



ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ/ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI: <https://doi.org/10.60797/IRJ.2026.172.47> EDN: MGORKK**DEVELOPMENT AND VALIDATION OF FAKEKAPCHA: INTELLIGENT TOOL FOR MISINFORMATION DETECTION ON SOCIAL MEDIA**

Research article

Kholopov R.A.¹, Raheem Q.², Kiryowa I.³, Kabanova D.A.^{4,*}¹ORCID : 0009-0009-5112-3960;²ORCID : 0009-0001-4526-4814;³ORCID : 0009-0007-3178-0424;^{1,4} St. Petersburg College of Telecommunications named after E.T. Krenkel, Saint-Petersburg, Russian Federation² SRM University, Andhra Pradesh, India³ ITMO University, Saint-Petersburg, Russian Federation

* Corresponding author (kabanova.da[at]sut.ru)

Suggested: 26.06.2026; Accepted: 21.09.2026; Published: 16.10.2026

Abstract

The rapid growth of social media has accelerated the spread of misinformation, creating a demand for reliable automated detection methods. This paper presents FakeKapcha, a software framework for misinformation detection in Russian-language social media using natural language processing and ensemble machine learning. The proposed system was trained and validated on a dataset of more than 350,000 VK publications supplemented with the WELFake dataset. An ensemble of 11 machine learning models provided robust and accurate classification while producing probabilistic estimates of misinformation. The developed desktop application enables automated analysis of social media content and supports practical misinformation monitoring. Experimental results demonstrate high classification performance and confirm the effectiveness of the proposed ensemble approach for large-scale Russian-language social media analysis.

Keywords: misinformation detection, social media analysis, ensemble learning, natural language processing, Russian-language news, VK API.

РАЗРАБОТКА И ВАЛИДАЦИЯ ФАКЕКАРШНА: ИНТЕЛЛЕКТУАЛЬНЫЙ ИНСТРУМЕНТ ДЛЯ ВЫЯВЛЕНИЯ ДЕЗИНФОРМАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ

Научная статья

Холопов Р.А.¹, Рахим К.², Кирьова И.³, Кабанова Д.А.^{4,*}¹ORCID : 0009-0009-5112-3960;²ORCID : 0009-0001-4526-4814;³ORCID : 0009-0007-3178-0424;^{1,4} Санкт-Петербургский колледж телекоммуникаций им. Э.Т. Кренкеля, Санкт-Петербург, Российская Федерация² Университет SRM, Андхра-Прадеш, Индия³ Университет ИТМО, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (kabanova.da[at]sut.ru)

Предложена: 26.06.2026; Принята: 21.09.2026; Опубликовано: 16.10.2026

Аннотация

Быстрый рост социальных сетей ускорил распространение дезинформации, что привело к потребности в надежных методах автоматического обнаружения. В данной статье представлена FakeKapcha — программная платформа для обнаружения дезинформации в русскоязычных социальных сетях с использованием обработки естественного языка и ансамблевого машинного обучения. Предлагаемая система была обучена и проверена на наборе данных, включающем более 350 000 публикаций в «ВКонтакте», дополненном набором данных WELFake. Ансамбль из 11 моделей машинного обучения обеспечил надёжную и точную классификацию, одновременно давая вероятностные оценки дезинформации. Разработанное настольное приложение позволяет автоматизировать анализ контента социальных сетей и поддерживает практический мониторинг дезинформации. Результаты экспериментов демонстрируют высокую эффективность классификации и подтверждают результативность предложенного ансамблевого подхода для крупномасштабного анализа русскоязычных социальных сетей.

Ключевые слова: обнаружение дезинформации, анализ социальных сетей, ансамблевое обучение, обработка естественного языка, новости на русском языке, API ВКонтакте.

Introduction

The rise of digital communication platforms has transformed how information is created, shared, and consumed. Social media now serves as a key channel for news, providing users with immediate access to global events. However, the absence of strict verification processes and the rapid spread of information have led to the widespread distribution of misinformation and manipulated narratives. This can distort public perception, destabilize social discourse, and pose systemic risks online [1], [2].

The rise of unverified content highlights the shortcomings of traditional manual fact-checking methods, which are too slow and labor-intensive for the fast-paced data streams of social networks [3]. As a result, there is a pressing need for automated systems using AI and machine learning to improve digital information integrity and safety. While significant progress has been made, most existing frameworks focus on English-language data, leaving regional social media, particularly Russian-language content, inadequately addressed due to its unique linguistic and communication challenges [4]. Current solutions struggle on platforms like VK due to differing linguistic structures, user behaviors, and content dynamics compared to mainstream datasets.

Table 1 - Comparison of existing methodologies for misinformation detection across diverse linguistic contexts and digital infrastructures

DOI: <https://doi.org/10.60797/IRJ.2026.172.47.1>

Methodological Approach	Data Scope	Language Focus	Limitations in Scalability	Key Advantages of the Proposed FakeKapcha Framework	Reference
Supervised Learning	Global Datasets	Primarily English	High computational cost and data imbalance	11-model ensemble; Russian dataset	[5]
Automated Heuristic Analysis	Localized Communities	Restricted Linguistic Support	Inefficient processing of non-standard vernacular and hybrid structures	Data-driven ML; VK-specific processing	[6]
Ensemble Machine Learning	Large-scale datasets	Multilingual/Regional	Adaptability issues regarding rapidly evolving disinformation tactics	Three-level output; integrated desktop tool	[7]

This research develops and validates an automated framework, "FakeKapcha", for detecting misinformation in Russian-language social networks like VK. Instead of relying on manual verification or binary filters, it utilizes a scalable, multi-tiered probability assessment pipeline that addresses localized language and inauthentic behavior [9]. The system uses automated data collection from the VK API, compiling over 350,000 records, then text cleaning and employing high-dimensional TF-IDF feature extraction, the data undergoes processing through a stacked ensemble of eleven machine learning and deep learning algorithms. By combining linguistic signals with metadata, the framework computes a misinformation probability score and classifies content into three categories: Reliable, Potential Misinformation, or Highly Probable Fake Content presented via a user-friendly desktop GUI for real-time analysis [8], [10]. This work presents significant advancements in automated text classification through several interconnected contributions. It introduces a newly curated dataset of over 350,590 publications to address gaps in open-source regional language tools. Our heterogeneous ensemble architecture combines 11 diverse machine learning and deep learning models, establishing a robust predictive baseline for high-dimensional text. This framework offers unique non-binary interpretability, effectively capturing epistemic uncertainty in live data. The findings provide insights into how large-margin linear classifiers and well-regularized probabilistic models manage semantic noise without relying on resource-intensive transformer models. Additionally, this framework is a desktop application for non-technical users like fact-checkers and policy analysts, aiding their efforts against disinformation. This paper details the workflow, pipeline mechanics, classifier performance, current system limitations, and future directions, including user behavior signals and adaptive model retraining.

