ijtimoiy-iqtisodiy, innovatsion texnik, fan va ta'limga oid ilmiy-amaliy jurnal

2026-YIL
SENTYABR / 9-SON, I-QISM

Milliy nashrlar

OAK: https://oak.uz/pages/4802

05.00.00 – Texnika fanlari
08.00.00 – Iqtisodiyot fanlar

Google Scholar

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

OpenAIRE

ISSN: 3060-463X

РЭУ.РФ
РОССИЙСКИЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Г.В. ПЛЕХАНОВА
ФИЛИАЛ В Г. ТАШКЕНТЕ

TOSHKENT DAVLAT TEXNIKA UNIVERSITETI

TOSHKENT TO'QIMACHILIK VA YENGIL SANOAT UNIVERSITETI

muhandislik & iqtisodiyot

ijtimoiy-iqtisodiy, innovatsion texnik,
fan va ta'limga oid ilmiy-amaliy jurnal

Elektron nashr, 2026-yil, sentyabr.

muhandislik & iqtisodiyot

ijtimoiy-iqtisodiy, innovatsion texnik,
fan va ta'limga oid ilmiy-amaliy jurnal

05.01.00 – Axborot texnologiyalari, boshqaruv va kompyuter grafikasi
05.01.01 – Muhandislik geometriyasi va kompyuter grafikasi. Audio va video texnologiyalari
05.01.02 – Tizimli tahlil, boshqaruv va axborotni qayta ishlash
05.01.03 – Informatikaning nazariy asoslari
05.01.04 – Hisoblash mashinalari, majmualari va kompyuter tarmoqlarining matematik va dasturiy ta'minoti
05.01.05 – Axborotlarni himoyalash usullari va tizimlari. Axborot xavfsizligi
05.01.06 – Hisoblash texnikasi va boshqaruv tizimlarining elementlari va qurilmalari
05.01.07 – Matematik modellashtirish
05.01.11 – Raqamli texnologiyalar va sun'iy intellekt
05.02.00 – Mashinasozlik va mashinashunoslik
05.02.08 – Yer usti majmualari va uchish apparatlari
05.03.02 – Metrologiya va metrologiya ta'minoti
05.04.01 – Telekommunikasiya va kompyuter tizimlari, telekommunikasiya tarmoqlari va qurilmalari. Axborotlarni taqsimlash
05.05.03 – Yorug'lik texnikasi. Maxsus yoritish texnologiyasi
05.05.05 – Issiqlik texnikasining nazariy asoslari
05.05.06 – Qayta tiklanadigan energiya turlari asosidagi energiya qurilmalari
05.06.01 – To'qimachilik va yengil sanoat ishlab chiqarishlari materialshunosligi
05.08.03 – Temir yo'l transportini ishlatish
05.08.06 – "G'ildirakli va gusenisali mashinalar va ularni ishlatish" (texnika fanlari)
05.09.01 – Qurilish konstruksiyalari, bino va inshootlar
05.09.04 – Suv ta'minoti. Kanalizatsiya. Suv havzalarini muhofazalovchi qurilish tizimlari
10.00.06 – Qiyosiy adabiyotshunoslik, chog'ishtirma tilshunoslik va tarjimashunoslik
10.00.04 – Yevropa, Amerika va Avstraliya xalqlari tili va adabiyoti
08.00.01 – Iqtisodiyot nazariyasi
08.00.02 – Makroiqtisodiyot
08.00.03 – Sanoat iqtisodiyoti
08.00.04 – Qishloq xo'jaligi iqtisodiyoti
08.00.05 – Xizmat ko'rsatish tarmoqlari iqtisodiyoti
08.00.06 – Ekonometrika va statistika
08.00.07 – Moliya, pul muomalasi va kredit
08.00.08 – Buxgalteriya hisobi, iqtisodiy tahlil va audit
08.00.09 – Jahon iqtisodiyoti
08.00.10 – Demografiya. Mehnat iqtisodiyoti
08.00.11 – Marketing
08.00.12 – Mintaqaviy iqtisodiyot
08.00.13 – Menejment
08.00.14 – Iqtisodiyotda axborot tizimlari va texnologiyalari
08.00.15 – Tadbirkorlik va kichik biznes iqtisodiyoti
08.00.16 – Raqamli iqtisodiyot va xalqaro raqamli integratsiya
08.00.17 – Turizm va mehmonxona faoliyati