Research methods and principles

The proposed FakeKapcha system detects fake news from social network communities (e.g., VK) using machine learning and natural language processing. It functions as a sequential pipeline, converting raw social media content into structured classification results. Figure 1 illustrates the system architecture, detailing the flow from data collection to output generation.

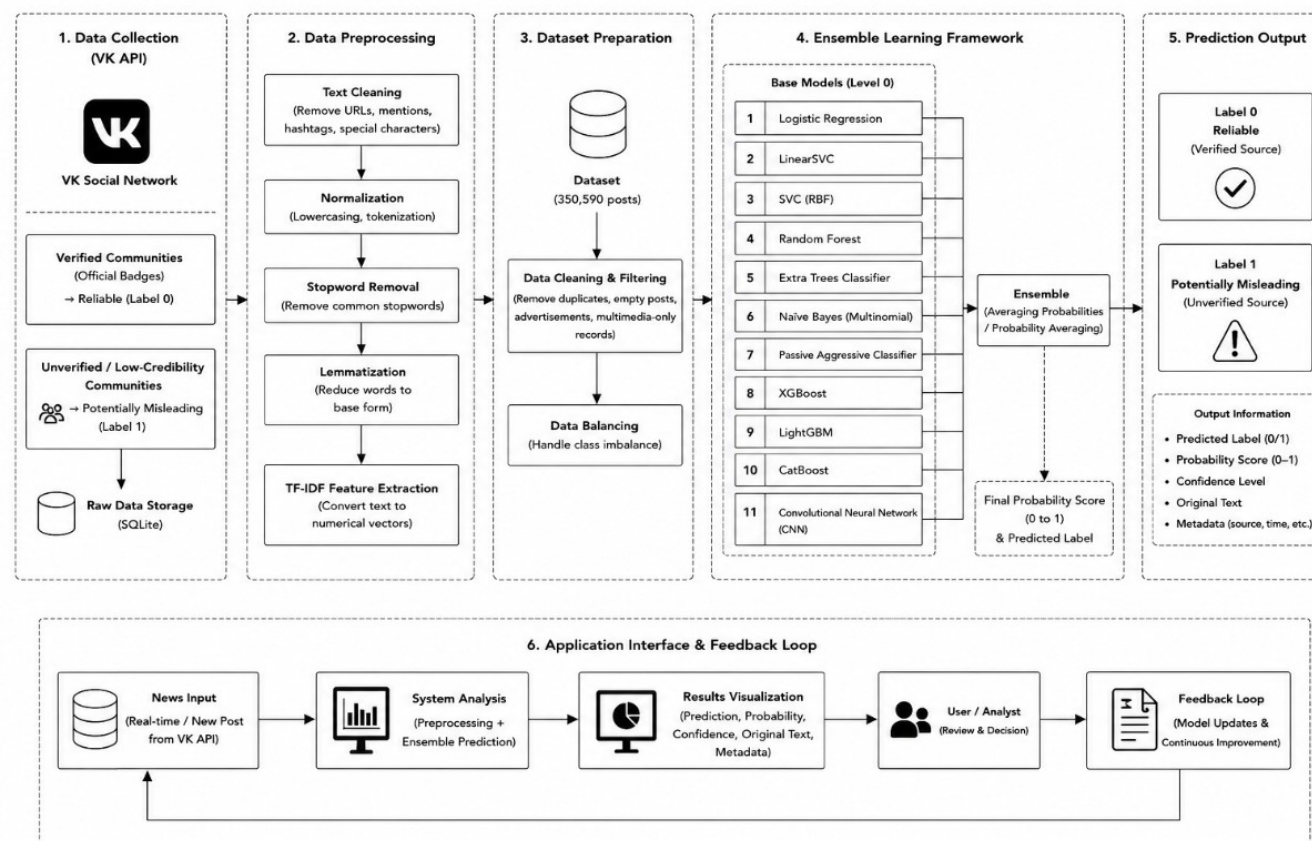


Figure 1 - Overall Architecture of the Proposed FakeKapcha Misinformation Detection Framework

DOI: <https://doi.org/10.60797/IRJ.2026.172.47.2>

Data acquisition: textual data and structured metadata, such as content, publication time, and engagement metrics (likes, comments, reposts, views), are fetched from VK communities through authenticated requests to the wall via the VK API. All records are stored in a local SQLite database for processing.

Dataset construction and labelling: the training dataset was created from VK social network data via the official VK API, focusing on verified communities and unverified, low-credibility channels. Low-credibility communities were identified based on the lack of official verification, frequent sharing of unverified content, and manual reputation assessments. Communities focused on entertainment or advertisements were excluded. Publications were labeled: Label 0 for reliable content from verified sources and Label 1 for potentially misleading posts from unverified sources. The final dataset included 350,590 text samples, with duplicates, empty posts, ads, and multimedia-only entries removed to ensure quality.

Experimental Design and Data Partitioning: the dataset, comprising 350,590 labeled VK publications, was divided into training, validation, and test subsets using stratified sampling to preserve class distributions. Specifically, 70% was used for training, 15% for validation, and 15% for testing. This approach ensured balanced representation of both reliable and misleading classes, minimizing sampling bias.

All preprocessing, including text normalization and TF-IDF feature extraction, was performed only on the training set to avoid information leakage. Machine learning models were trained on the same partitions for fair comparisons, and performance was evaluated using Accuracy, Precision, Recall, and F1-score on the unseen test dataset.

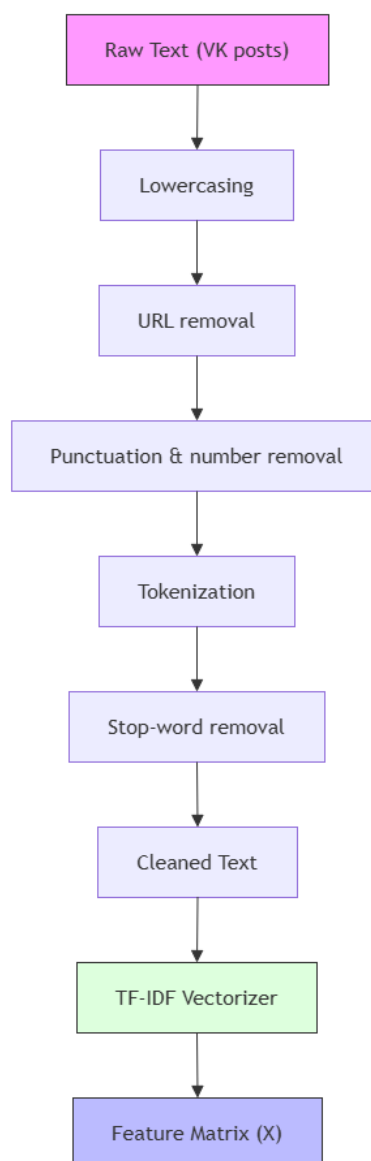


Figure 2 - Text Preprocessing and TF-IDF Feature Matrix Construction Pipeline

DOI: <https://doi.org/10.60797/IRJ.2026.172.47.3>

Dataset balancing: to address class imbalance, a balancing procedure is applied before model training to ensure that neither class dominates the learning process. This minimizes bias toward the majority class, enhancing model stability and classification reliability.

Text preprocessing: raw text is preprocessed to reduce noise and ensure consistency. This involves converting text to lowercase, removing URLs, punctuation, special characters, and numbers, and then tokenizing, filtering stop words, and normalizing whitespace.

Feature extraction: the cleaned text is vectorized using TF-IDF, a method that emphasizes informative terms while reducing the weight of frequently occurring words to create suitable input for the machine learning models.

Machine learning classification: the prediction engine integrates eleven diverse machine learning classifiers, including Logistic Regression, Support Vector Machines, Random Forest, Extra Trees, Naive Bayes, XGBoost, LightGBM, CatBoost, and a Convolutional Neural Network. Each model generates an independent probability score indicating whether a publication is fake, and the final decision is made by averaging these scores:

$$P(\text{final}) = (1/N) \sum P_i,$$

where N equals 11 [11]. The ensemble approach enhances stability and reduces overfitting risks compared to individual models (Fig. 3). By aggregating diverse algorithmic predictions, this stacked ensemble improves classification robustness and generalizability, this minimizes model biases and maintains a balanced diagnostic performance [12].

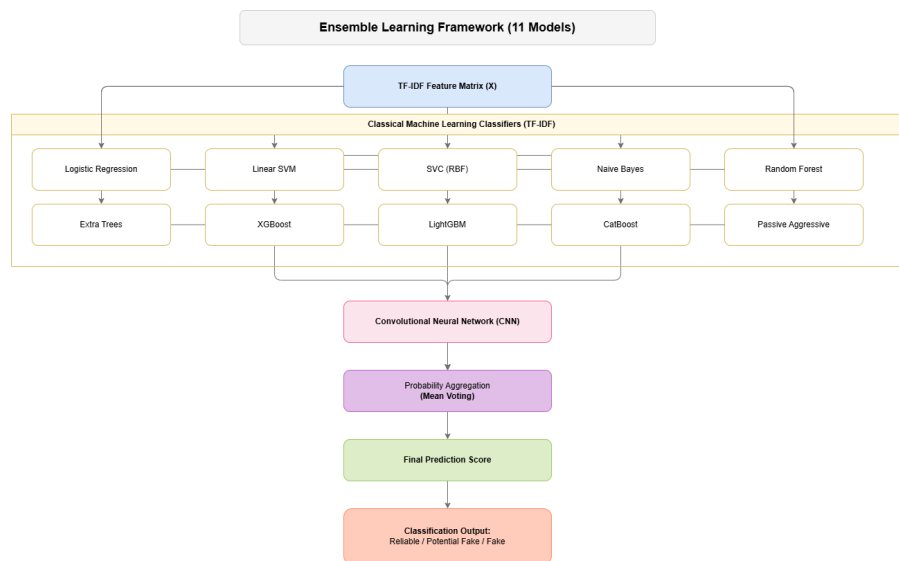


Figure 3 - Ensemble learning architecture of the proposed FakeKapcha system

DOI: <https://doi.org/10.60797/IRJ.2026.172.47.4>

Hardware and Software Specifications: system development and experimental validation were conducted in a Python 3.10 environment under Windows 11 (64-bit). The graphical user interface (GUI) was built using PyQt6, while local analytical history and user settings are managed via an embedded SQLite 3 database. The core execution pipeline relies exclusively on central processing units (CPU), eliminating the requirement for dedicated high-performance GPUs. The minimum hardware specifications required for stable deployment include a dual-core processor (Intel Core i3 / AMD Ryzen 3 equivalent), 4–8 GB of RAM, and up to 3 GB of available disk space (to store serialized .joblib model artifacts and vectorizers). Peak memory consumption during full ensemble initialization didn't exceed 2 GB.

Training Time and Resource Consumption: the training of the base ensemble classifiers was conducted on a balanced dataset consisting of approximately 70,000 records. Text vectorization was performed using TF-IDF, which created 10,000 features based on unigrams and bigrams, taking between 2 and 5 minutes. The complete sequential training and validation pipeline for all 11 classical models on a CPU required between 10 and 30 minutes. Experimental fine-tuning of transformer-based architectures, such as BERT and RoBERTa, resulted in a slight increase in accuracy (approximately 95–96%). However, this process demanded several hours of CPU computation and increased model storage to 4,4 GB. As a result, a classical machine learning ensemble was chosen for the production version to effectively balance predictive precision with low hardware requirements.