Ma'lumot uchun, OAK Rayosatining 2024-yil 28-avgustdagi 360/5-son qarori bilan "Dissertatsiyalar asosiy ilmiy natijalarini chop etishga tavsiya etilgan milliy ilmiy nashrlar ro'yxati"ga texnika va iqtisodiyot fanlari bo'yicha "Muhandislik va iqtisodiyot" jurnali ro'yxatga kiritilgan.

Muassis: "Tadbirkor va ishbilarmon" MChJ

Hamkorlarimiz:

1. Toshkent shahridagi G.V.Plexanov nomidagi Rossiya iqtisodiyot universiteti
2. Toshkent davlat iqtisodiyot universiteti
3. Toshkent irrigatsiya va qishloq xo'jaligini mexanizatsiyalash muhandislari instituti" milliy tadqiqot universiteti
4. Islom Karimov nomidagi Toshkent davlat texnika universiteti
5. Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
6. Toshkent davlat transport universiteti
7. Toshkent arxitektura-qurilish universiteti
8. Toshkent kimyo-texnologiya universiteti
9. Jizzax politexnika instituti

MUNDARIJA • СОДЕРЖАНИЕ • CONTENTS

MUNDARIJA

MUNDARIJA • СОДЕРЖАНИЕ • CONTENTS

RESEARCHING METHODS FOR THE AUTOMATIC DETECTION OF OFFENSIVE LANGUAGE BASED ON ARTIFICIAL INTELLIGENCE

Fayzullaev Nurbek
Sambhram University, Jizzakh, Uzbekistan

Kuchkarov Muslimjon
Tashkent University of Information Technologies
named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
E-mail: nurbekfayzullayev89@gmail.com

Abstract. This study explores artificial intelligence-based methods for the automatic detection of offensive language. The research focuses on the application of Natural Language Processing (NLP), machine learning, and deep learning techniques to identify offensive, abusive, and inappropriate content in textual data. Particular attention is given to text preprocessing, feature extraction, text classification, and neural network-based approaches for improving the accuracy and efficiency of offensive language detection. The study highlights the potential of artificial intelligence to support safer digital communication and enhance automated content moderation across online platforms.

Keywords: Artificial intelligence, offensive language detection, natural language processing, machine learning, deep learning, neural networks, text classification, sentiment analysis, automated content moderation, abusive language, language models, social media.

Аннотация. В данной работе исследуются методы автоматического обнаружения оскорбительной лексики с использованием технологий искусственного интеллекта. Исследование направлено на применение методов обработки естественного языка (NLP), машинного обучения и глубокого обучения для выявления оскорбительного, агрессивного и неприемлемого контента в текстовых данных. Особое внимание уделяется предварительной обработке текста, извлечению признаков, классификации текстов и нейросетевым подходам к обнаружению оскорбительной лексики. Рассматривается потенциал технологий искусственного интеллекта в повышении точности и эффективности автоматической модерации контента, а также в формировании более безопасной и конструктивной цифровой коммуникации.

Ключевые слова: Искусственный интеллект, обнаружение оскорбительной лексики, обработка естественного языка, машинное обучение, глубокое обучение, нейронные сети, классификация текстов, анализ тональности, автоматизированная модерация контента, агрессивная лексика, языковые модели, социальные сети.