Result generation and visualization: the system outputs the predicted class, probability score, confidence level, original text, and metadata. According to P. Kulkarni et al. (2021), the use of a three-level output scheme: reliable information, potential misinformation, and fake information, enhances interpretability for uncertain cases [13]. All results are stored in a local SQLite database and presented through a graphical user interface for real-time analysis and visualization. The desktop interface loads the model ensemble in up to 60 seconds. For classifying individual publications, the process includes text preprocessing, vectorization, and model voting, with core latency between 100 and 500 milliseconds. If automatic translation through an external API is needed, processing takes 1 to 3 seconds. In batch processing of community feeds, the system can handle over 5 posts per second, and re-analyzed texts return cached results instantly.

Results and Discussion

The FakeKapcha system was validated using a dataset of 350,590 publications from VK communities via the VK API. Eleven machine learning algorithms were compared under identical conditions using TF-IDF features, with performance assessed by Accuracy, Precision, Recall, and F1-score. The Passive Aggressive algorithm achieved the highest score (0,972), followed by LinearSVC and LightGBM (0,971), XGBoost (0,963), and CNN (0,961). Logistic Regression and CatBoost were around 0,955, while Extra Trees and Random Forest scored between 0,922 and 0,928. Naive Bayes had the lowest performance at 0,896. These results highlight the effectiveness of ensemble learning in creating a strong classification framework that outperforms individual models, as shown by the comparative heat map in Figure 4 [14], [15].

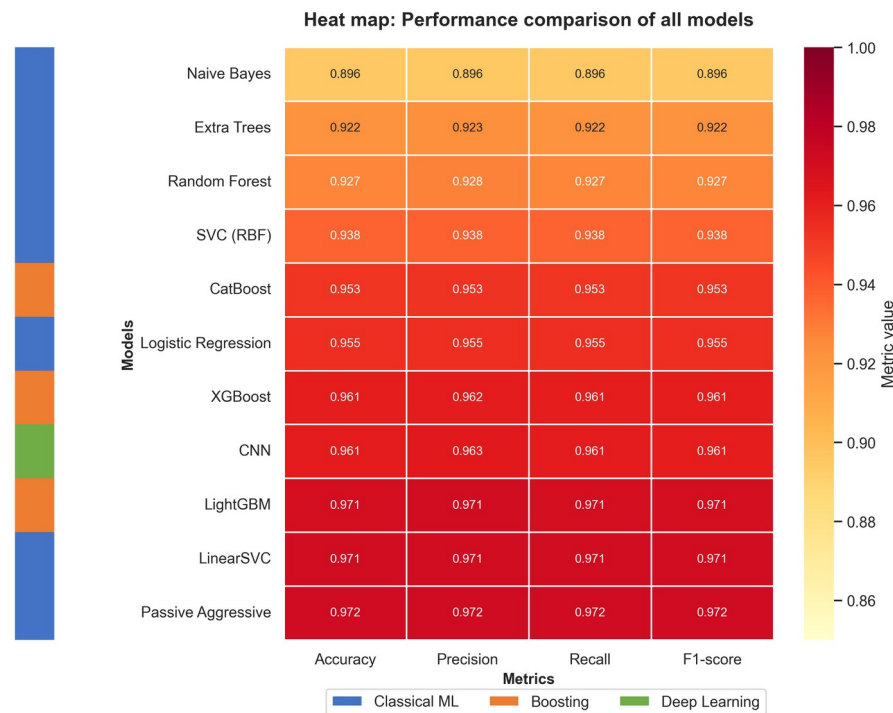


Figure 4 - Comparative evaluation of machine learning models based on Accuracy, Precision, Recall, and F1-score
DOI: <https://doi.org/10.60797/IRJ.2026.172.47.5>

The warmer tones in the heat map indicate high performance for models like Passive Aggressive, LinearSVC, LightGBM, and XGBoost, scoring between 0,955 and 0,972. In contrast, cooler tones for Naive Bayes, Extra Trees, and Random Forest reflect lower performance, ranging from 0,896 to 0,928. This visualization facilitates comparative evaluation of model performance by highlighting outliers and performance gaps without numerical comparisons [16]. The minimal variation across evaluation metrics within each algorithm suggests that top models strike a balance between Precision and Recall, rather than excelling in just one area. However, the uniformity across the heat map confirms that the ensemble improves accuracy while maintaining consistency across all evaluation dimensions, showing that selected base-learners using TF-IDF can effectively classify misinformation without complex architectures [17].

The radar chart (Fig. 5) displays the distribution of Accuracy, Precision, Recall, and F1-score for the top three classifiers: Passive Aggressive, LinearSVC, and LightGBM. Passive Aggressive demonstrates the highest overall performance across all evaluation metrics, indicating superior classification quality, while LinearSVC and LightGBM have slightly lower values. This comparison illustrates that all models maintain balanced performance across metrics, essential for reliable classification and preventing bias. These results align with literature indicating that ensemble architectures, especially with robust TF-IDF representations, enhance discriminatory power compared to individual classifiers [12], [14].

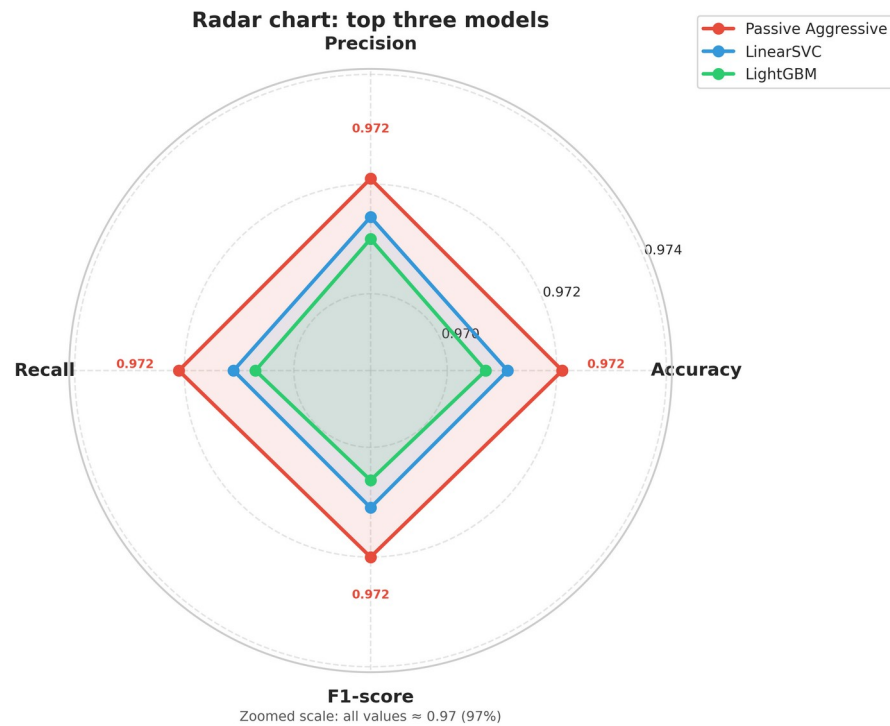


Figure 5 - Radar chart comparison of the highest-performing machine learning classifiers

DOI: <https://doi.org/10.60797/IRJ.2026.172.47.6>

Precision–Recall curves illustrate classification stability across varying decision thresholds. Passive Aggressive and Logistic Regression achieve high precision, with an Average Precision (AP) of 0,989, while Naive Bayes shows a decline in precision as recall increases, resulting in a lower AP of 0,917. This sensitivity to class imbalance is evident in the curves, where Naive Bayes exhibits a more pronounced decline in precision as recall increases before reaching high recall, while Passive Aggressive and Logistic Regression maintain nearly horizontal paths near unity. This demonstrates their ability to preserve high precision even when lowering thresholds to capture more positive instances, contrasting with Naive Bayes' performance, influenced by feature independence and skewed posterior estimates [18], [19], [20].

PR curves provide a more diagnostic view than ROC plots for the imbalanced VK corpus, as evidenced by the significant AP difference of around seven percentage points between the upper curves and the Naive Bayes trajectory [19]. Passive Aggressive and Logistic Regression demonstrate robustness due to their mathematical foundations: Passive Aggressive effectively adapts to high-dimensional TF-IDF vectors, while Logistic Regression models stable, calibrated probabilities across thresholds. In contrast, Naive Bayes struggles with rare co-occurring terms [18]. Overall, the PR visualization supports the view that large-margin linear classifiers and well-regularized probabilistic models maintain consistent classification quality, while those relying on strong independence assumptions perform poorly as recall increases, highlighting the need for residual-based and ensemble architectures in the FakeKapcha pipeline [21].

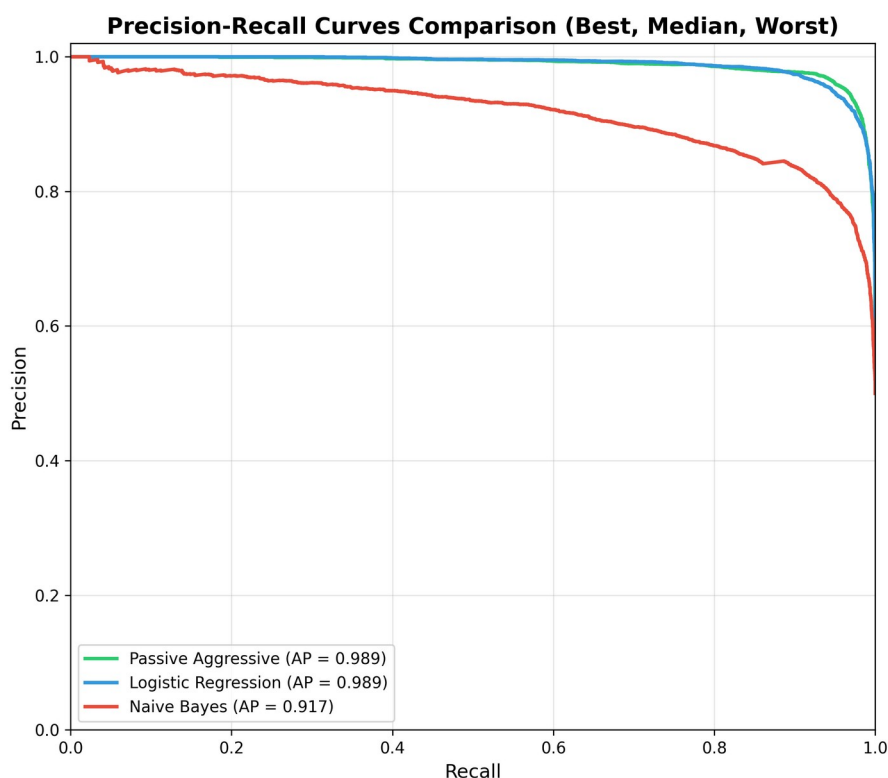


Figure 6 - Precision-Recall performance analysis of representative classification models
DOI: <https://doi.org/10.60797/IRJ.2026.172.47.7>

The developed application's operation is shown after completing publication analysis (see Fig. 7). The graphical interface displays 20 analyzed VK posts, including 16 reliable and 4 suspicious publications identified by the classification system. It also includes confidence scores and voting statistics from the ensemble of 11 machine learning algorithms. All evaluation metrics were computed on an independent testing dataset, confirming high classification performance and stable operation of the software when processing large volumes of social network publications.