Annotatsiya. Ushbu tadqiqot sun’iy intellekt asosidagi texnologiyalar yordamida haqoratli va noo‘rin til birliklarini avtomatik aniqlash usullarini o‘rganishga bag‘ishlangan. Tadqiqot matnli ma’lumotlarda haqoratli, tajovuzkor va noo‘rin kontentni aniqlashda tabiiy tilni qayta ishlash (NLP), mashinaviy o‘qitish va chuqur o‘qitish usullarini qo‘llashga qaratilgan. Matnni oldindan qayta ishlash, xususiyatlarni ajratib olish, matnlarni tasniflash hamda haqoratli til birliklarini aniqlashda neyron tarmoqlarga asoslangan yondashuvlarga alohida e’tibor beriladi. Shuningdek, sun’iy intellekt texnologiyalarining aniqlash jarayonining aniqligi va samaradorligini oshirish, avtomatlashtirilgan kontent moderatsiyasini takomillashtirish hamda raqamli muhitda xavfsiz va konstruktiv muloqotni qo‘llab-quvvatlashdagi imkoniyatlari yoritiladi.

Kalit so‘zlar: Sun’iy intellekt, haqoratli til birliklarini aniqlash, tabiiy tilni qayta ishlash, mashinaviy o‘qitish, chuqur o‘qitish, neyron tarmoqlar, matnlarni tasniflash, hissiy holatni tahlil qilish, avtomatlashtirilgan kontent moderatsiyasi, tajovuzkor til birliklari, til modellari, ijtimoiy tarmoqlar.

INTRODUCTION

The rapid development of digital technologies and social media has significantly transformed the ways in which people communicate, exchange information, and express their opinions. Online platforms, including social networks, discussion forums, messaging applications, and content-sharing websites, generate

enormous amounts of user-generated textual data every day. These platforms create valuable opportunities for communication, knowledge sharing, collaboration, and public engagement. At the same time, the growing volume and diversity of online communication highlight the importance of developing effective approaches to identifying offensive, abusive, insulting, and harmful language. The timely detection of such content can contribute to improving the quality of online communication, supporting user well-being, and assisting platform administrators and content moderators in maintaining a safer and more constructive digital environment [1].

Traditional approaches to offensive language detection mainly rely on manual moderation and rule-based filtering systems. Although these approaches can be useful in specific situations, they may require considerable time and human resources when applied to large-scale and continuously evolving online content. Furthermore, offensive language can occur in various linguistic forms, including slang, abbreviations, intentional spelling variations, sarcasm, implicit insults, and context-dependent expressions. As a result, simple keyword-based approaches may have difficulties in accurately distinguishing offensive content from legitimate expressions, potentially leading to false-positive and false-negative classifications. These limitations emphasize the need for more flexible and context-aware computational methods.

Artificial intelligence (AI) provides promising opportunities to address these challenges through automated and scalable text analysis. In particular, Natural Language Processing (NLP), machine learning, and deep learning techniques can be applied to identify linguistic patterns associated with offensive language. Supervised learning algorithms can classify texts into predefined categories, while deep neural networks can learn complex semantic and contextual relationships from large-scale datasets. More recently, transformer-based language models have demonstrated strong capabilities in capturing the contextual meaning of words and sentences, making them particularly valuable for offensive language detection. The integration of these technologies can improve the accuracy, efficiency, and scalability of automated content moderation while supporting the development of more respectful, safe, and constructive digital communication environments.

LITERATURE REVIEW

Research on the automatic detection of offensive language has developed rapidly alongside advances in Natural Language Processing (NLP), machine learning, and deep learning. Existing studies have focused on developing classification models, creating annotated datasets, and enhancing the ability of artificial intelligence systems to recognize offensive content across different linguistic and social contexts.

Davidson et al. [2] investigated the classification of offensive language on social media and developed a labeled dataset containing tweets categorized as offensive, hate speech, or neither. Their study demonstrated that machine learning algorithms can effectively distinguish between different types of potentially harmful content. The authors also highlighted the importance of contextual and linguistic characteristics when differentiating offensive language from hate speech.

Waseem and Hovy [3] examined hate speech and offensive language on Twitter using supervised machine learning techniques. Their research emphasized the relevance of linguistic characteristics and demographic information in identifying abusive online communication. The findings demonstrated that automatic classification approaches can support large-scale analysis and monitoring of social media content, while also highlighting the influence of linguistic complexity and variability on classification performance.

Zampieri et al. [4] introduced the OLID (Offensive Language Identification Dataset), which was specifically designed for offensive language identification. The dataset provided a structured framework for distinguishing offensive from non-offensive content and identifying different types of offensive expressions. This work made an important contribution to the development of standardized evaluation tasks and enabled researchers to compare different computational approaches to offensive language detection.

Zampieri et al. [5] further developed the OffensEval shared task, which encouraged researchers to design and evaluate automatic systems for offensive language identification. The task incorporated multiple levels of classification, enabling researchers to determine whether a text was offensive and, where applicable, identify the target and type of offensive content. The results demonstrated the effectiveness of supervised machine learning and neural approaches while also emphasizing the importance of contextual interpretation in achieving reliable classification.

Fortuna and Nunes [6] reviewed computational approaches to hate speech detection and examined different definitions, datasets, features, and classification techniques. Their analysis showed that the absence of a universally accepted definition of offensive and hateful language can create challenges in dataset construction and in the comparison of research findings. The authors emphasized the importance of contextual information, annotation quality, linguistic diversity, and appropriate evaluation metrics in developing reliable and effective detection systems.

Fortuna et al. [7] investigated the application of deep learning methods to the detection of offensive and abusive language. Their research demonstrated that neural architectures can learn complex textual representations and reduce reliance on manually designed linguistic features. Deep learning approaches, particularly those capable of capturing contextual relationships, showed considerable potential for improving the accuracy and efficiency of automated text classification. At the same time, the study indicated that model performance is influenced by dataset size, data quality, language characteristics, and the representativeness of training examples.

Overall, the reviewed studies demonstrate a gradual transition from traditional machine learning and manually engineered linguistic features toward more advanced neural and contextual approaches. The development of high-quality annotated datasets and standardized evaluation frameworks has created favorable conditions for further progress in offensive language detection. In particular, deep learning and transformer-based models provide new opportunities to improve contextual understanding, classification accuracy, and the scalability of automated content moderation systems.

RESEARCH METHODOLOGY

This study employs a quantitative experimental design to evaluate AI-based offensive language detection on a balanced dataset of 5,000 online comments (80% training, 20% testing). The methodology consists of text preprocessing (cleaning, lowercasing, tokenization), TF-IDF feature extraction, and classification using machine learning algorithms and deep neural networks. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics.

ANALYSIS AND RESULTS

The study employed a quantitative experimental approach to investigate the effectiveness of artificial intelligence methods for the automatic detection of offensive language. A textual dataset consisting of 5,000 online comments was used for the analysis. The dataset was divided into two categories: offensive and non-offensive content. The research process included several sequential stages: data preprocessing, feature extraction, model training, text classification, and performance evaluation.

The data preprocessing stage included the removal of unnecessary symbols, text normalization, tokenization, and transformation of textual data into numerical representations suitable for computational analysis. Machine learning and deep learning approaches were applied to classify the textual data and identify patterns associated with offensive language. The effectiveness of the models was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score.

The experimental framework provided a systematic basis for comparing the performance of artificial intelligence approaches in offensive language detection. The use of multiple evaluation metrics enabled a more comprehensive assessment of classification performance and provided an objective basis for determining the strengths and practical applicability of the developed detection approaches (Table 1).

Table 1
Dataset characteristics

Indicator	Value
Total texts	5,000
Offensive texts	2,500
Non-offensive texts	2,500
Training data	4,000 (80%)
Test data	1,000 (20%)

The table summarizes the main characteristics of the dataset used in the study. The dataset consists of 5,000 texts, equally distributed between two categories: 2,500 offensive texts and 2,500 non-offensive texts. This balanced class distribution provides a reliable foundation for model development, as it minimizes potential class imbalance and enables the model to learn the distinguishing characteristics of both categories more effectively. It also supports a more objective and consistent evaluation of the model's ability to differentiate between offensive and non-offensive language.