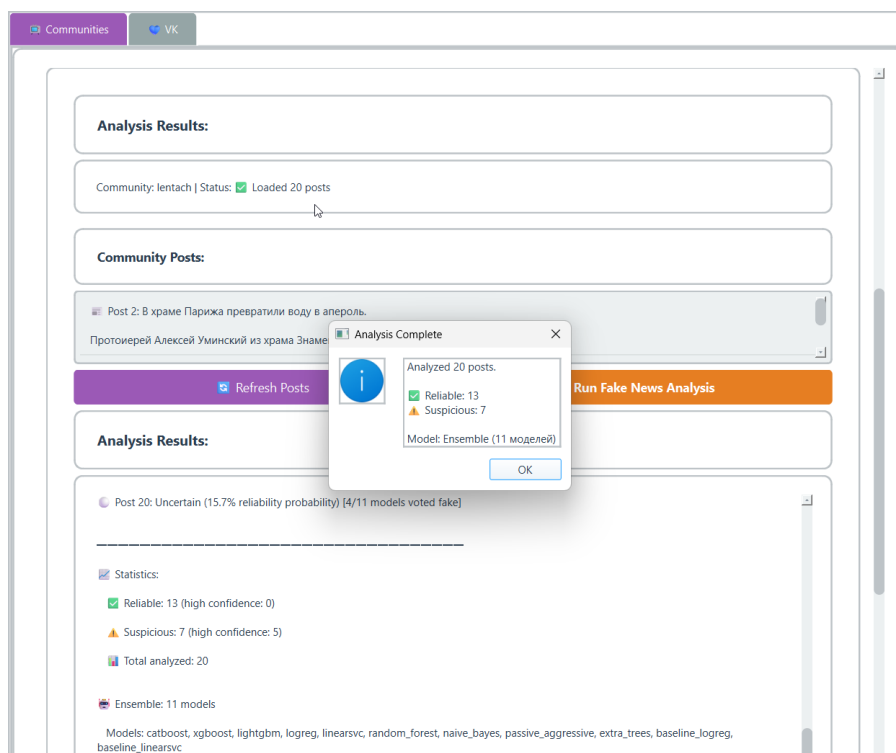


Figure 7 - Graphical interface of the FakeKapcha system demonstrating misinformation detection results
DOI: <https://doi.org/10.60797/IRJ.2026.172.47.8>



Practical Utility and Cost-Effectiveness

With the increasing costs of server-grade hardware and limited access to paid cloud AI APIs, the proposed tool offers significant practical benefits for everyday use. Operating as an autonomous desktop application, it allows for real-time verification of VK social media streams without the need for cloud infrastructure, monthly API subscriptions, or expensive GPU hardware. This system serves as a lightweight, privacy-preserving decision-support tool for editors, community moderators, and general users who want to verify local content on VK.

Ethical consideration

Even though automated classification tools offer valuable support, they should not be misconstrued as instruments for determining absolute truth [22]. Machine learning models rely on statistical patterns and may produce errors when handling satire, irony, opinion-based articles, rapidly changing news, or content that lacks sufficient context. As a result, these systems function as decision-support tools that assist users rather than supplanting human judgment. This distinction is vital due to the risks of false positives, which can erode trust in legitimate sources, and false negatives, which may allow misleading content to persist. Transparency remains an ethical imperative; users need to understand that the results are probabilistic estimates, not definitive verifications. Thus, incorporating confidence indicators, probability scores, and a "Potential Misinformation" category helps convey the uncertainty that accompanies automated predictions [23].

Future Directions

Further research and development of the FakeKapcha tool should assertively expand beyond VK, creating a unified platform capable of processing content from other platforms such as Telegram, X, Facebook, Reddit, and other online news portals [4], [13], [22]. Thereby, our integration of transformer-based language models specifically designed for Russian-language content is expected to improve detection accuracy.

The tools should be allowed to tap into and expand a dataset and self-model retrain, which will ensure that the system remains resilient and adaptive to evolving information ecosystems.

To maintain long-term reliability, future versions of the system will incorporate continuous performance monitoring using data drift indicators and predictive quality metrics. The ensemble performance will be periodically evaluated through Accuracy, Precision, Recall, and F1-score on newly labeled validation samples. A decrease in the F1-score of more than 5% relative to the baseline performance or statistically significant changes in feature distributions will be used as indicators for initiating ensemble retraining. The retraining process can be planned after accumulating sufficient newly verified data from supported platforms, ensuring that the framework remains robust against evolving misinformation patterns.

Conclusion

Modern information environments increasingly require automated verification technologies. This article demonstrates the successful application of machine learning to detect misinformation in Russian-language social media content. The FakeKapcha system implements a comprehensive pipeline, including automated data collection via the VK API, text preprocessing, TF-IDF feature extraction, ensemble classification, and result visualization. Using a custom dataset of over 350,000 publications, we confirmed that this ensemble strategy enhances prediction stability and robustness for heterogeneous textual data. A key innovation is a three-level classification mechanism that categorizes content as Reliable Information, Potential Misinformation, or Fake Information, improving interpretability by enabling nuanced, non-binary assessments alongside transparent confidence scores. The system effectively integrates automated data acquisition, natural language processing, and user-oriented design to assess information reliability on social media. Future development aims to expand platform support, create cross-platform mobile applications, and incorporate multimodal analysis and advanced language models to broaden the system's applicability.

Благодарности

Авторы выражают глубокую признательность Санкт-Петербургскому колледжу телекоммуникаций имени Е. Т. Кренкеля за огромные усилия и оказанную поддержку, а также студентам магистратуры Университета ИТМО и Университета SRM (Индия) за сотрудничество; компании «VK Technologies» — за предоставление открытого доступа к данным.

Конфликт интересов

Не указан.

Рецензия

Сообщество рецензентов Международного научно-исследовательского журнала
DOI: <https://doi.org/10.60797/IRJ.2026.172.47.9>

Acknowledgement

Authors acknowledge the substantial efforts, and guidance from St. Petersburg College of Telecommunications named after E. T. Krenkel, and collaboration of graduate students from ITMO University and SRM University (India); VK technologies for the open data availability.

Conflict of Interest

None declared.

Review

Community of Reviewers of the International Research Journal
DOI: <https://doi.org/10.60797/IRJ.2026.172.47.9>

Список литературы / References

1. Меркушина Е.А. Пользовательский контент в социальной сети «ВКонтакте»: вызов медиаграмотности пользователей / Е.А. Меркушина // Динамика медиасистем. — 2025. — Т. 5. — № 1. — С. 128–139. — DOI: 10.47475/2949-3390-2025-5-1-128-139. — EDN ECQXVI.
2. Tomassi A. Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society / A. Tomassi, A. Falegnami, E. Romano // PLOS ONE. — 2024. — Vol. 19. — № 5. — DOI: 10.1371/journal.pone.0303183.