For model development and evaluation, the dataset was divided into two subsets using a standard 80:20 training-test split. A total of 4,000 texts (80%) were allocated to the training set, allowing the machine learning or

deep learning model to identify relevant patterns and linguistic characteristics associated with both categories. The remaining 1,000 texts (20%) were reserved for testing. These texts were not used during the training process and were subsequently employed to assess the model's performance and generalization capability on previously unseen data.

Overall, the dataset demonstrates a well-balanced class distribution and follows a widely adopted 80:20 training-test allocation. This structure provides a solid and methodologically appropriate basis for developing, training, and objectively evaluating an offensive language detection model. The balanced composition and clear separation of training and test data contribute to the reliability, consistency, and generalizability of the study's results (Table 2).

Table 2
Research methods

Method	Purpose
Text preprocessing	Cleaning and normalization
TF-IDF	Feature extraction
Machine learning	Text classification
Neural network	Offensive language detection
Accuracy, Precision, Recall, F1	Model evaluation

The table presents the main methods employed in the study and describes the specific purpose of each method within the offensive language detection process. Together, these methods form a structured methodological framework covering the key stages of text preparation, feature extraction, classification, and model evaluation.

Text preprocessing represents the initial stage of the methodology and focuses on cleaning, standardizing, and preparing the textual data for further analysis. During this stage, unnecessary characters, punctuation, duplicate spaces, and other irrelevant elements are removed where appropriate. Text normalization may also involve converting words into a consistent format, such as transforming all letters to lowercase. This preprocessing stage contributes to data quality and consistency, thereby creating a reliable foundation for subsequent feature extraction and classification.

The second method, TF-IDF (Term Frequency–Inverse Document Frequency), is applied for feature extraction. TF-IDF transforms textual information into numerical representations that can be effectively processed by machine learning algorithms. The method assigns greater importance to terms that occur frequently within a particular text but are relatively less frequent across the dataset. As a result, TF-IDF enables the identification of informative linguistic features and supports the model in distinguishing patterns associated with offensive and non-offensive language.

Machine learning methods are subsequently applied to perform text classification. Using the extracted TF-IDF features, classification algorithms learn the distinguishing characteristics of the two text categories. In this study, the primary objective of the classification stage is to determine whether an input text contains offensive language. This approach provides an efficient and systematic mechanism for automatically categorizing textual content.

A neural network is also employed as an advanced approach to offensive language detection. Compared with conventional machine learning methods, neural networks have the ability to learn more complex relationships and patterns from textual data. This capability can support the identification of contextual and semantic characteristics of offensive expressions, including patterns that may be less effectively represented through conventional word-based features. Consequently, the neural network approach provides an additional perspective for evaluating the effectiveness of AI-based text classification.

Finally, the developed models are evaluated using Accuracy, Precision, Recall, and F1-score. These metrics provide a comprehensive assessment of classification performance from different perspectives. Accuracy represents the overall proportion of correctly classified texts, while Precision indicates the proportion of texts identified as offensive that are correctly classified as such. Recall measures the model's ability to identify relevant offensive texts, and the F1-score provides a balanced evaluation by combining Precision and Recall into a single measure.

The experimental results were compared to identify which AI-based approach provides more effective and reliable automatic identification of offensive language. The proposed methodological framework enables the performance of different approaches to be assessed systematically and provides a clear basis for evaluating the potential of artificial intelligence in automated content moderation. Overall, the combination of text

preprocessing, TF-IDF feature extraction, machine learning, neural networks, and comprehensive evaluation metrics establishes a structured and effective approach for offensive language detection.

The results of the study demonstrate that artificial intelligence can be effectively applied to the automatic detection of offensive language in online textual content. The findings indicate that effective preprocessing and appropriate text representation play an important role in supporting classification performance. The removal of unnecessary symbols, text normalization, and sentence tokenization contributed to reducing data noise and improving the consistency and quality of the input data.