3. Rashid M.T. CovidSens: a vision on reliable social sensing for COVID-19 / M.T. Rashid, D. Wang // *Artificial Intelligence Review*. — 2020. — Vol. 54. — № 1. — P. 1–25. — DOI: 10.1007/s10462-020-09852-3.
4. Park C.Y. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media / C.Y. Park, J. Mendelsohn, A. Field [et al.] // *Findings of the Association for Computational Linguistics: EMNLP 2022*. — 2022. — P. 5209–5235. — DOI: 10.18653/v1/2022.findings-emnlp.382.
5. Madrap B. Design Architecture/Layout for Detecting Fake News and Misinformation / B. Madrap // *Zenodo*. — 2026. — DOI: 10.5281/zenodo.18921511.
6. Wang X. Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey / X. Wang, W. Zhang, S. Rajtmajer // *arXiv*. — 2024. — DOI: 10.48550/arXiv.2410.18390.
7. Vysotska V. Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods / V. Vysotska, M. Nazarkevych // *Eastern-European Journal of Enterprise Technologies*. — 2025. — Vol. 4. — № 2 (136). — P. 35–51. — DOI: 10.15587/1729-4061.2025.335501.
8. Mini T.V. Natural Language Processing Techniques for Misinformation Detection in Social Media: A Comparative Analysis / T.V. Mini, K. Abraham, P.A. Julie [et al.] // *International Scientific Journal of Engineering and Management*. — 2025. — Vol. 4. — № 11. — P. 1–9. — DOI: 10.55041/isjem05147.
9. Cinus F. Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 U.S. Election / F. Cinus, M. Minici, L. Luceri [et al.] // *SSRN Electronic Journal*. — 2025. — DOI: 10.2139/ssrn.5018877.
10. Ahmad I. Fake News Detection Using Machine Learning Ensemble Methods / I. Ahmad, M.M. Yousaf, S. Yousaf [et al.] // *Complexity*. — 2020. — Vol. 2020. — P. 1–11. — DOI: 10.1155/2020/8885861.
11. Padalko H. Ensemble machine learning approaches for fake news classification / H. Padalko, V. Chomko, S. Yakovlev [et al.] // *Radioelectronic and Computer Systems*. — 2023. — № 4. — P. 5–19. — DOI: 10.32620/reks.2023.4.01.
12. Aslam Z. Advancing fake news combating using machine learning: a hybrid model approach / Z. Aslam, M.M.S. Missen, A.A. Ghaffar [et al.] // *Knowledge and Information Systems*. — 2025. — Vol. 67. — № 12. — P. 12137–12177. — DOI: 10.1007/s10115-025-02588-y.
13. Kulkarni P. Fake News Detection using Machine Learning / P. Kulkarni, S. Karwande, R. Keskar [et al.] // *ITM Web of Conferences*. — 2021. — Vol. 40. — DOI: 10.1051/itmconf/20214003003.
14. Yigezu M.G. Ethio-Fake: Cutting-Edge Approaches to Combat Fake News in Under-Resourced Languages Using Explainable AI / M.G. Yigezu, M.A. Mersha, G.Y. Bade [et al.] // *arXiv*. — 2024. — DOI: 10.48550/arXiv.2410.02609.
15. Verma S. Comparative Analysis of Image Classification Algorithms / S. Verma // *International Journal for Research in Applied Science & Engineering Technology*. — 2023. — Vol. 11. — № 12. — P. 1513–1520. — DOI: 10.22214/ijraset.2023.57662.
16. Ansari G.A. Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection / G.A. Ansari, S. Shafi, M.D. Ansari [et al.] // *Frontiers in Medicine*. — 2025. — Vol. 12. — DOI: 10.3389/fmed.2025.1620268.
17. Suh A. Are Metrics Enough? Guidelines for Communicating and Visualizing Predictive Models to Subject Matter Experts / A. Suh, G. Appleby, E. Anderson [et al.] // *IEEE Transactions on Visualization and Computer Graphics*. — 2023. — Vol. 30. — № 7. — P. 4137–4153. — DOI: 10.1109/TVCG.2023.3259341.
18. Terven J. A comprehensive survey of loss functions and metrics in deep learning / J. Terven, D.M. Córdova-Esparza, J.A. Romero-González [et al.] // *Artificial Intelligence Review*. — 2025. — Vol. 58. — № 7. — DOI: 10.1007/s10462-025-11198-7.
19. Saito T. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets / T. Saito, M. Rehmsmeier // *PLOS ONE*. — 2015. — Vol. 10. — № 3. — DOI: 10.1371/journal.pone.0118432.
20. Rennie J.D.M. Tackling the poor assumptions of naive bayes text classifiers / J.D.M. Rennie, L. Shih, J. Teevan [et al.] // *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*. — 2003. — P. 616–623.
21. Johnson J. Survey on deep learning with class imbalance / J. Johnson, T.M. Khoshgoftaar // *Journal of Big Data*. — 2019. — Vol. 6. — № 1. — DOI: 10.1186/s40537-019-0192-5.
22. Arranz-Escudero O. Enhancing misinformation countermeasures: a multimodal approach to twitter bot detection / O. Arranz-Escudero, L. Quijano-Sánchez, F. Liberatore // *Social Network Analysis and Mining*. — 2025. — Vol. 15. — № 1. — DOI: 10.1007/s13278-025-01435-w.
23. Mujahid Z.M. Profiling News Media for Factuality and Bias Using LLMs and the Fact-Checking Methodology of Human Experts / Z.M. Mujahid, D. Azizov, M.T. Agro [et al.] // *arXiv*. — 2025. — DOI: 10.48550/arXiv.2506.12552.

Список литературы на английском языке / References in English

1. Merkushina E.A. Pol'zovatel'skij kontent v social'noj seti "Vkontakte": vyzov mediagramotnosti pol'zovatelej [User-generated content on the VKontakte social network: a challenge to users' media literacy] / E.A. Merkushina // *Dinamika mediasistem [Dynamics of Media Systems]*. — 2025. — Vol. 5. — № 1. — P. 128–139. — DOI: 10.47475/2949-3390-2025-5-1-128-139. — EDN ECQXVI. [in Russian]
2. Tomassi A. Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society / A. Tomassi, A. Falegnami, E. Romano // *PLOS ONE*. — 2024. — Vol. 19. — № 5. — DOI: 10.1371/journal.pone.0303183.
3. Rashid M.T. CovidSens: a vision on reliable social sensing for COVID-19 / M.T. Rashid, D. Wang // *Artificial Intelligence Review*. — 2020. — Vol. 54. — № 1. — P. 1–25. — DOI: 10.1007/s10462-020-09852-3.



4. Park C.Y. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media / C.Y. Park, J. Mendelsohn, A. Field [et al.] // Findings of the Association for Computational Linguistics: EMNLP 2022. — 2022. — P. 5209–5235. — DOI: 10.18653/v1/2022.findings-emnlp.382.
5. Madrap B. Design Architecture/Layout for Detecting Fake News and Misinformation / B. Madrap // Zenodo. — 2026. — DOI: 10.5281/zenodo.18921511.
6. Wang X. Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey / X. Wang, W. Zhang, S. Rajtmajer // arXiv. — 2024. — DOI: 10.48550/arXiv.2410.18390.
7. Vysotska V. Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods / V. Vysotska, M. Nazarkevych // Eastern-European Journal of Enterprise Technologies. — 2025. — Vol. 4. — № 2 (136). — P. 35–51. — DOI: 10.15587/1729-4061.2025.335501.
8. Mini T.V. Natural Language Processing Techniques for Misinformation Detection in Social Media: A Comparative Analysis / T.V. Mini, K. Abraham, P.A. Julie [et al.] // International Scientific Journal of Engineering and Management. — 2025. — Vol. 4. — № 11. — P. 1–9. — DOI: 10.55041/isjem05147.
9. Cinus F. Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 U.S. Election / F. Cinus, M. Minici, L. Luceri [et al.] // SSRN Electronic Journal. — 2025. — DOI: 10.2139/ssrn.5018877.
10. Ahmad I. Fake News Detection Using Machine Learning Ensemble Methods / I. Ahmad, M.M. Yousaf, S. Yousaf [et al.] // Complexity. — 2020. — Vol. 2020. — P. 1–11. — DOI: 10.1155/2020/8885861.
11. Padalko H. Ensemble machine learning approaches for fake news classification / H. Padalko, V. Chomko, S. Yakovlev [et al.] // Radioelectronic and Computer Systems. — 2023. — № 4. — P. 5–19. — DOI: 10.32620/reks.2023.4.01.
12. Aslam Z. Advancing fake news combating using machine learning: a hybrid model approach / Z. Aslam, M.M.S. Missen, A.A. Ghaffar [et al.] // Knowledge and Information Systems. — 2025. — Vol. 67. — № 12. — P. 12137–12177. — DOI: 10.1007/s10115-025-02588-y.
13. Kulkarni P. Fake News Detection using Machine Learning / P. Kulkarni, S. Karwande, R. Keskar [et al.] // ITM Web of Conferences. — 2021. — Vol. 40. — DOI: 10.1051/itmconf/20214003003.
14. Yigezu M.G. Ethio-Fake: Cutting-Edge Approaches to Combat Fake News in Under-Resourced Languages Using Explainable AI / M.G. Yigezu, M.A. Mersha, G.Y. Bade [et al.] // arXiv. — 2024. — DOI: 10.48550/arXiv.2410.02609.
15. Verma S. Comparative Analysis of Image Classification Algorithms / S. Verma // International Journal for Research in Applied Science & Engineering Technology. — 2023. — Vol. 11. — № 12. — P. 1513–1520. — DOI: 10.22214/ijraset.2023.57662.
16. Ansari G.A. Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection / G.A. Ansari, S. Shafi, M.D. Ansari [et al.] // Frontiers in Medicine. — 2025. — Vol. 12. — DOI: 10.3389/fmed.2025.1620268.
17. Suh A. Are Metrics Enough? Guidelines for Communicating and Visualizing Predictive Models to Subject Matter Experts / A. Suh, G. Appleby, E. Anderson [et al.] // IEEE Transactions on Visualization and Computer Graphics. — 2023. — Vol. 30. — № 7. — P. 4137–4153. — DOI: 10.1109/TVCG.2023.3259341.
18. Terven J. A comprehensive survey of loss functions and metrics in deep learning / J. Terven, D.M. Córdova-Esparza, J.A. Romero-González [et al.] // Artificial Intelligence Review. — 2025. — Vol. 58. — № 7. — DOI: 10.1007/s10462-025-11198-7.
19. Saito T. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets / T. Saito, M. Rehmsmeier // PLOS ONE. — 2015. — Vol. 10. — № 3. — DOI: 10.1371/journal.pone.0118432.
20. Rennie J.D.M. Tackling the poor assumptions of naive bayes text classifiers / J.D.M. Rennie, L. Shih, J. Teevan [et al.] // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). — 2003. — P. 616–623.
21. Johnson J. Survey on deep learning with class imbalance / J. Johnson, T.M. Khoshgoftaar // Journal of Big Data. — 2019. — Vol. 6. — № 1. — DOI: 10.1186/s40537-019-0192-5.
22. Arranz-Escudero O. Enhancing misinformation countermeasures: a multimodal approach to twitter bot detection / O. Arranz-Escudero, L. Quijano-Sánchez, F. Liberatore // Social Network Analysis and Mining. — 2025. — Vol. 15. — № 1. — DOI: 10.1007/s13278-025-01435-w.
23. Mujahid Z.M. Profiling News Media for Factuality and Bias Using LLMs and the Fact-Checking Methodology of Human Experts / Z.M. Mujahid, D. Azizov, M.T. Agro [et al.] // arXiv. — 2025. — DOI: 10.48550/arXiv.2506.12552.