The comparison of machine learning and deep learning approaches further indicates that AI-based models are capable of identifying linguistic patterns associated with offensive expressions. Traditional machine learning methods can achieve reliable classification results when the dataset is properly prepared and informative features are extracted. At the same time, deep learning approaches offer strong potential for capturing more complex linguistic patterns, contextual dependencies, and semantic relationships within textual data. This highlights the complementary value of both approaches in developing effective offensive language detection systems.

An important finding of the study is that offensive language is strongly influenced by context. The meaning and intention of the same word can vary depending on the sentence in which it is used. In addition, linguistic phenomena such as sarcasm, slang, abbreviations, intentional spelling variations, and implicit expressions may present additional challenges for automated classification. These characteristics emphasize the importance of developing context-aware models capable of interpreting language beyond individual keywords.

Overall, the findings suggest that combining effective text preprocessing, informative feature representation, and advanced AI-based classification approaches can provide a strong foundation for automated offensive language detection. The results also highlight opportunities for further improving model performance through enhanced contextual analysis, larger and more diverse datasets, and more advanced deep learning techniques. Such improvements can contribute to the development of more accurate, adaptable, and reliable AI-based content moderation systems for online communication.

CONCLUSIONS AND RECOMMENDATIONS

The study concludes that artificial intelligence (AI) and Natural Language Processing (NLP) methods provide effective and promising tools for the automatic detection of offensive language in online communication. The integration of text preprocessing, machine learning, and deep learning approaches enables offensive and non-offensive content to be classified efficiently, supporting the development of automated and scalable content moderation systems.

The findings further demonstrate that contextual understanding plays a significant role in enhancing detection accuracy, particularly when processing sarcasm, slang, implicit insults, and ambiguous expressions. Although AI-based systems provide valuable support for automated content moderation, human oversight remains an important complementary component for interpreting complex or context-sensitive cases. The combination of automated analysis and human judgment can therefore contribute to more reliable and balanced moderation outcomes.

Future research can focus on expanding the dataset with larger and more diverse multilingual collections, applying transformer-based language models, and developing more context-sensitive classification techniques. These directions can further strengthen the adaptability, accuracy, and reliability of offensive language detection systems. Overall, the development and continuous improvement of AI-based moderation technologies can contribute to safer, more respectful, and more inclusive digital communication environments.

REFERENCES

1. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, *11*(1), 512–515.
2. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of the NAACL Student Research Workshop, 88–93.
3. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). Proceedings of the 13th International Workshop on Semantic Evaluation, 75–86.
4. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., et al. (2020). SemEval-2020 Task 12: Multilingual offensive language identification in social media (OffensEval 2020). Proceedings of the 14th International Workshop on Semantic Evaluation, 1425–1447.
5. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys, *51*(4), Article 85, 1–30.
6. Fortuna, P., Soler-Company, J., Wanner, L., et al. (2020). Toxic, hateful, offensive or abusive? What are we really detecting? Proceedings of the 12th Language Resources and Evaluation Conference *(LREC)*, 6230–6239.

muhandislik & iqtisodiyot

ijtimoiy-iqtisodiy, innovatsion texnik,
fan va ta'limga oid ilmiy-amaliy jurnal

Ingliz tili muharriri: Feruz Hakimov

Musahhih: Zokir Alibekov

Sahifalovchi va dizayner: Abdurahmon Qurbonov

2026. № 9

"Muhandislik va iqtisodiyot" jurnali 26.06.2023-yildan
O'zbekiston Respublikasi Prezidenti Adminstratsiyasi huzuridagi
Axborot va ommaviy kommunikatsiyalar agentligi tomonidan
№S–5669245 reyestr raqami tartibi bo'yicha ro'yxatdan o'tkazilgan.
Litsenziya raqami: №095310.

**Manzilimiz: Toshkent shahri Yunusobod
tumani 15-mavze 19-uy**

+998 93 718 40 07 https://muhandislik-iqtisodiyot.uz/index.php/journal t.me/yait_2100